盘古大模型

用户指南

文档版本 01

发布日期 2025-08-30



华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2025。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 盘古大模型服务使用流程.....1
2 准备工作.....13
2.1 申请试用 ModelArts Studio 大模型开发平台.....13
2.2 订购盘古大模型服务.....13
2.3 配置服务访问授权.....22
2.4 创建并管理工作空间.....22
2.4.1 工作空间介绍.....22
2.4.2 创建并管理工作空间.....22
2.4.3 管理工作空间成员.....25
3 在模型广场查看模型.....31
4 使用数据工程构建数据集.....35
4.1 数据工程介绍.....35
4.2 数据工程使用流程.....40
4.3 数据集格式要求.....41
4.3.1 文本类数据集格式要求.....41
4.3.2 图片类数据集格式要求.....46
4.3.3 视频类数据集格式要求.....69
4.3.4 音频类数据集格式要求.....72
4.3.5 气象类数据集格式要求.....73
4.3.6 预测类数据集格式要求.....75
4.3.7 其他类数据集格式要求.....76
4.4 导入数据至盘古平台.....77
4.5 加工数据集.....80
4.5.1 数据集加工场景介绍.....80
4.5.2 数据集加工算子介绍.....83
4.5.2.1 文本类加工算子介绍.....83
4.5.2.2 视频类加工算子介绍.....106
4.5.2.3 图片类加工算子介绍.....121
4.5.2.4 气象类加工算子介绍.....132
4.5.2.5 音频类加工算子介绍.....133
4.5.2.6 通用类加工算子介绍.....137
4.5.3 加工文本类数据集.....138

 版权所有 © 华为云计算技术有限公司

4.5.3.1 加工文本类数据集 139
4.5.3.2 合成文本类数据集 141
4.5.3.3 标注文本类数据集 149
4.5.3.4 配比文本类数据集 155
4.5.4 加工图片类数据集 157
4.5.4.1 加工图片类数据集 158
4.5.4.2 标注图片类数据集 161
4.5.4.3 配比图片类数据集 164
4.5.5 加工视频类数据集 166
4.5.5.1 加工视频类数据集 166
4.5.5.2 标注视频类数据集 168
4.5.5.3 配比视频类数据集 172
4.5.6 加工气象类数据集 174
4.5.7 加工预测类数据集 176
4.5.7.1 加工预测类数据集 176
4.5.7.2 配比预测类数据集 178
4.5.8 加工音频类数据集 181
4.5.8.1 加工音频类数据集 181
4.5.8.2 标注音频类数据集 183
4.5.9 加工其他类数据集 185
4.5.10 自定义数据集加工算子 187
4.5.10.1 算子配置文件规范 188
4.5.10.2 算子包开发规范 195
4.5.10.3 算子开发典型样例 208
4.5.10.4 管理自定义算子 211
4.5.11 使用日志查看数据集加工进度 217
4.5.12 管理加工后的数据集 219
4.5.13 管理加工任务资源 221
4.5.14 加工任务生成数据集 221
4.6 发布数据集 222
4.6.1 数据集发布场景介绍 223
4.6.2 发布文本类数据集 224
4.6.2.1 评估文本类数据集 224
4.6.2.2 发布文本类数据集 226
4.6.3 发布图片类数据集 230
4.6.3.1 评估图片类数据集 230
4.6.3.2 发布图片类数据集 232
4.6.4 发布视频类数据集 234
4.6.4.1 评估视频类数据集 234
4.6.4.2 发布视频类数据集 236
4.6.5 发布音频类数据集 237
4.6.6 发布气象类数据集 238

 版权所有 © 华为云计算技术有限公司

4.6.7 发布预测类数据集 239
4.6.8 发布其他类数据集 240
4.6.9 管理发布后的数据集 241
4.7 数据集格式转换 243
4.7.1 使用限制 243
4.7.2 格式转换操作流程 244
4.7.3 转换算子说明 246
4.8 数据工程常见报错与解决方案 251
5 开发盘古 NLP 大模型 256
5.1 使用数据工程构建 NLP 大模型数据集 256
5.2 训练 NLP 大模型 259
5.2.1 NLP 大模型训练流程与选择建议 260
5.2.2 NLP 大模型默认 prompt 266
5.2.3 创建 NLP 大模型训练任务 279
5.2.4 查看 NLP 大模型训练状态与指标 297
5.2.5 发布训练后的 NLP 大模型 301
5.2.6 管理 NLP 大模型训练任务 302
5.2.7 NLP 大模型训练常见报错与解决方案 303
5.3 压缩 NLP 大模型 305
5.4 部署 NLP 大模型 307
5.4.1 创建 NLP 大模型部署任务 307
5.4.2 查看 NLP 大模型部署任务详情 309
5.4.3 管理 NLP 大模型部署任务 310
5.4.4 NLP 大模型推理常见错误及解决方案 313
5.5 评测 NLP 大模型 315
5.5.1 创建 NLP 大模型评测集 316
5.5.2 创建 API 服务（可选） 318
5.5.3 创建 NLP 大模型评测任务 323
5.5.4 查看 NLP 大模型评测报告 329
5.5.5 管理 NLP 大模型评测任务 332
5.6 调用 NLP 大模型 333
5.6.1 使用“体验中心”调用 NLP 大模型 333
5.6.2 使用 API 调用 NLP 大模型 334
5.6.3 统计 NLP 大模型调用信息 338
6 开发盘古 CV 大模型 340
6.1 使用数据工程构建 CV 大模型数据集 340
6.2 训练 CV 大模型 342
6.2.1 CV 大模型训练流程与选择建议 342
6.2.2 创建 CV 大模型训练任务 344
6.2.3 查看 CV 大模型训练状态与指标 352
6.2.4 发布训练后的 CV 大模型 356
6.2.5 管理 CV 大模型训练任务 357

 版权所有 © 华为云计算技术有限公司

6.2.6 CV 大模型训练常见报错与解决方案 357
6.3 部署 CV 大模型 358
6.3.1 创建 CV 大模型部署任务 358
6.3.2 查看 CV 大模型部署任务详情 360
6.3.3 管理 CV 大模型部署任务 360
6.4 评测 CV 大模型 363
6.4.1 创建 CV 大模型评测数据集 363
6.4.2 创建 CV 大模型评测任务 364
6.4.3 查看 CV 大模型评测报告 365
6.4.4 管理 CV 大模型评测任务 367
6.5 调用 CV 大模型 367
6.5.1 使用 API 调用 CV 大模型 367
6.5.2 统计 CV 大模型调用信息 369
7 开发盘古预测大模型 371
7.1 使用数据工程构建预测大模型数据集 371
7.2 训练预测大模型 373
7.2.1 预测大模型训练流程与选择建议 373
7.2.2 创建预测大模型训练任务 375
7.2.3 查看预测大模型训练状态与指标 390
7.2.4 发布训练后的预测大模型 394
7.2.5 管理预测大模型训练任务 395
7.2.6 预测大模型训练常见报错与解决方案 396
7.3 部署预测大模型 397
7.3.1 创建预测大模型部署任务 397
7.3.2 查看预测大模型部署任务详情 398
7.3.3 管理预测大模型部署任务 399
7.4 评测预测大模型 402
7.4.1 创建预测大模型评测数据集 402
7.4.2 创建预测大模型评测任务 403
7.4.3 查看预测大模型评测报告 405
7.4.4 管理预测大模型评测任务 407
7.5 调用预测大模型 408
7.5.1 使用 API 调用预测大模型 408
7.5.2 统计预测大模型调用信息 412
8 开发盘古科学计算大模型 413
8.1 气象/海洋数据获取指南 414
8.2 使用数据工程构建科学计算大模型数据集 419
8.3 训练科学计算大模型 425
8.3.1 科学计算大模型训练流程与选择建议 425
8.3.2 创建科学计算大模型训练任务 431
8.3.3 查看科学计算大模型训练状态与指标 456
8.3.4 发布训练后的科学计算大模型 461

 版权所有 © 华为云计算技术有限公司

8.3.5 管理科学计算大模型训练任务 462
8.3.6 科学计算大模型训练常见报错与解决方案 462
8.4 部署科学计算大模型 463
8.4.1 创建科学计算大模型部署任务 463
8.4.2 查看科学计算大模型部署任务详情 465
8.4.3 管理科学计算大模型部署任务 466
8.5 调用科学计算大模型 469
8.5.1 使用“体验中心”调用科学计算大模型 469
8.5.2 使用 API 调用科学计算大模型 472
9 开发盘古搜索规划模型 475
9.1 搜索规划模型介绍 475
9.2 部署搜索规划大模型 475
9.2.1 创建 Pangu-SearchPlan-ZH_EN 模型部署任务 475
9.2.2 查看 Pangu-SearchPlan-ZH_EN 模型部署任务详情 476
9.2.3 管理 Pangu-SearchPlan-ZH_EN 模型部署任务 477
9.3 调用搜索规划大模型 480
10 开发盘古向量&重排模型 484
10.1 向量&重排模型介绍 484
10.2 部署向量&重排大模型 484
10.2.1 创建 Pangu-EmbeddingRank-zh 模型部署任务 484
10.2.2 查看 Pangu-EmbeddingRank-zh 模型部署任务详情 485
10.2.3 管理 Pangu-EmbeddingRank-zh 模型部署任务 486
10.3 调用向量&重排大模型 489
11 开发盘古行业大模型 493
11.1 盘古行业 NLP 大模型 493
11.1.1 盘古行业 NLP 大模型介绍 493
11.1.2 使用数据工程构建盘古行业 NLP 大模型数据集 493
11.1.3 训练盘古行业 NLP 大模型 495
11.1.3.1 创建盘古行业 NLP 大模型训练任务 498
11.1.3.2 查看盘古行业 NLP 大模型训练任务状态与指标 506
11.1.3.3 发布训练后的盘古行业 NLP 大模型 509
11.1.3.4 管理盘古行业 NLP 大模型训练任务 510
11.1.4 压缩盘古行业 NLP 大模型 511
11.1.5 部署盘古行业 NLP 大模型 512
11.1.6 调用盘古行业 NLP 大模型 514
11.1.6.1 使用“体验中心”调用盘古行业 NLP 大模型 514
11.1.6.2 使用 API 调用盘古行业 NLP 大模型 515
12 开发 Deepseek 大模型 519
12.1 Deepseek 大模型介绍 519
12.2 使用数据工程构建 Deepseek 大模型数据集 520
12.3 部署 Deepseek 大模型 520

 版权所有 © 华为云计算技术有限公司

12.4 评测 Deepseek 大模型 521
12.4.1 创建 Deepseek 大模型评测数据集 521
12.4.2 创建 API 服务 524
12.4.3 创建 Deepseek 大模型评测任务 529
12.4.4 查看 Deepseek 大模型评测报告 535
12.4.5 管理 Deepseek 大模型评测任务 537
12.5 调用 Deepseek 大模型 538
13 开发三方大模型 540
13.1 使用数据工程构建三方大模型数据集 540
13.2 部署三方大模型 541
13.2.1 模型部署形态和部署说明 541
13.2.2 创建三方大模型部署任务 542
13.2.3 查看三方大模型部署任务详情 543
13.2.4 管理三方大模型部署任务 544
13.3 评测三方大模型 548
13.3.1 创建三方大模型评测数据集 548
13.3.2 创建 API 服务 550
13.3.3 创建三方大模型评测任务 555
13.3.4 查看三方大模型评测报告 561
13.3.5 管理三方大模型评测任务 563
13.4 调用三方大模型 564
13.4.1 使用 API 调用三方大模型 564
13.5 三方大模型训推高可用性 568
13.5.1 训练日志失败分析 568
14 开发盘古多模态图像问答大模型 570
14.1 使用数据工程构建图像问答大模型数据集 570
14.2 训练图像问答大模型 573
14.2.1 图像问答大模型训练流程与选择建议 573
14.2.2 创建图像问答大模型训练任务 575
14.2.3 查看图像问答大模型训练状态与指标 582
14.2.4 发布训练后的多模态大模型 583
14.2.5 管理多模态大模型训练任务 584
14.2.6 多模态大模型训练常见报错与解决方案 584
14.3 部署图像问答大模型 585
14.3.1 创建多模态大模型部署任务 585
14.3.2 查看多模态大模型部署任务详情 586
14.3.3 管理多模态大模型部署任务 587
14.4 评测图像问答大模型 588
14.4.1 创建多模态大模型评测数据集 588
14.4.2 创建多模态大模型评测任务 589
14.4.3 查看多模态大模型评测报告 592
14.4.4 管理多模态大模型评测任务 594

 版权所有 © 华为云计算技术有限公司

14.5 调用图像问答大模型 595
14.5.1 使用“体验中心”调用多模态大模型 595
14.5.2 使用 API 调用多模态大模型 596
15 开发图像搜索模型 599
15.1 图像搜索模型介绍 599
15.2 使用数据工程构建图像搜索模型数据集 599
15.3 训练图像搜索模型 600
15.3.1 创建图像搜索模型训练任务 600
15.3.2 查看图像搜索模型训练状态与指标 601
15.3.3 发布训练后的图像搜索模型 603
15.3.4 管理图像搜索模型训练任务 603
15.4 部署图像搜索模型 604
15.4.1 创建图像搜索模型部署任务 604
15.4.2 查看图像搜索模型部署任务详情 605
15.4.3 管理图像搜索模型部署任务 605
15.5 调用图像搜索模型 606
15.5.1 使用 API 调用图像搜索模型 607
16 开发提示词工程 609
16.1 什么是提示词工程 609
16.2 获取提示词模板 611
16.3 撰写提示词 611
16.3.1 创建提示词工程 611
16.3.2 撰写提示词 611
16.3.3 预览提示词效果 613
16.4 横向比较提示词效果 614
16.4.1 设置候选提示词 614
16.4.2 横向比较提示词效果 615
16.5 批量评估提示词效果 616
16.5.1 创建提示词评估数据集 616
16.5.2 创建提示词评估任务 618
16.5.3 查看提示词评估结果 619
16.6 发布提示词 620
17 开发盘古大模型垂域应用 621
17.1 工业应用编排介绍 621
17.2 编排工业应用 621
17.2.1 创建应用组件 621
17.2.1.1 预置组件功能介绍 621
17.2.1.2 创建脚本组件 627
17.2.1.3 创建算法包组件 632
17.2.1.4 部署算法包组件 638
17.2.2 创建静态应用 641

 版权所有 © 华为云计算技术有限公司

17.2.3 部署静态应用 642
17.2.4 调用静态应用 644
17.2.5 工业应用编排实践 645
17.2.5.1 示例：使用内置组件创建容器数据存取应用 645
18 开发 Agent 应用 649
18.1 Agent 开发平台介绍 649
18.1.1 平台简介 649
18.1.2 使用流程 650
18.2 使用百宝箱中预置的 Agent 应用 650
18.3 快速搭建 Agent 应用 652
18.4 快速搭建工作流 659
18.5 单智能体应用开发 681
18.5.1 基础配置 681
18.5.1.1 配置提示词 681
18.5.1.2 配置智能体调度模式 684
18.5.2 为应用添加技能 685
18.5.2.1 配置插件 685
18.5.2.2 配置工作流 686
18.5.2.3 配置知识库 688
18.5.2.4 配置 MCP 服务 691
18.5.3 提升应用对话体验 694
18.5.4 调试与发布应用 695
18.5.4.1 调试应用 695
18.5.4.2 发布应用为 API 服务 698
18.5.5 使用 API 调用应用 700
18.6 多智能体应用开发（Beta） 703
18.6.1 多智能体应用介绍 703
18.6.2 创建多智能体应用 704
18.6.3 调试与发布多智能体应用 707
18.6.4 使用 API 调用多智能体应用 707
18.7 工作流应用开发 709
18.7.1 工作流介绍 709
18.7.2 对话型工作流和任务型工作流 710
18.7.3 创建工作流 711
18.7.4 配置单智能体应用工作流 714
18.7.5 配置多智能体应用工作流 715
18.7.6 配置聊天记忆 718
18.7.7 调试与发布工作流 720
18.7.8 通过 API 调用工作流 721
18.7.9 工作流节点配置参考 724
18.7.9.1 开始和结束节点 724
18.7.9.2 大模型节点 727

 版权所有 © 华为云计算技术有限公司

18.7.9.3 知识检索节点 729
18.7.9.4 意图识别节点 731
18.7.9.5 插件节点 736
18.7.9.6 判断节点 741
18.7.9.7 代码节点 744
18.7.9.8 消息节点 746
18.7.9.9 提问器节点 747
18.7.9.10 循环节点 751
18.7.9.11 变量赋值节点 753
18.7.9.12 变量聚合节点 755
18.7.9.13 输入节点 757
18.7.9.14 工作流节点 758
18.7.9.15 MCP 服务节点 759
18.7.9.16 Agent 节点配置说明 761
18.7.10 管理工作流 763
18.8 管理 Agent 平台插件 763
18.8.1 插件管理 763
18.8.1.1 插件介绍 764
18.8.1.2 创建插件 764
18.8.1.2.1 基于 API 创建一个插件 764
18.8.1.2.2 通过 JSON 文件导入插件 770
18.8.1.3 在单智能体应用中使用插件 770
18.8.1.4 在工作流中使用插件 771
18.8.1.5 导出插件 772
18.8.1.6 管理插件 773
18.8.2 知识库管理 774
18.8.2.1 知识库介绍 774
18.8.2.2 创建知识库及上传文档 776
18.8.2.3 知识库命中测试 777
18.8.2.4 在单智能体中使用知识库 778
18.8.2.5 在工作流中使用知识库 779
18.8.2.6 管理知识库 779
18.8.3 MCP 服务管理 780
18.8.3.1 MCP 服务介绍 780
18.8.3.2 创建 MCP 服务 780
18.8.3.3 在单智能体应用中使用 MCP 服务 786
18.8.3.4 在工作流中使用 MCP 服务 787
18.8.3.5 管理 MCP 服务 790
18.8.3.6 开通 MCP 服务 792
18.9 Agent 开发常见报错与解决方案 793
18.10 管理 Agent 应用 796
19 管理空间资产 800

 版权所有 © 华为云计算技术有限公司

19.1 盘古大模型空间资产介绍 800
19.2 管理盘古数据资产 800
19.3 管理盘古模型资产 805
19.4 管理盘古组件资产 807
20 管理资源池 809
20.1 创建边缘资源池 809

 版权所有 © 华为云计算技术有限公司

1 盘古大模型服务使用流程

盘古大模型服务简介

盘古大模型服务包含盘古基础大模型、盘古行业大模型与ModelArts Studio大模型开发平台。凭借“模型+开发平台”的产品组合，助力各行业客户实现一站式的大模型开发与应用能力。

ModelArts Studio大模型开发平台是盘古大模型服务推出的包括数据管理、模型训练、模型部署于一体的综合平台，专为开发和应用大模型而设计，旨在为开发者提供简单、高效的大模型开发和部署方式。平台配备数据工程、模型开发、应用开发三大工具链，帮助开发者充分利用盘古大模型的功能。通过该平台，企业可根据需求选择合适的盘古NLP大模型、CV大模型、预测大模型、科学计算大模型、专业大模型等服务，便捷地构建自己的模型和应用。

- **数据工程工具链**：数据是大模型训练的核心基础。数据工程工具链作为平台的重要组成部分，具备数据获取、数据加工和数据发布等功能，确保数据的高质量与一致性。工具链能够高效收集并处理各种格式的数据，满足不同训练任务的需求，并提供强大的数据存储和管理能力，为大模型训练提供坚实的数据支持。
- **模型开发工具链**：模型开发工具链是盘古大模型服务的核心组件，提供从模型创建到部署的一站式解决方案，涵盖模型训练、压缩、部署、评测、调用等功能，保障模型的高效应用。
- **应用开发工具链**：应用开发工具链是盘古大模型平台的重要模块，支持提示词工程、Agent开发，显著加速大模型应用的开发流程，帮助企业快速应对复杂业务需求。

预置模型使用流程

ModelArts Studio大模型开发平台提供了不同类型的预置模型，包括NLP大模型和科学计算大模型。用户可将**预置模型**部署为**预置服务**，用于后续的调用操作。

其中，**NLP预置模型**使用流程见**图1-1**、**表1-1**，**科学计算预置模型**使用流程见**图1-2**、**表1-2**。

版权所有 © 华为云计算技术有限公司

图 1-1 NLP 预置模型使用流程图

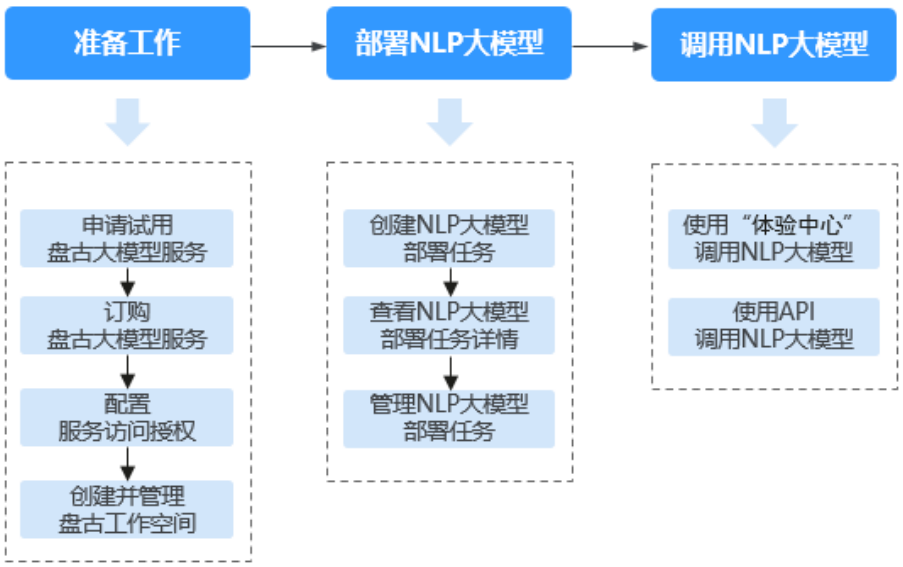


表 1-1 NLP 预置模型使用流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常地存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理工作空间
部署NLP大模型	创建NLP大模型部署任务	部署后的模型可用于后续调用操作。	创建NLP大模型部署任务
	查看NLP大模型部署任务详情	查看部署任务的详情，包括部署的模型基本信息、任务日志等。	查看NLP大模型部署任务详情
	管理NLP大模型部署任务	可对部署任务执行描述、删除等操作。	管理NLP大模型部署任务
调用NLP大模型	使用“体验中心”调用NLP大模型	使用该功能调用部署后的预置服务进行文本对话，支持设置人设和参数等。	使用“体验中心”调用NLP大模型
	使用API调用NLP大模型	可调用API接口与NLP预置服务进行文本对话。	使用API调用NLP大模型

版权所有 © 华为云计算技术有限公司

图 1-2 科学计算预置模型使用流程图

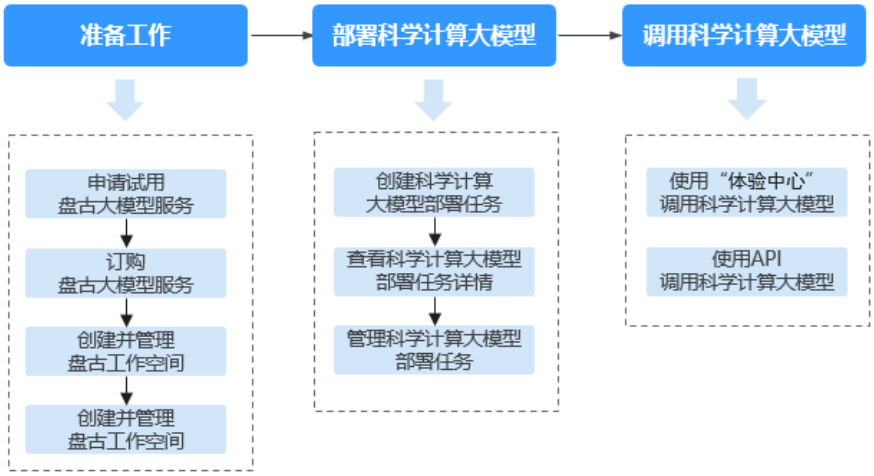


表 1-2 科学计算预置模型使用流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常地存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理工作空间
部署科学计算大模型	创建科学计算大模型部署任务	部署后的模型可用于后续调用操作。	创建科学计算大模型部署任务
	查看科学计算大模型部署任务详情	查看部署任务的详情，包括部署的模型基本信息、任务日志等。	查看科学计算大模型部署任务详情
	管理科学计算大模型部署任务	可对部署任务执行描述、删除等操作。	管理科学计算大模型部署任务
调用科学计算大模型	使用“体验中心”调用科学计算大模型	使用该功能调用部署后的预置服务对区域海洋要素等场景进行预测。	使用“体验中心”调用科学计算大模型
	使用API调用科学计算大模型	可调用科学计算API接口对区域海洋要素等场景进行预测。	使用API调用科学计算大模型

版权所有 © 华为云计算技术有限公司

数据工程使用流程

ModelArts Studio大模型开发平台提供了数据工程能力，帮助用户构造高质量的数据集，助力模型进行更好地预测和决策。

数据工程使用流程见图1-3、表1-3。

图 1-3 数据工程使用流程图

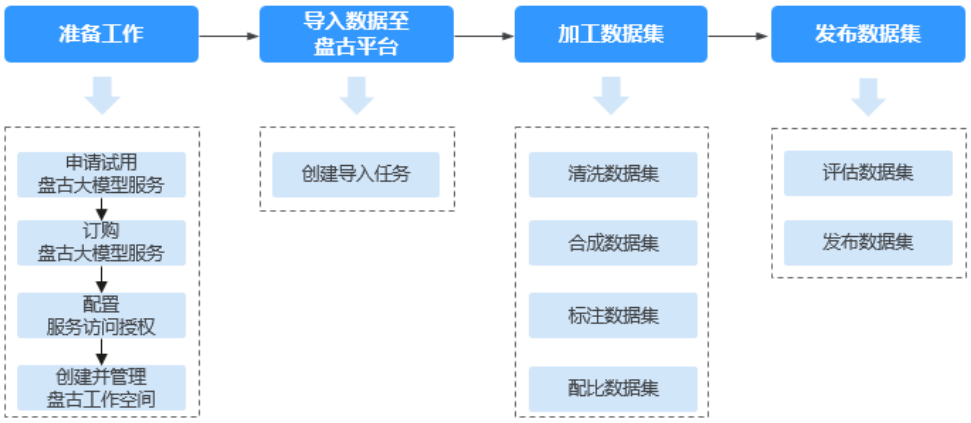


表 1-3 数据工程使用流程表

流程	子流程	说明
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。
加工数据集	加工数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。
	合成数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。

版权所有 © 华为云计算技术有限公司

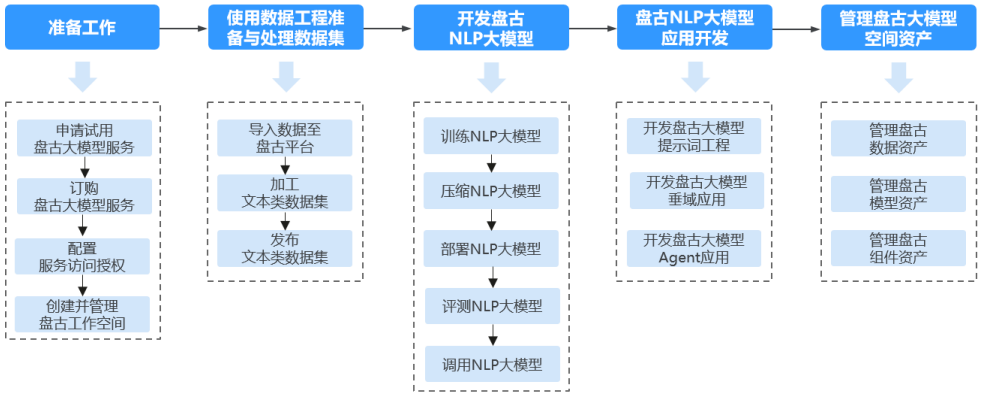
流程	子流程	说明
	标注数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。
	配比数据集	数据配比是将多个数据集按特定比例组合并生成为“加工数据集”的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。
发布数据集	评估数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。
	发布数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。 • 标准格式：平台默认的格式。该格式的数据集可发布为资产，但不可应用于盘古大模型的开发中。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。

NLP 大模型开发流程

ModelArts Studio大模型开发平台提供了NLP大模型的全流程开发支持，涵盖了从数据处理到模型训练、压缩、部署、评测、调用的各个环节。

NLP大模型开发流程见**图1-4**、**表1-4**。

图 1-4 NLP 大模型开发流程图



 版权所有 © 华为云计算技术有限公司

表 1-4 NLP 大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	**申请试用ModelArts Studio大模型开发平台**
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	**订购盘古大模型服务**
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	**配置服务访问授权**
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	**创建并管理工作空间**
使用数据工程构建NLP大模型数据集	导入数据至盘古平台	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	**导入数据至盘古平台**
	加工文本类数据集	对文本类数据集进行加工，包括加工、合成、标注、配比操作。	**加工文本类数据集**
	发布文本类数据集	对文本类数据集进行发布，包括评估、发布操作。	**发布文本类数据集**
开发盘古NLP大模型	训练NLP大模型	进行模型的训练，如预训练、微调、强化学习等训练方式。	**训练NLP大模型**
	压缩NLP大模型	通过模型压缩可以降低推理显存占用，节省推理资源提高推理性能。	**压缩NLP大模型**
	部署NLP大模型	将模型部署用于后续模型的调用操作。	**部署NLP大模型**
	评测NLP大模型	评测NLP大模型的效果。	**评测NLP大模型**
	调用NLP大模型	支持“体验中心”功能与API两种方式调用大模型。	**调用NLP大模型**
盘古NLP大模型应用开发	开发盘古大模型提示词工程	辅助用户进行提示词撰写、比较和评估等操作，并对提示词进行保存和管理。	**开发提示词工程**
	开发盘古大模型Agent应用	基于NLP大模型，致力打造智能时代集开发、调测和运行为一体的AI应用平台。无论开发者是否拥有大模型应用的编程经验，都可以通过Agent平台快速创建各种类型的智能体。	**开发Agent应用**

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	开发盘古大模型垂域应用	工业应用编排功能是一个基于脚本组件和算法包组件的编排工具。其中，组件是用来构成应用编排的最小业务单元，即将封装好的业务功能作为应用编排中的一个节点，实现特定的业务功能。	开发盘古大模型垂域应用
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产
	管理盘古组件资产	管理预置或创建的组件。	管理盘古组件资产

CV 大模型开发流程

ModelArts Studio大模型开发平台提供了CV大模型的全流程开发支持，涵盖了从数据处理到模型训练、部署环节。

CV大模型开发流程见图1-5、表1-5。

图 1-5 CV 大模型开发流程图

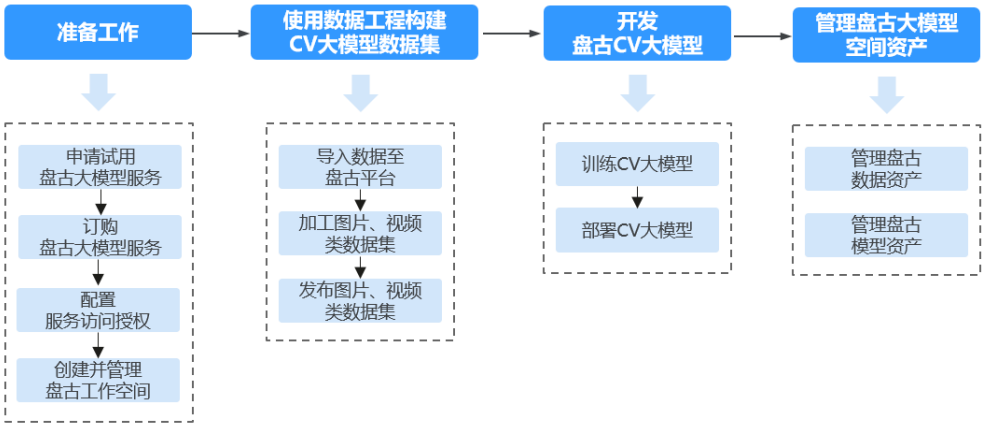


表 1-5 CV 大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理工作空间
使用数据工程构建CV大模型数据集	导入数据至盘古平台	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
	加工图片、视频类数据集	对图片、视频类数据集进行加工，包括加工、标注、配比操作。	加工图片类数据集、加工视频类数据集
	发布图片、视频类数据集	对图片、视频类数据集进行发布，包括评估、发布操作。	发布图片类数据集、发布视频类数据集
开发盘古CV大模型	训练CV大模型	进行模型的训练，如微调等训练方式。	训练CV大模型
	部署CV大模型	模型训练完成后，执行模型的部署操作。	部署CV大模型
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产

预测大模型开发流程

ModelArts Studio大模型开发平台提供了预测大模型的全流程开发支持，涵盖了从数据处理到模型训练、部署环节。

预测大模型开发流程见图1-6、表1-6。

版权所有 © 华为云计算技术有限公司

图 1-6 预测大模型开发流程图

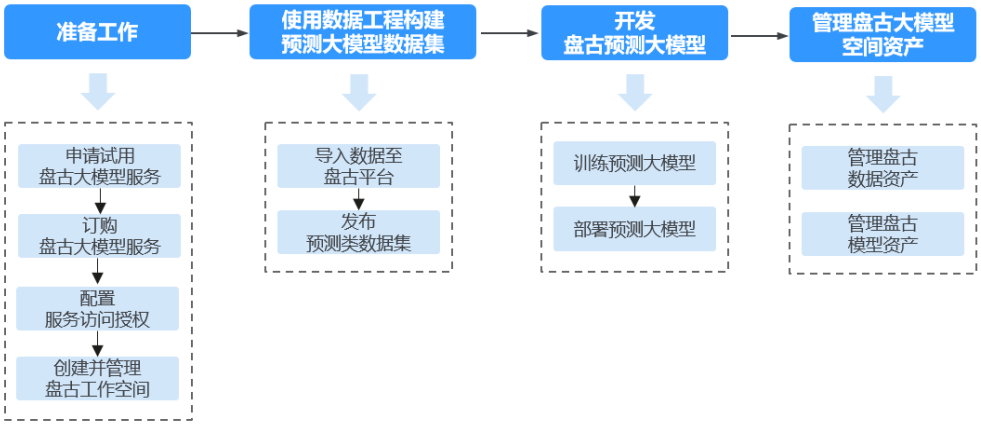


表 1-6 预测大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理工作空间
使用数据工程构建预测大模型数据集	导入数据至盘古平台	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
	发布预测类数据集	对预测类数据集进行发布操作。	发布预测类数据集
开发盘古预测大模型	训练预测大模型	进行模型的训练，如微调训练方式。	训练预测大模型
	部署预测大模型	模型训练完成后，执行模型的部署操作。	部署预测大模型
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产

版权所有 © 华为云计算技术有限公司

科学计算大模型开发流程

ModelArts Studio大模型开发平台提供了科学计算大模型的全流程开发支持，涵盖了从数据处理到模型训练、部署、调用的各个环节。

科学计算大模型开发流程见图1-7、表1-7。

图 1-7 科学计算大模型开发流程图

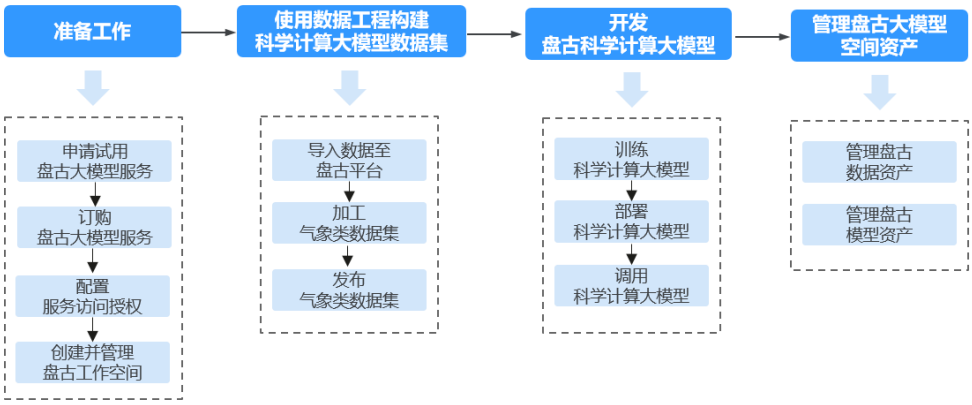


表 1-7 科学计算大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理工作空间
使用数据工程构建科学计算大模型数据集	导入数据至盘古平台	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
	加工气象类数据集	对气象类数据集进行加工操作。	加工气象类数据集
	发布气象类数据集	对气象类数据集进行发布操作。	发布气象类数据集
开发盘古科学计算大模型	训练科学计算大模型	进行模型的训练，如预训练、微调等训练方式。	训练科学计算大模型

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	部署科学计算大模型	将模型部署用于后续模型的调用操作。	部署科学计算大模型
	调用科学计算大模型	支持“体验中心”功能与API两种方式调用大模型。	调用科学计算大模型
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产

专业大模型开发流程

ModelArts Studio大模型开发平台提供了专业大模型的部署功能。

专业大模型开发流程见图1-8、表1-8。

图 1-8 专业大模型开发流程图

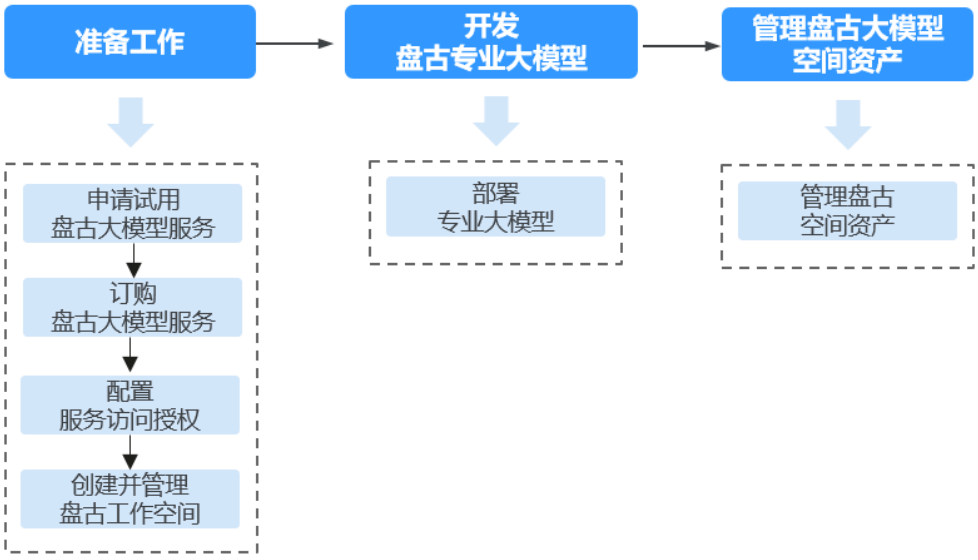


表 1-8 专业大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用ModelArts Studio大模型开发平台
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	**配置服务访问授权**
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	**创建并管理工作空间**
开发盘古专业大模型	部署专业大模型	支持专业大模型的部署操作。	**部署向量&重排大模型**
管理盘古大模型空间资产	管理盘古模型资产	管理预置的专业大模型。	**管理盘古模型资产**

版权所有 © 华为云计算技术有限公司

2 准备工作

申请试用ModelArts Studio大模型开发平台
订购盘古大模型服务
配置服务访问授权
创建并管理工作空间

2.1 申请试用 ModelArts Studio 大模型开发平台

盘古大模型服务为用户提供了服务试用，需提交试用申请。

试用申请步骤如下：

1. 登录**ModelArts Studio大模型开发平台**。
2. 单击“试用咨询”，进入华为云售前咨询页面。

图 2-1 申请试用



3. 填写姓名、联系电话等用户信息，单击“提交申请”进行表单预约。

2.2 订购盘古大模型服务

订购说明

ModelArts Studio大模型开发平台支持订购**模型资产、数据资源、训练资源、推理资源**，支持模型资产的包年/包月订购、资源的包年/包月和按需计费订购。

版权所有 © 华为云计算技术有限公司

- 模型资产：**模型资产**可用于模型开发、应用开发等模块。当前支持订购NLP大模型、多模态大模型、CV大模型、预测大模型、科学计算大模型，专业大模型。
- 数据资源：**数据通算单元**适用于数据加工，用于正则类算子加工、**数据智算单元**适用于数据加工，用于AI类算子加工，**数据托管单元**适用于数据工程，用于存储数据集。
- 训练资源：**训练单元**可用于所有大模型的模型训练、模型压缩功能。
- 推理资源：**推理单元**可用于NLP、多模态大模型和BI专业大模型在基础平台和Agent开发的推理服务场景， **模型实例**可用于预测、科学计算大模型的模型推理功能。

订购场景

ModelArts Studio大模型开发平台提供**免费模型、付费模型**，订购相应的模型和资源后支持在平台使用**数据工程**功能，**模型开发**功能，**Agent**功能。

用户可以基于使用场景进行按需进行订购。

表 2-1 订购场景介绍

编号	场景	场景说明
1	数据场景	仅订购数据资源不订购模型，支持数据的加工、合成、标注、配比存储等场景。
2	推理场景	纯推理场景，用户可以使用免费模型或订购模型资产（可选）+推理资源，部署后进行推理。
3	训推场景	支持模型的训练和推理，订购模型资产（可选）+训练资源+推理资源，支持用户使用已订购的付费模型或免费模型进行训练，将训练后的模型部署后再进行推理，获得更好的效果。
4	Agent场景	订购模型资产（可选）+训练资源（可选）+推理资源，支持使用已订购的付费模型或免费模型部署后在Agent中使用。

场景1：数据场景，仅使用数据工程的能力

订购“数据通算单元”、“数据智算单元”用于数据的加工、合成、标注、配比存储等场景，订购数据托管单元用于数据存储（包括数据集，prompt模板等）。



 版权所有 © 华为云计算技术有限公司

场景2：推理场景，使用已订购的付费模型或模型广场的免费模型，进行模型部署推理操作

订购 “ModelArts Studio-推理单元（NLP、多模态、BI）”用于NLP大模型，多模态大模型和BI专业大模型在基础平台和Agent开发的免费和付费模型的部署推理。

订购“ModelArts Studio-推理单元（CV）”和“CV大模型”资产用于CV大模型的部署推理。

订购“ModelArts Studio-模型实例（预测）”和“预测大模型”模型资产用于预测大模型的部署推理。

订购“ModelArts Studio-模型实例（科学计算）”和“科学大模型”模型资产用于科学计算大模型的部署推理。





场景3：训推场景，订购模型资产+训练资源+推理资源，支持用户训练和部署模型

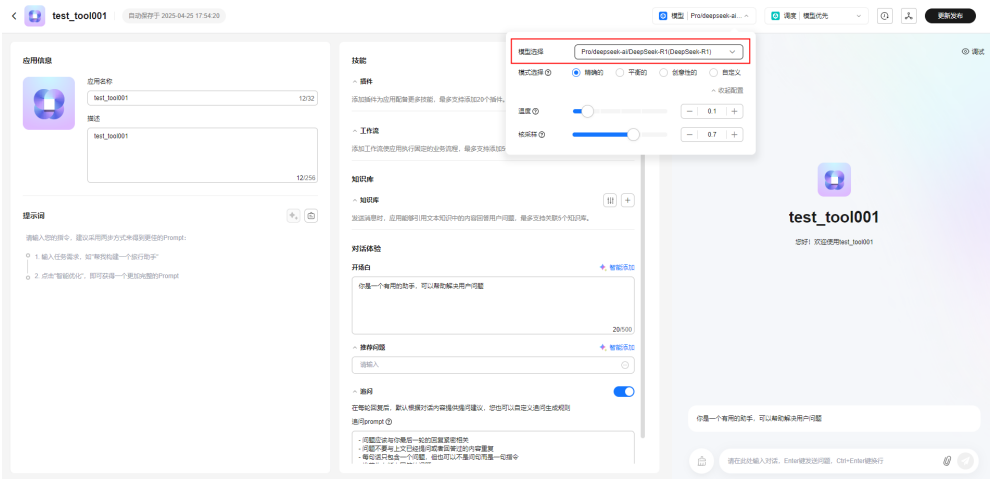
订购“ModelArts Studio-训练单元”，使用已订购的付费模型或模型广场的免费模型进行模型训练操作，训练完成后进行模型部署。



场景4：Agent场景，使用已经部署的模型用于Agent应用

版权所有 © 华为云计算技术有限公司

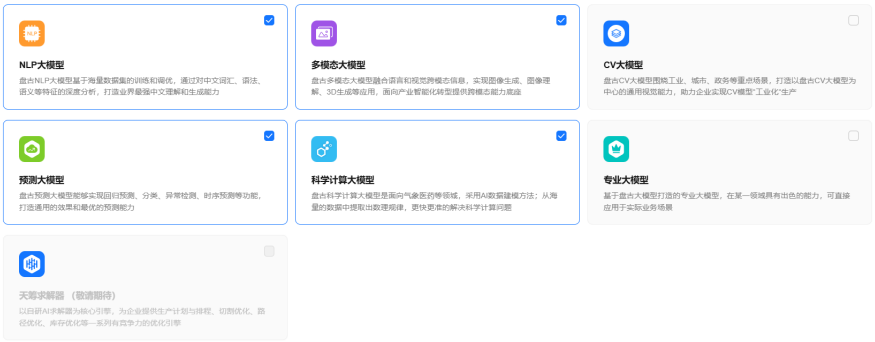
订购模型资产（可选）+训练资源（可选）+推理资源，支持使用已订购的付费模型或免费模型部署后在Agent中使用。



订购步骤

1. 使用主账户登录**ModelArts Studio大模型开发平台**，单击“立即订购”进入“订购”页面。
2. 在“开发场景”中勾选需要订购的大模型（可多选），页面将根据勾选情况适配具体的订购项。

图 2-2 选择开发场景



3. 在“模型资产”页面，参考**表2-2**完成模型资产的订购。

版权所有 © 华为云计算技术有限公司

表 2-2 模型资产订购说明

模型分类	模型订阅	模型资产	计费方式
NLP大模型	盘古-NLP-N1-基模型 盘古-NLP-N1-基础功能模型	Pangu-NLP-N1-32K Pangu-NLP-N1-128K Pangu-NLP-N1-Reasoner-128K 赠送如下模型： Pangu-EmbeddingRank-zh Pangu-RAG-N1-32k Pangu-SearchPlan-ZH_EN Pangu-AgentExpert-N1 Pangu-AgentExpert-N2	包年/包月（1~9个月，包年为1年）
	盘古-NLP-N2-基模型 盘古-NLP-N2-基础功能模型	Pangu-NLP-N2-4K Pangu-NLP-N2-32K Pangu-NLP-N2-128K Pangu-NLP-N2-256K Pangu-NLP-N2-Reasoner-32K Pangu-NLP-N2-Reasoner-128K Pangu-NLP-N2-Reasoner-TTC-32K Pangu-Medical-NLP-N2-32K Pangu-Medical-NLP-N2-reasoner-32K 赠送如下模型： Pangu-EmbeddingRank-zh Pangu-RAG-N1-32k Pangu-SearchPlan-ZH_EN Pangu-AgentExpert-N1 Pangu-AgentExpert-N2	包年/包月（1~9个月，包年为1年）

版权所有 © 华为云计算技术有限公司

模型分类	模型订阅	模型资产	计费方式
	盘古-NLP-N4-基模型 盘古-NLP-N4-基础功能模型	Pangu-NLP-N4-4K Pangu-NLP-N4-32K Pangu-NLP-N4-Reasoner-32K Pangu-NLP-N4-Reasoner-128K Pangu-NLP-N4-MoE-Reasoner-32K 赠送如下模型： Pangu-EmbeddingRank-zh Pangu-RAG-N1-32k Pangu-SearchPlan-ZH_EN Pangu-AgentExpert-N1 Pangu-AgentExpert-N2	包年/包月（1~9个月，包年为1年）
多模态大模型	盘古-多模态-万物理解模型 盘古-多模态-图生文模型	Pangu-MM-M2-Img2Txt-16K	包年/包月（1~9个月，包年为1年）
	盘古-多模态-万物生成模型	Pangu-MM-M2-AIGVideo Pangu-MM-M2-Text2Video	包年/包月（1~9个月，包年为1年）
CV大模型	盘古-CV-基础模型	Pangu-CV-ObjectDetection-S Pangu-CV-ObjectDetection-N Pangu-CV-ImageClassification Pangu-CV-SemanticSegmentation Pangu-CV-InstanceSegmentation-1.1.0	包年/包月（1~9个月，包年为1年）
预测大模型	盘古-预测-基础模型	Pangu-Predict-Anom-Table Pangu-Predict-Cla-TS Pangu-Predict-Cla-Table Pangu-Predict-Reg-TS Pangu-Predict-Reg-Table	包年/包月（1~11个月，包年为1年）

版权所有 © 华为云计算技术有限公司

模型分类	模型订阅	模型资产	计费方式
科学计算大模型	盘古-天气气象-基础版	Pangu-AI4S-Global-Ocean Pangu-AI4S-Global-Weather Pangu-AI4S-Global-Precip	包年/包月（1~9个月，包年为1年）
	盘古-天气气象-专业版	Pangu-AI4S-Global-Ocean Pangu-AI4S-Global-Weather Pangu-AI4S-Global-Precip Pangu-AI4S-Regional-Ocean Pangu-AI4S-Global-Ecology Pangu-AI4S-Global-Swell Pangu-AI4S-Regional-Pollution Pangu-AI4S-Regional-Wind Pangu-AI4S-Regional-Photovoltaic	包年/包月（1~9个月，包年为1年）

说明

NLP大模型支持模型按参数兼容订购，具体兼容订购规则如下：

1. 订购了盘古N2模型后可以使用盘古N1所有模型，用户无需再单独订购N1系列模型。
2. 订购了盘古N4模型后可以使用盘古N1、N2所有模型，用户无需再单独订购N1、N2系列模型。

资源订购说明

资源名称	订购项	适用场景	计费方式
数据资源	ModelArts Studio-数据托管单元	用于数据存储（包括数据集，prompt模板等）。	包年/包月
	ModelArts Studio-数据通算单元	适用于数据加工，用于正则类算子加工。 不同数据加工算子所需数据资源类型详见**数据集加工算子介绍**。	按需（时长）计费、包年/包月
	ModelArts Studio-数据智算单元	适用于数据加工，用于AI类算子加工。 不同数据加工算子所需数据资源类型详见**数据集加工算子介绍**。	按需（时长）计费、包年/包月

版权所有 © 华为云计算技术有限公司

资源名称	订购项	适用场景	计费方式
训练资源	ModelArts Studio-训练单元	用于所有模型的模型训练、模型压缩。	按需（时长）计费、包年/包月
推理资源	ModelArts Studio-推理单元（NLP、多模态）	适用于NLP大模型、多模态大模型在基础平台和Agent开发的推理服务场景。	包年/包月
	ModelArts Studio-推理单元（CV、遥感）	适用于CV大模型在基础平台的推理服务场景。	包年/包月
	ModelArts Studio-模型实例（预测）	适用于预测大模型在基础平台的推理服务场景。	包年/包月
	ModelArts Studio-模型实例（科学计算）	适用于气象大模型在基础平台的推理服务场景。	包年/包月

续订模型资产

ModelArts Studio大模型开发平台支持以**包年/包月**方式续订模型资产，即在当前订购的模型资产基础上延长使用时间。

续订模型资产的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“模型订购”页签，在订阅模型列表单击操作列“续订”。
3. 在“续费管理”页面根据提示完成模型资产的续费操作。

退订模型资产

退订模型资产的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“模型订购”页签，在订阅模型列表单击操作列“退订”。

说明

模型资产退订后不影响运行中的模型训练、压缩、评测、部署等任务，但退订之后将无法再选择该模型创建任务，请谨慎操作。

退订属于高危操作，在退订模型资产前，请确保您已保存所有必要的数据和进度，以避免不必要的损失。

增购模型资产

ModelArts Studio大模型开发平台支持增购模型资产，即在当前模型资产基础上订购新的模型资产，增购完成后支持使用多个模型资产。

版权所有 © 华为云计算技术有限公司

增购模型资产的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击右上角“新增订购”。
3. 在“订购”页面，参考**订购步骤**完成新的模型资产增购。

续订资源

ModelArts Studio大模型开发平台支持以**包年/包月**方式续订数据资源、训练资源、推理资源，即在当前资源基础上延长使用时间。

续订资源的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“资源订购”页签，在资源列表单击操作列“续订”。
3. 在“续费管理”页面根据提示完成资源的续费操作。

说明

按需计费方式暂不支持续订。

退订资源

退订资源的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“资源订购”页签，在资源列表单击操作列“退订”。

说明

退订后会停止运行中的服务，且之后将无法再选择该资源使用。

退订属于高危操作，请确保您已保存所有必要的数据和进度，以避免不必要的损失。

扩缩容资源

ModelArts Studio大模型开发平台支持数据资源、训练资源、推理资源的扩缩容，即在当前资源的基础上扩充或缩小对应的资源。

资源扩缩容的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“资源订购”页签，在资源列表单击操作列“扩缩容”。
3. 在“扩缩容”页面完成当前资源的扩缩容操作，平台将根据扩缩容前后的规格差异支付或退还费用差价。

说明

缩容可能会影响进行中的任务以及后续任务的创建，缩容前，请先确认需要缩容的资源已释放。

版权所有 © 华为云计算技术有限公司

2.3 配置服务访问授权

配置 OBS 访问授权

ModelArts Studio大模型开发平台使用对象存储服务（Object Storage Service，简称OBS）进行数据存储，实现安全、高可靠和低成本的存储需求。因此，为了能够顺利进行存储数据、训练模型等操作，需要用户配置访问OBS服务的权限。

配置OBS访问授权步骤如下：

1. 登录ModelArts Studio大模型开发平台首页。
2. 配置OBS访问授权。
   - 方式1：在首页顶部单击“此处”，在弹窗中选择授权项，并单击“确认授权”。

     图 2-3 配置 OBS 访问授权

   - 方式2：单击首页右上角“设置”，在“授权管理”页签，单击“一键授权”。

2.4 创建并管理工作空间

2.4.1 工作空间介绍

工作空间功能旨在为用户提供灵活、高效的资产管理与协作方式。平台支持用户根据业务需求或团队结构，自定义创建独立的工作空间。

每个工作空间在资产层面完全隔离，确保资产的安全性和操作的独立性，有效避免交叉干扰或权限错配带来的风险。用户可以结合实际使用场景，如不同的项目管理、部门运营或特定的研发需求，划分出多个工作空间，实现资产的精细化管理与有序调配，帮助用户高效地规划和分配任务，使团队协作更加高效。

此外，平台配备了完善的角色权限体系，覆盖超级管理员、管理员、模型开发工程师等多种角色。通过灵活的权限设置，每位用户能够在其对应的权限范围内安全高效地操作平台功能，从而最大程度保障数据的安全性与工作效率。

2.4.2 创建并管理工作空间

约束限制

如果当前用户所在的用户组未取得**全部云服务管理员（Tenant Administrator）**授权，则无法创建工作空间。

版权所有 © 华为云计算技术有限公司

创建空间前，请参考**创建用户组**及**创建盘古子用户**，赋予当前用户组**全部云服务管理员**权限，或创建具有**全部云服务管理员**权限的用户组和对应用户。

创建工作空间

创建工作空间步骤如下：

1. 登录**ModelArts Studio大模型开发平台**，在“我的空间”模块，单击“创建空间”。

图 2-4 我的空间



2. 填写空间名称、描述，标签（可选），单击“确认”，完成空间的创建。

图 2-5 创建空间



版权所有 © 华为云计算技术有限公司

3. 单击创建好的空间，进入ModelArts Studio大模型开发平台。

   如果用户具备多个空间的访问权限，可在页面左上角单击 切换空间。

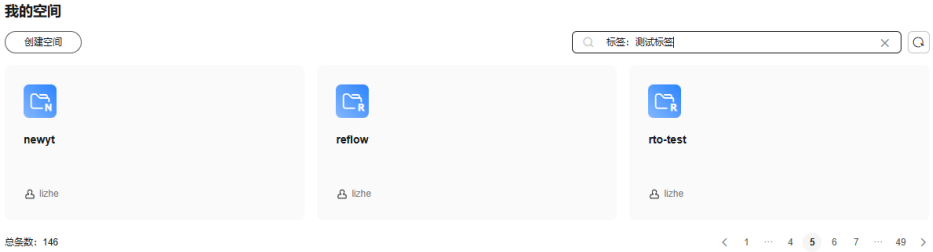
   图 2-6 切换空间



4. 按标签过滤工作空间。

   在ModelArts Studio大模型开发平台首页的“我的空间”模块，空间搜索支持按照标签过滤空间。

   图 2-7 按标签过滤工作空间



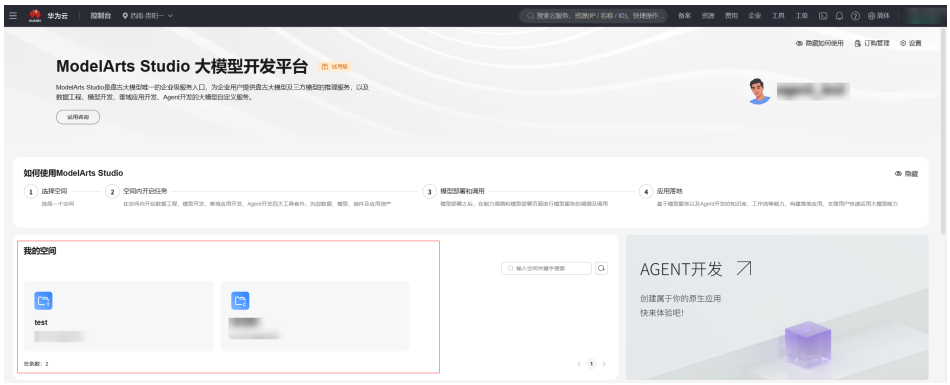
管理盘古工作空间

盘古工作空间支持用户查看当前空间详情，修改空间名称与描述，还可以对不需要的空间实现删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 2-8 我的空间



2. 单击左侧导航栏的“空间管理”，在“空间设置”页签可执行如下操作：
   - 修改当前空间名称与描述。
   - 可查看当前空间的创建时间。
   - 单击右上角“删除”，可删除当前空间。

   说明

   删除空间前请确保当前空间不再进行使用。

2.4.3 管理工作空间成员

如果您需要为企业员工设置不同的访问权限，以实现功能使用权限和资产的权限隔离，可以为不同员工配置相应的角色，以确保资产的安全和管理的高效性。

如果华为云账号已经能满足您的要求，不需要创建独立的IAM用户（子用户）进行权限管理，您可以**跳过本章节**，不影响您使用盘古的其他功能。

您可以使用统一身份认证服务（IAM）并结合ModelArts Studio大模型开发平台提供的“成员管理”功能实现子用户精细的权限管理。

创建用户组

管理员可以创建用户组，并给用户组授予策略或角色，然后将用户加入用户组，使得用户组中的用户获得相应的权限。

创建用户组的步骤如下：

1. 使用主账号登录**IAM服务控制台**。
2. 左侧导航栏中，选择“用户组”页签，单击右上方的“创建用户组”。

图 2-9 创建用户组



版权所有 © 华为云计算技术有限公司

3. 在“创建用户组”页面，输入“用户组名称”，单击“确定”，创建用户组。
4. 返回用户组列表，单击操作列的“授权”。

图 2-10 用户组授权



5. 参考表2-3，在搜索框中搜索授权项，为用户组设置权限，选择后单击“下一步”。

表 2-3 授权项

授权项	说明
Agent Operator	拥有该权限的用户可以切换角色到委托方账号中，访问被授权的服务。
Tenant Administrator	全部云服务管理员（除IAM管理权限）。
Security Administrator	统一身份认证服务（除切换角色外）所有权限。

图 2-11 添加用户组权限



6. 设置最小授权范围。

   根据授权项策略，系统会自动推荐授权范围方案。

   - 可以选择“所有资源”，即用户组内的IAM用户可以基于设置的授权项使用账号中所有的企业项目、区域项目、全局服务资源。
   - 可以选择“指定区域项目资源”，如指定“西南-贵阳一”区域，即用户组内的IAM用户仅可使用该区域项目中的资源。
   - 可以选择“全局服务资源”，即服务部署时不区分区域，访问全局级服务，不需要切换区域，全局服务不支持基于区域项目授权。如对象存储服务（OBS）、内容分发网络（CDN）等。

   选择完成后，单击“确定”。

版权所有 © 华为云计算技术有限公司

图 2-12 设置最小授权范围



7. 单击“完成”，完成用户组授权。

图 2-13 完成授权

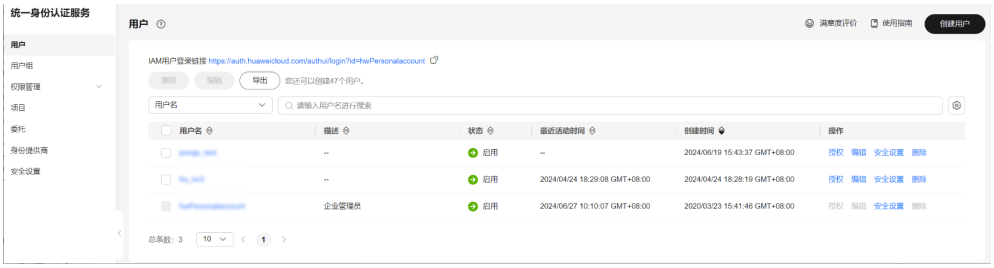


创建盘古子用户

创建盘古子用户步骤如下：

1. 使用主账号登录**IAM服务控制台**。
2. 左侧导航窗格中，选择“用户”页签，单击右上方的“创建用户”。

图 2-14 创建用户

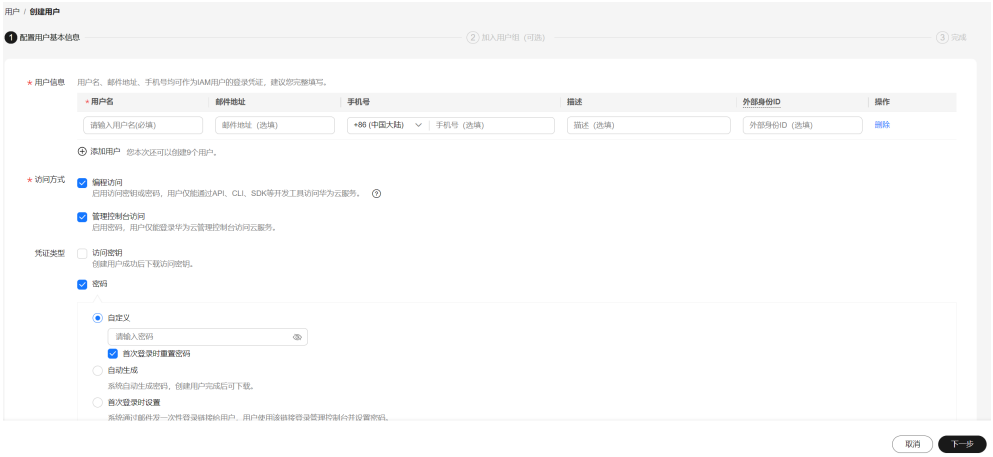


3. 配置用户基本信息，单击“下一步”。

   配置用户信息时，需要勾选“编程访问”，如果未勾选此项，会导致IAM用户无法使用盘古服务API、SDK。

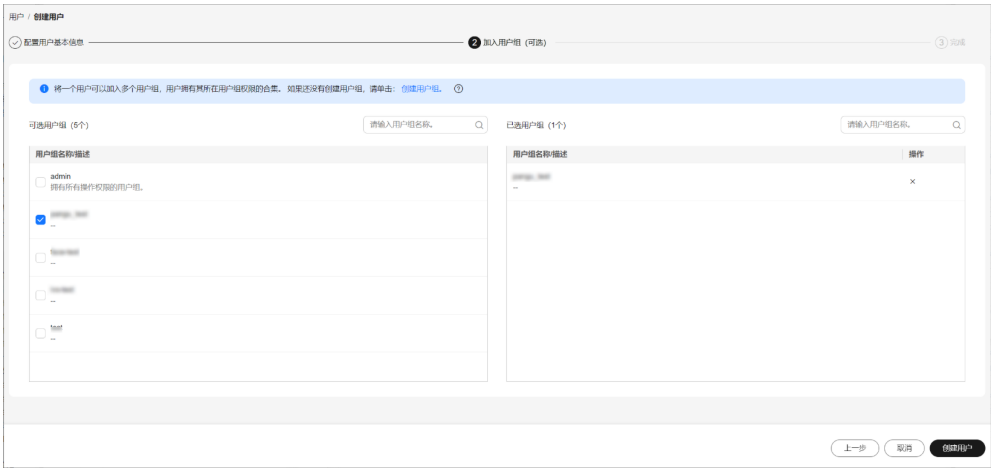
版权所有 © 华为云计算技术有限公司

图 2-15 配置用户基本信息



4. 将用户添加至**创建用户组**步骤中创建的用户组，单击“创建用户”，完成IAM用户的创建。

图 2-16 加入用户组



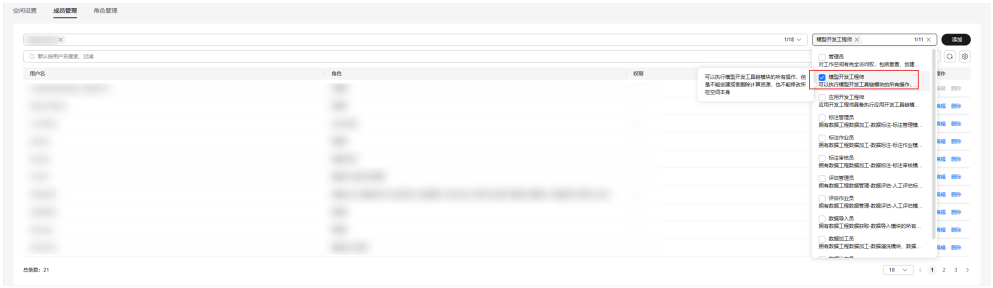
添加盘古子用户至工作空间

在添加盘古子用户至工作空间前，请先完成**创建盘古子用户**。

1. 登录ModelArts Studio大模型开发平台。
2. 进入需要添加子用户的空间，在空间内单击左侧导航栏“空间管理”，并进入“成员管理”页签。
3. 如图，以添加子用户为“模型开发工程师”角色为例。在搜索框中搜索子用户名称，在“请选择角色”选项栏中设置用户角色，设置完成后单击右侧“添加”，将该用户添加至本空间。

版权所有 © 华为云计算技术有限公司

图 2-17 添加成员为“模型开发工程师”角色

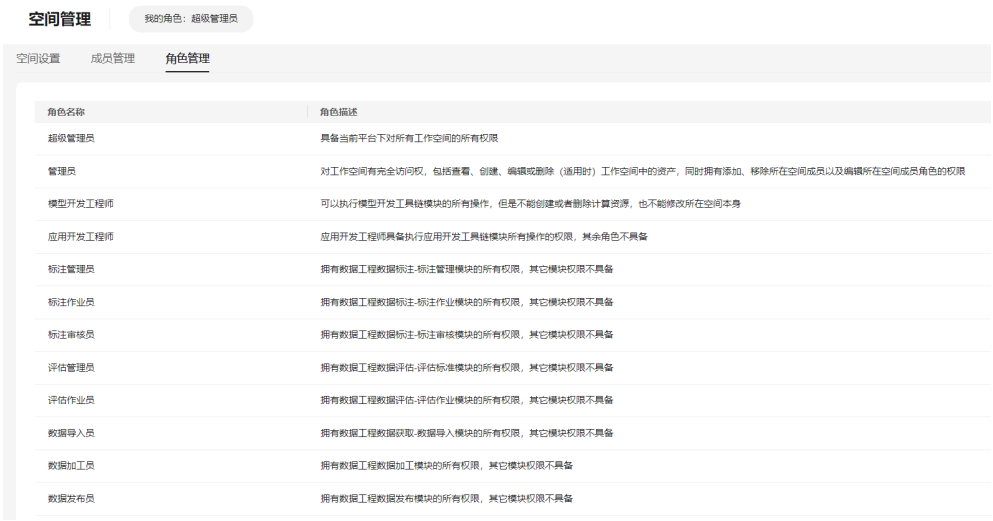


修改盘古子用户权限

当需要修改空间内某个子用户权限时可以按如下步骤操作：

1. 登录ModelArts Studio大模型开发平台。
2. 进入需要修改子用户权限的空间，在空间内单击左侧导航栏“空间管理”，在“角色管理”页签，可以查看各角色名称及其权限的描述。

图 2-18 角色管理



3. 单击进入“成员管理”页签。
4. 如图，以授权子用户“模型开发工程师”权限为例。单击用户列表操作栏的“编辑”，勾选需要赋予用户的角色，单击“确认”。

图 2-19 授权子用户“模型开发工程师”权限



移除盘古子用户

当需要删除空间内某个子用户时，可以按如下步骤操作：

版权所有 © 华为云计算技术有限公司

1. 登录ModelArts Studio大模型开发平台。
2. 进入需要删除子用户的空间，在空间内单击左侧导航栏“空间管理”，并进入“成员管理”页面。
3. 单击用户列表操作栏的“删除”。

图 2-20 成员管理



4. 单击“确定”进行二次确认，即可删除空间子用户。

图 2-21 删除操作二次确认



版权所有 © 华为云计算技术有限公司

3 在模型广场查看模型

ModelArts Studio平台模型广场页面提供一站式的模型资源库，为用户提供了多种模型选择，包括盘古系列模型、三方开源模型、行业领域模型等，您可以依据业务需求选择合适的模型执行训练推理任务。

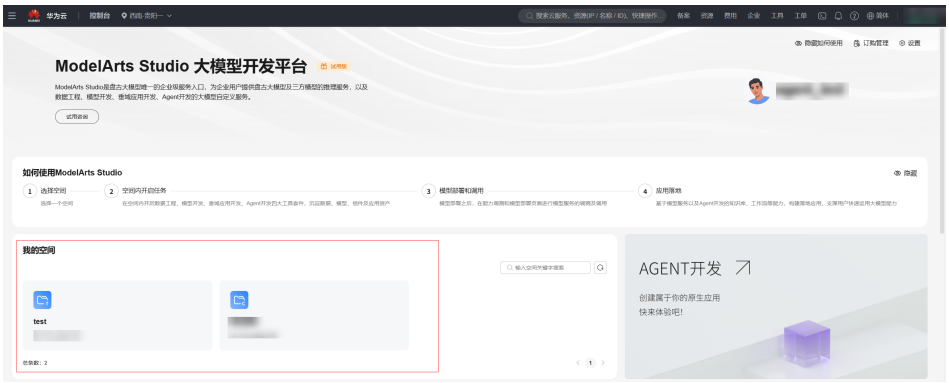
模型广场左侧区域是搜索与筛选区，支持客户勾选不同筛选条件，过滤出所需要的模型；并且支持按照关键字搜索模型。右侧区域是模型展示区，以卡片的形式展示模型内容，包括模型名称、模型描述、应用场景标签，支持单击“查看详情”，进入模型详情页，查看版本号、版本描述、版本详情等信息，支持单击“训练”或者“部署”，单击跳转到模型训练或模型部署界面，对该模型创建训练或部署任务。

平台支持的大模型清单请参见《产品介绍》“模型能力与规格 > 盘古NLP大模型”章节、“模型能力与规格 > 盘古CV大模型”章节、“模型能力与规格 > 盘古预测大模型”章节、“模型能力与规格 > 盘古科学计算大模型”章节、“模型能力与规格 > 盘古专业大模型”章节、“模型能力与规格 > 盘古行业大模型”章节、“模型能力与规格 > 三方大模型”章节、“模型能力与规格 > 盘古多模态大模型”章节。

访问模型广场

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击空间名称进入操作空间。

图 3-1 我的空间



2. 在左侧导航栏中选择“模型广场”。
3. 在“模型广场”页面左侧区域，可以按照“模型类型”、“模型系列”、“应用场景”、“支持功能”、“订阅类型”进行模型筛选。

版权所有 © 华为云计算技术有限公司

表 3-1 模型筛选说明

筛选项	说明
模型类型	包括大语言类型、向量&重排模型、搜索规划模型、多模态大模型、CV大模型、预测大模型、科学计算大模型、图像搜索等。
模型系列	包括盘古基础大模型、盘古行业大模型、盘古专业大模型、Deepseek、AIKits。
应用场景	包括长文本推理、逻辑推理、行业增强、物体检测、深度思考、文本生成、数据分析、意图分类、图像分类等。
支持功能	包括支持训练、支持部署。
订阅类型	包括免费和付费。

查看模型版本信息

在“模型广场”中，鼠标指向目标模型，单击“查看详情”，进入到模型详情界面，查看模型详情。

图 3-2 模型广场



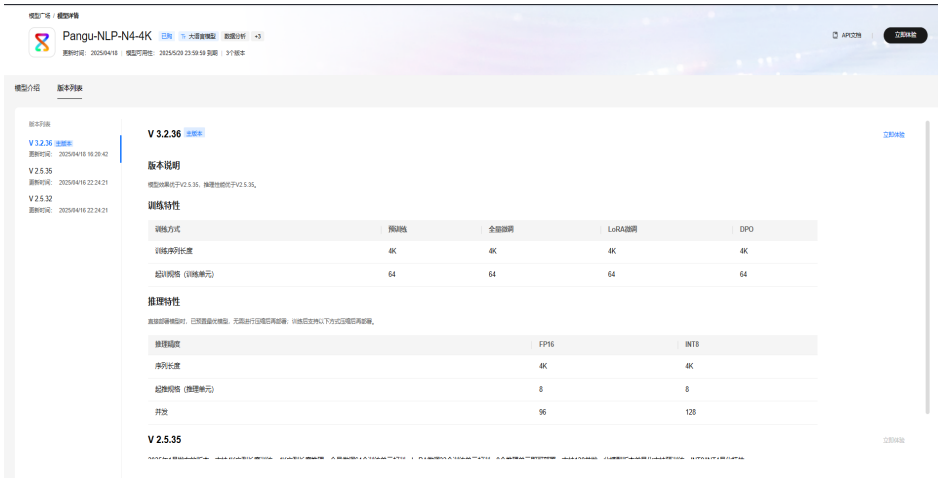
模型详情界面展示信息分为模型介绍以及版本列表页。模型介绍页主要展示对应模型的介绍内容、模型使用申明、平台服务协议。版本列表页展示当前模型下多个版本的特性介绍。

图 3-3 模型介绍



版权所有 © 华为云计算技术有限公司

图 3-4 版本列表



训练、部署模型

参考开发各大模型章节，模型训练和部署的操作步骤。

以NLP为例，可参考**训练NLP大模型**和**部署NLP大模型**完成模型的训练和部署。

体验模型功能

对于已有模型进行部署操作。平台提供预置部署服务，可直接单击“体验中心”体验模型功能，对于没有提供预置服务的模型，需要到“模型开发 > 模型部署”页面进行模型部署。

1. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
2. 在“创建部署”页面，选择模型来源、模型类型，完成部署参数设置，启动模型部署。

   部署完成后，模型广场界面对应模型的卡片会展示“立即体验”按钮，单击后跳转到“体验中心”体验模型功能。

图 3-5 立即体验



版权所有 © 华为云计算技术有限公司

图 3-6 体验模型功能



 版权所有 © 华为云计算技术有限公司

4 使用数据工程构建数据集

数据工程介绍
数据工程使用流程
数据集格式要求
导入数据至盘古平台
加工数据集
发布数据集
数据集格式转换
数据工程常见报错与解决方案

4.1 数据工程介绍

数据工程介绍

数据工程是ModelArts Studio大模型开发平台（下文简称“平台”）为用户提供的一站式数据处理与管理功能，旨在通过系统化的数据获取、加工、发布等过程，确保数据能够高效、准确地为大模型的训练提供支持，帮助用户高效管理和处理数据，提升数据质量和处理效率，为大模型开发提供坚实的数据基础。

数据工程包含的具体功能如下：

- **数据获取**：数据获取是数据工程的第一步，支持将不同来源和格式的数据导入平台，并生成“原始数据集”。
  - 支持的接入方式：通过OBS服务导入数据。
  - 支持的数据类型：文本、图片、视频、音频、气象、预测、其他。
  - 自定义格式：用户可以根据业务需求上传自定义格式的数据，提升数据获取的灵活性和可扩展性。

  通过这些功能，用户可以轻松将大量数据导入平台，为后续的数据加工和模型训练等操作做好准备。
- **数据加工**：平台提供了数据加工、数据合成、数据标注、数据配比的加工操作，旨在确保原始数据能够满足各种业务需求和模型训练的标准，生成“加工数据集”。

 版权所有 © 华为云计算技术有限公司

  - 数据加工：数据加工旨在通过使用数据集加工算子对数据进行预处理操作，针对不同类型的数据集，平台设计了专用的加工算子，以确保数据符合模型训练的标准和业务需求。
  - 数据合成：数据合成利用预置或自定义的数据指令对原始数据集进行处理，并根据设定的轮数生成新的数据。
  - 数据标注：数据标注旨在为无标签的数据集添加准确的标签，标注数据的质量直接影响模型的训练效果和精度。针对不同数据集平台支持人工标注与AI预标注两种形式。

    其中，图片Caption、视频Caption标注项支持AI预标注功能。
  - 数据配比：将多个数据集按照特定比例关系组合并为一个“加工数据集”的过程，确保数据的多样性、平衡性和代表性。

  通过数据加工操作，平台能够有效清理噪声数据、标准化数据格式，提升数据集的整体质量。
- **数据发布：**平台提供了数据发布操作，支持发布不同模态及不同格式的数据集，并生成“发布数据集”。
  - 数据发布：将单个数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

    支持发布的数据集格式为标准格式、盘古格式（适用于训练盘古大模型时）。目前，仅文本类和图片类数据集支持发布为“盘古格式”。
- **数据管理：**平台提供了数据集管理，数据评估操作，旨在管理各类型数据集以及通过数据质量评估确保数据满足大模型训练的多样性、平衡性和代表性需求，并促进数据的高效流通与应用。
  - 数据集：管理“原始数据集”、“加工数据集”、“发布数据集”等不同类别的数据集。
  - 数据评估：数据评估通过对数据集进行系统的质量检查，依据评估标准评估数据的多个维度，旨在发现潜在问题并加以解决。

在集成了数据获取、数据加工、数据发布功能外，平台还支持对原始数据集、加工数据集、发布数据集、数据合成指令进行一站式管理。在大规模数据集的构建过程中，ModelArts Studio大模型开发平台的数据工程功能为用户提供了极大的灵活性和高效性，确保了数据处理的各个环节都能紧密协作，快速响应不断变化的业务需求和技术要求。

平台支持的数据类型

ModelArts Studio大模型开发平台支持的数据类型见表4-1各类型数据格式详细要求请参考数据集格式要求。

表 4-1 平台支持的数据类型

数据类型	数据内容	支持的文件格式
文本类	文档	txt、mobi、epub、docx、pdf
	网页	html
	预训练文本	jsonl
	单轮问答	jsonl、csv

版权所有 © 华为云计算技术有限公司

数据类型	数据内容	支持的文件格式
	单轮问答（人设）	jsonl、csv
	多轮问答	jsonl
	多轮问答（人设）	jsonl
	问答排序	jsonl、csv
	偏好优化DPO	jsonl
	偏好优化DPO（人设）	jsonl
图片类	图片	• 图片格式支持：jpg、jpeg、png、bmp。支持图片和jsonl文件导入，jsonl为非必须文件类型。 • 压缩包格式：所有图片保存为tar包。图片格式支持：jpg、jpeg、png、bmp。支持图片和jsonl文件导入，jsonl为非必须文件类型。 • 标注文件格式：可选，当存在jsonl文件时，图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。注意：jsonl文件仅支持UTF-8编码。
	图片+Caption	• 图片格式支持：jpg、jpeg、png、bmp，所有图片需保存为tar包。 • Caption格式支持：jsonl，jsonl文件仅支持UTF-8编码。
	图片+QA对	• 图片格式支持：jpg、jpeg、png、bmp，所有图片需保存为tar包。 • QA对格式支持：jsonl，jsonl文件仅支持UTF-8编码。
	物体检测	• 图片格式支持：jpg、jpeg、png、bmp、tif、tiff • 标注格式支持：xml
	图像分类	• 图片格式支持：jpg、jpeg、png、bmp、tif、tiff • 标注格式支持：txt
	异常检测	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：txt

版权所有 © 华为云计算技术有限公司

数据类型	数据内容	支持的文件格式
	语义分割	● 图片+xml： – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：xml ● 原始图片+标注图片+json – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：标注图片+json ● 原始图片+标注图片+txt – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：标注图片+txt ● 原始图片+标注图片：原始/标注图片格式支持：jpg、jpeg、png、bmp ● 图片+png
	姿态估计	● 图片+json（一对一） – 图片格式支持：jpg、jpeg、png、bmp – 标注格式支持：json ● 图片+json（多对一） – 图片格式支持：jpg、jpeg、png、bmp – 标注格式支持：json
	实例分割	● 图片格式支持：jpg、jpeg、png、bmp ● 标注格式支持：xml
	变化检测	● 图片+xml – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：xml ● 原始图片+标注图片+json – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：png+json ● 原始图片+标注图片 – 图片格式支持：jpg、jpeg、png、bmp – 标注文件格式：png
	旋转检测	● 图片格式支持：jpg、jpeg、png、bmp ● 标注格式支持：xml
视频类	视频	mp4、avi
	视频+标注	● 视频格式支持：mp4、avi。 ● 标注文件格式：jsonl，jsonl文件仅支持UTF-8编码。

版权所有 © 华为云计算技术有限公司

数据类型	数据内容	支持的文件格式
	视频分类	文件格式支持视频+txt，视频格式支持：mp4、avi。标注文件格式：txt，视频与标注文件一一对应。
	事件检测	• 图片+json（一对一） – 视频格式支持：mp4、avi，每个视频时长大于128s，FPS>=10。 – 标注文件格式：json。 • 图片+json（多对一） – 视频格式支持：mp4、avi，每个视频时长大于128s，FPS>=10。 – 标注文件格式：json。
音频	音频	• 音频文件：支持mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 • 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。
气象类	气象数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2
预测类	时序分类	csv
	时序回归	csv
	结构化分类	csv
	结构化回归	csv
其他类	自定义	支持构建用户自定义场景下所需的数据集类型。

各类数据支持的操作

各类型数据支持的数据工程操作见表4-2。

表 4-2 各类数据支持的操作

数据类型	数据获取	数据加工	数据合成	数据标注	数据配比	数据评估	数据发布
文本类	√	√	√	√	√	√	√
图片类	√	√	-	√	√	√	√
视频类	√	√	-	√	√	√	√
音频类	√	√	-	√	-	-	√
气象类	√	√	-	-	-	-	√

版权所有 © 华为云计算技术有限公司

数据类型	数据获取	数据加工	数据合成	数据标注	数据配比	数据评估	数据发布
预测类	√	√	-	-	√	-	√
其他类	√	√	-	-	-	-	√

4.2 数据工程使用流程

高质量数据是推动大模型不断迭代和优化的根基，它的质量直接决定了模型的性能、泛化能力以及应用场景的适配性。只有通过系统化地准备和处理数据，才能提取出有价值的信息，从而更好地支持模型训练。因此，数据的获取、加工、合成、标注、配比、评估、发布等环节，成为数据开发中不可或缺的重要步骤。

数据工程操作流程见图4-1、表4-3。

图 4-1 数据集构建流程图

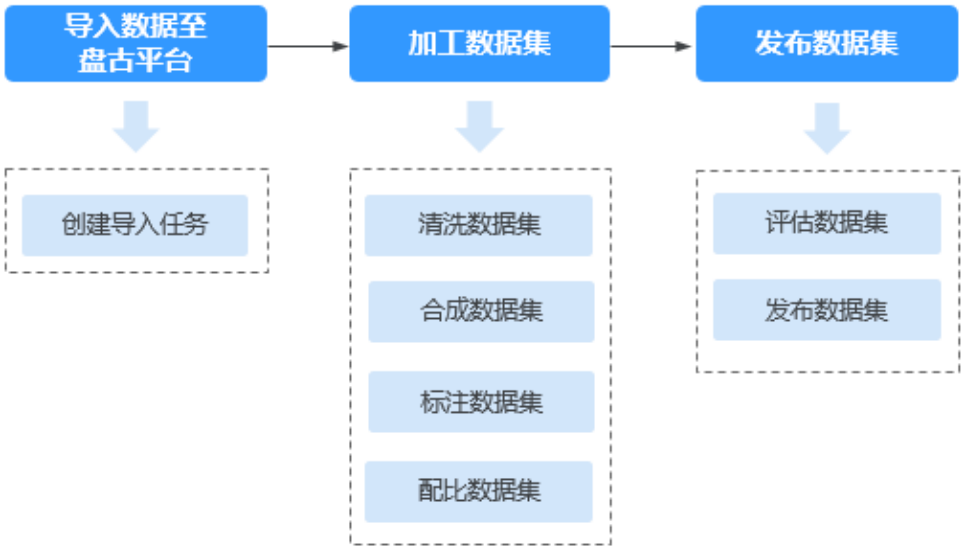


表 4-3 数据集构建流程表

流程	子流程	说明
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。
加工数据集	加工数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。
	合成数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。

版权所有 © 华为云计算技术有限公司

流程	子流程	说明
	标注数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。
	配比数据集	数据配比是将多个数据集按特定比例组合并生成为“加工数据集”的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。
发布数据集	评估数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。
	发布数据集	数据发布是将数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。 • 标准格式：平台默认的格式。该格式的数据集可发布为资产，但不可应用于盘古大模型的开发中。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。

4.3 数据集格式要求

4.3.1 文本类数据集格式要求

ModelArts Studio大模型开发平台支持创建文本类数据集，创建时可导入多种形式的数据，具体格式要求详见表4-4。

表 4-4 文本类数据集格式要求

文件内容	文件格式	文件要求
文档	txt、mobi、epub、docx、pdf	从OBS导入：单个文件大小不超过1GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。
网页	html	从OBS导入：单个文件大小不超过1GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
预训练文本	jsonl	• jsonl格式：text表示预训练所使用的文本数据，具体格式示例如下： {"text":"盘古大模型，是华为推出的盘古系列AI大模型，包括NLP大模型、多模态大模型、CV大模型、科学计算大模型、预测大模型。"} • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。
单轮问答	jsonl、csv	• jsonl盘古格式-非思维链：数据由问答对构成，context、target分别表示问题、答案，具体格式示例如下： {"context": ["你好，请介绍自己"], "target": "我是盘古大模型"} • jsonl盘古格式-思维链：数据由问答对构成，context、target分别表示问题、答案，并且target必须包含think标签对表示思考过程，具体格式示例如下： {"context": ["你好，请介绍自己"], "target": "<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>我是盘古大模型"} • csv盘古格式-非思维链：csv文件的第一列对应context，第二列对应target，具体格式示例如下： "你好，请介绍自己","我是盘古大模型" • csv盘古格式-思维链：csv文件的第一列对应context，第二列对应target，并且target必须包含think标签对，具体格式示例如下： "你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>我是盘古大模型" • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
单轮问答（人设）	jsonl、csv	• jsonl盘古格式-非思维链：system表示人设，context、target分别表示问题、答案。具体格式示例如下： {"system":"你是一个机智幽默问答助手","context":["你好，请介绍自己"],"target":"哈哈，你好呀，我是你的聪明助手。"} • jsonl盘古格式-思维链：system表示人设，context、target分别表示问题、答案， 并且target必须包含think标签对表示思考过程，具体格式示例如下： {"system":"你是一个机智幽默问答助手","context":["你好，请介绍自己"],"target":"<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手。"} • csv盘古格式-非思维链：csv文件的第一列对应system，第二三列分别对应context、target，具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","哈哈，你好呀，我是你的聪明助手。" • csv盘古格式-思维链：csv文件的第一列对应system，第二三列分别对应context、target，并且target必须包含think标签对表示思考过程，具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手。" • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。
多轮问答	jsonl	• jsonl盘古格式-非思维链：数组格式，由一轮或多轮问答对构成，其中context、target分别表示问题、答案，具体格式示例如下： [{"context":["你好"],"target":"你好，请问有什么可以帮助你的？"},{"context":["请介绍一下华为云的产品。"],"target":"华为云提供包括但不限于计算、存储、网络等产品服务。"}] • jsonl盘古格式-思维链：数组格式，由一轮或多轮问答对构成，其中context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下： [{"context":["你好"],"target":"<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>你好，请问有什么可以帮助你的？"},{"context":["请介绍一下华为云的产品。"],"target":"华为云提供包括但不限于计算、存储、网络等产品服务。"}] • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
多轮问答（人设）	jsonl	• jsonl盘古格式-非思维链：数组格式，由人设和一轮或多轮问答对构成。system表示人设，context、target分别表示问题、答案，具体格式示例如下： [{"system":"你是一位书籍推荐专家"},{"context":["你好"],"target":"嗨！你好，需要点什么帮助吗？"},{"context":["能给我推荐点书吗？"],"target":"当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"}] • jsonl盘古格式-思维链：数组格式，由人设一轮或多轮问答对构成。system表示人设，context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下： [{"system":"你是一位书籍推荐专家"},{"context":["你好"],"target":"<think>用户在打招呼，需要回复以及询问</think>嗨！你好，需要点什么帮助吗？"},{"context":["能给我推荐点书吗？"],"target":"<think>我需要以专家的身份给客户推荐书籍</think>当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"}] • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。
问答排序	jsonl、csv	• jsonl格式：context表示问题，targets答案1、2、3表示答案的优劣顺序，最好的答案排在最前面。 { "context":"context内容","targets":["回答1","回答2","回答3"]} • csv格式：csv文件的第一列对应context，其余列为答案。 "问题","回答1","回答2","回答3" • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
偏好优化DPO	jsonl	• jsonl盘古格式-非思维链：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，具体格式示例如下： 单轮问答 {"context": ["你好，请介绍自己"], "target":"我是盘古大模型", "bad_target":"我不会回答"} 多轮问答 {"context": ["你好，请介绍自己", "我是盘古大模型", "请介绍一下有哪些产品。"], "target":"提供包括但不限于计算、存储、网络等产品服务。", "bad_target":"我不会回答"} • jsonl盘古格式-思维链：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，答案中至少有一个包含think标签对表示思考过程，具体格式示例如下： 单轮问答 {"context": ["你好，请介绍自己"], "target":"我是盘古大模型", "bad_target":"我不会回答"} 多轮问答 {"context": ["你好，请介绍自己", "我是盘古大模型", "请介绍一下有哪些产品。"], "target":"<think>客户想要了解产品</think>提供包括但不限于计算、存储、网络等产品服务。", "bad_target":"我不会回答"} • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
偏好优化DPO（人设）	jsonl	• jsonl盘古格式-非思维链：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，具体格式示例如下： 带人设单轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己"], "target":"哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "bad_target":"我不会回答"} 带人设多轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己", "哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "请介绍一下有哪些产品。"], "target":"我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！ ", "bad_target":"我不会回答"} • jsonl盘古格式-思维链：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，答案中至少有一个包含think标签对表示思考过程，具体格式示例如下： 带人设单轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己"], "target":"哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "bad_target":"我不会回答"} 带人设多轮问答 {"system": "你是一位机智幽默的问答助手", "context": ["你好，请介绍自己", "哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "请介绍一下有哪些产品。"], "target":"<think>客户想要了解产品</think>我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！ ", "bad_target":"我不会回答"} • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

4.3.2 图片类数据集格式要求

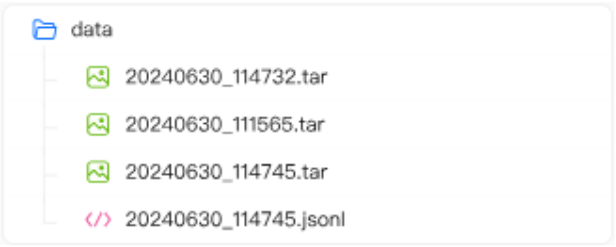
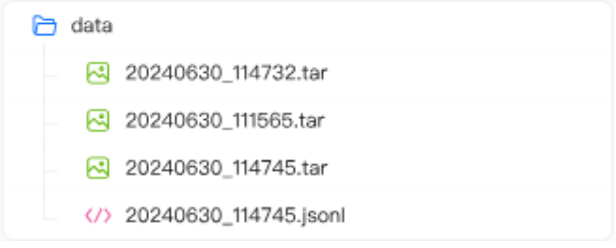
ModelArts Studio大模型开发平台支持创建图片类数据集，创建时可导入多种形式的数据，具体格式要求详见**表4-5**。

版权所有 © 华为云计算技术有限公司

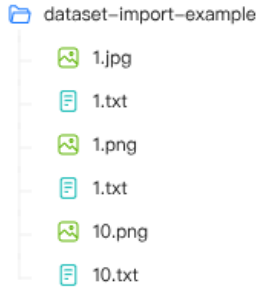
表 4-5 图片类数据集格式要求

文件内容	文件格式	文件要求
图片	tar、图片目录 可选：jsonl，名称固定为annotation.jsonl	• 图片：支持jpg、jpeg、png、bmp类型。  • tar：tar包内图片支持jpg、jpeg、png、bmp图片类型。  • 根目录下可存在单个annotation.jsonl文件，image_name字段必选。 {"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）"} • 从OBS导入：单个压缩包大小不超过50GB(只支持tar类型的压缩包)，单个文件大小不超过50GB，文件数量不限制。 本地上传：单个压缩包大小不超过10M(只支持tar类型的压缩包)，单个文件大小不超过10M，文件数量最多100个

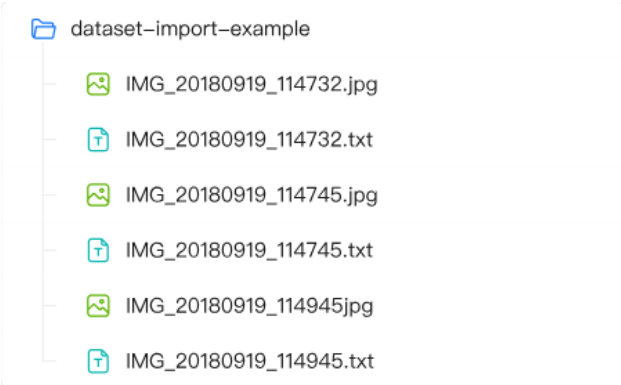
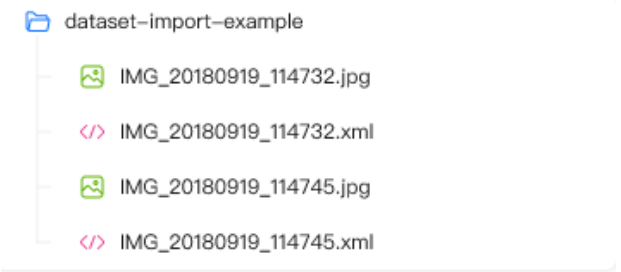
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
图片+Caption	图片支持tar，Caption支持jsonl	• 图片：图片以tar包格式存储，可以多个tar包。tar包存储原始的图片，每张图片命名要求唯一（如abc.jpg）。图片支持jpg、jpeg、png、bmp格式。 • jsonl：图片描述jsonl文件放在最外层目录，一个tar包对应一个jsonl文件，文件内容中每一行代表一段文本，形式为： {"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）","caption":"图片对应的文本描述"} • 从OBS导入：单个压缩包大小不超过50GB(只支持tar类型的压缩包)，单个文件大小不超过50GB，文件数量不限制。 本地上传：单个压缩包大小不超过10M(只支持tar类型的压缩包)，单个文件大小不超过10M，文件数量最多100个。 
图片+QA对	图片支持tar，QA对支持jsonl	• 图片：图片以tar包格式存储，可以多个tar包。tar包存储原始的图片，每张图片命名要求唯一（如abc.jpg）。图片支持jpg、jpeg、png、bmp格式。 • jsonl：图片描述jsonl文件放在最外层目录，一个tar包对应一个jsonl文件，文件内容中每一行代表一段文本，形式为： {"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）","conversations":[{"question":"问题1","answer":"回答1"},{"question":"问题2","answer","回答2"}]} • 从OBS导入：单个压缩包大小不超过50GB(只支持tar类型的压缩包)，单个文件大小不超过50GB，文件数量不限制。 本地上传：单个压缩包大小不超过10M(只支持tar类型的压缩包)，单个文件大小不超过10M，文件数量最多100个 

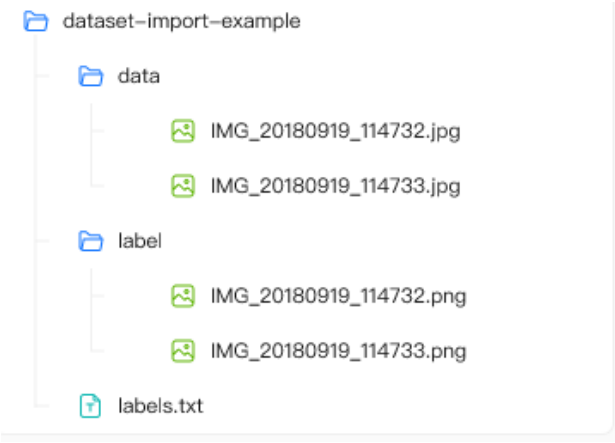
 版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
物体检测	PASCAL VOC	• 由图片文件和对应的标注文件构成，标注文件需要满足PASCAL VOC文件格式。要求用户将标注对象和标注文件存储在同一目录，并且相互对应，如标注对象文件名为“IMG_2.jpg”，那么标注文件的文件名应为“IMG_2.xml” • 图片支持jpg、jpeg、png、bmp、tif、tiff格式，标注文件为xml格式，标注文件说明请参见**物体检测数据集标注文件说明**。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 
图像分类	图片+txt	• 由图片文件和对应的标注文件构成，要求用户将标注对象和标注文件存储在同一目录，并且相互对应。 • 图片支持jpg、jpeg、png、bmp、tif、tiff格式，标注文件为txt格式，标注文件说明请参见**图像分类数据集标注文件说明**。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

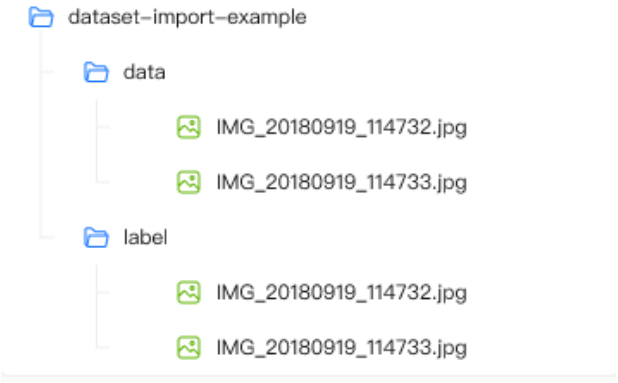
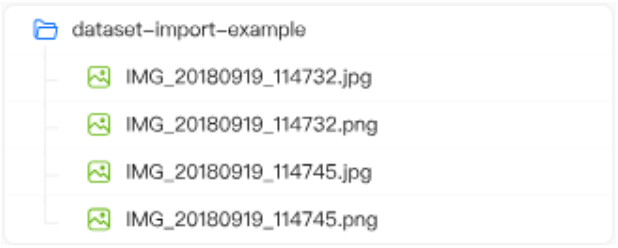
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
异常检测	图片+txt	• 文件存放方式要求满足异常检测格式，即标注文件和图片存于同一文件夹。 • 图片支持jpg、jpeg、png、bmp格式，标注文件为txt格式，标注文件说明请参见**异常检测数据集标注文件说明**。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 
语义分割	图片+xml	• 语义分割标注文件为xml格式，图片和标注文件一一对应。 图片格式支持：jpg、jpeg、png、bmp。 标注文件格式：xml。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

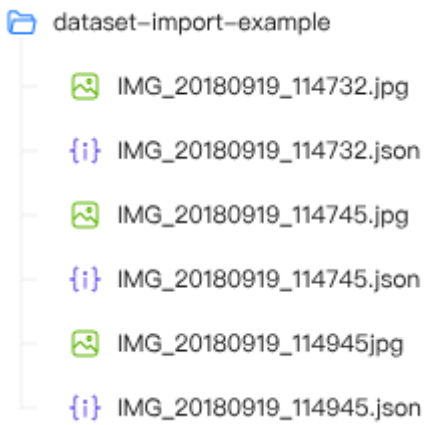
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
	原始图片+标注图片+json	• 语义分割标注方式为标注图片+json。 图片格式支持：jpg、jpeg、png、bmp。 标注文件格式：标注图片+json。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 
	原始图片+标注图片+txt	• 语义分割标注方式为标注图片+txt。 图片格式支持：jpg、jpeg、png、bmp。 标注文件格式：标注图片+txt。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

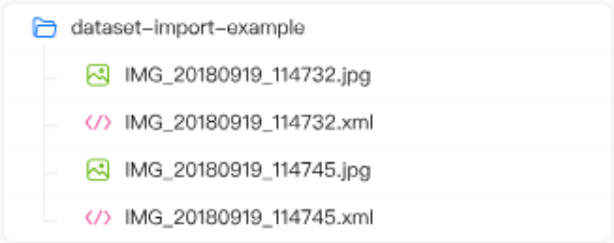
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
	原始图片+标注图片	• 语义分割标注方式为标注图片。 图片格式支持：jpg、jpeg、png、bmp。 标注文件格式：标注图片。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 
	图片+png；	• 文件存放方式要求满足语义分割格式，即原图为jpg文件，标注图采用同名同尺寸的png文件。其中，标注图上的每个像素值对应原图中像素的类别，且每个类别的值需连续且从0开始，表示不同的物体或区域类别。例如，假设有一张原图为IMG_20180919_114732.jpg，对应的标注图为IMG_20180919_114732.png，其中标注图的不同像素值代表不同的类别，标注图的每个像素值直接对应原图中相应位置的类别信息，类别需连续并从0开始。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

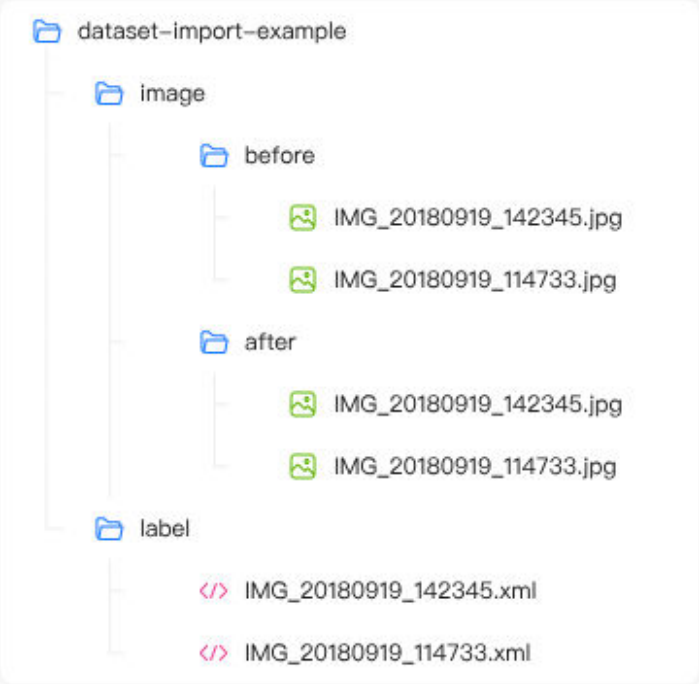
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
姿态估计	图片+json（一对一）	• 文件存放方式要求满足姿态估计格式。 图片格式支持：jpg、jpeg、png、bmp。 标注格式支持：json。 图片与标注文件一一对应。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 
	图片+json（多对一）	• 文件存放方式要求满足姿态估计格式。 图片格式支持：jpg、jpeg、png、bmp。 标注格式支持：json。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

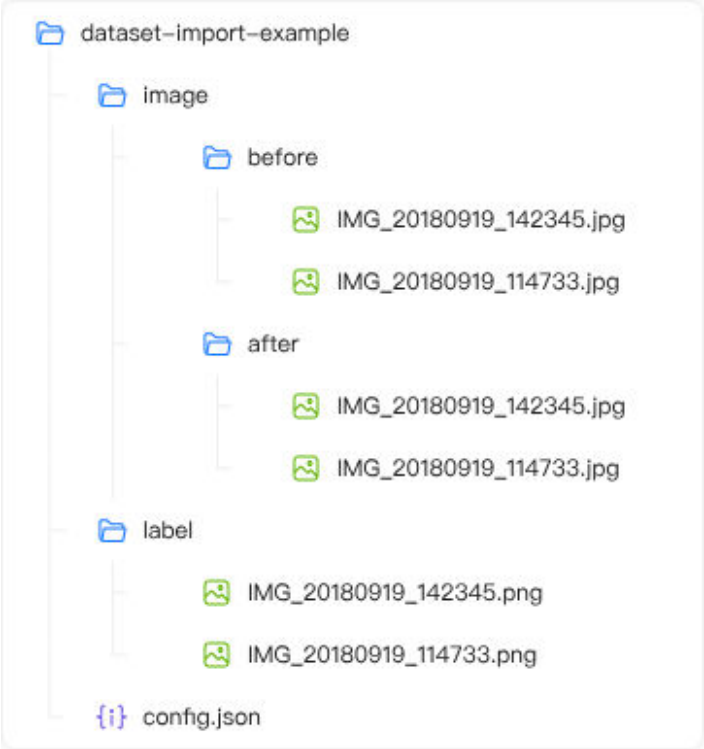
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
实例分割	图片+xml	• 文件存放方式要求满足万物分割/实例分割格式。 • 图片格式支持：jpg、jpeg、png、bmp，标注格式支持：xml • 基于PASCAL VOC矩形框格式进行标识，标注和图片同名并放在同一文件夹下。 • xml标注文件的详细说明请参见**实例分割数据集标注文件说明**。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

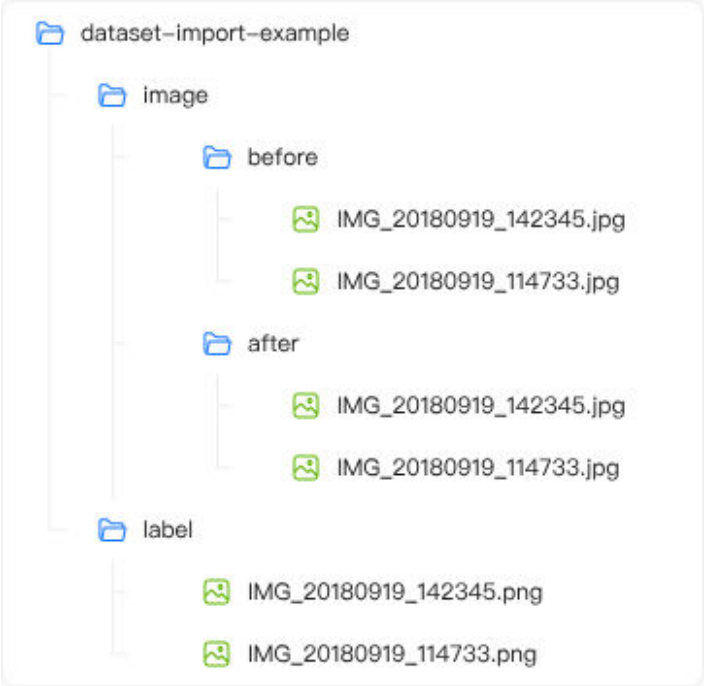
版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
变化检测	图片+xml	• 变化检测标注文件为xml格式。 • 图片格式支持：jpg、jpeg、png、bmp。 • 标注文件格式：xml。 • 前后两幅图片需要同名同尺寸。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。  其中，**before**文件夹：包含变化前的图片，每幅图片需与变化后的图片同名、同尺寸。 **after**文件夹：包含变化后的图片，每幅图片需与变化前的图片同名、同尺寸。 **label**文件夹：包含与变化前和变化后图片同名、同尺寸的PNG文件。每个像素值代表该位置对应的类别信息，类别应是连续的且从0开始。

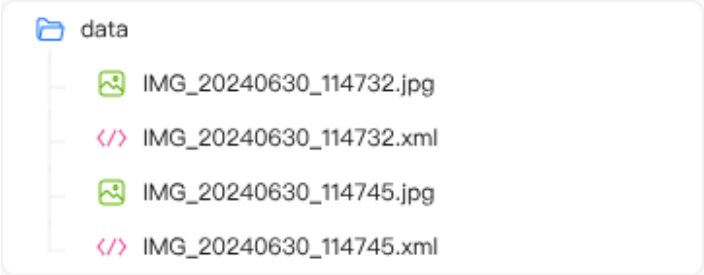
 版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
	原始图片+标注图片+json	• 变化检测标注方式为标注图片+json。 • 图片格式支持：jpg、jpeg、png、bmp。 • 标注文件格式：png+json。 • 前后两幅图片需要同名同尺寸。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。  其中，**before**文件夹：包含变化前的图片，每幅图片需与变化后的图片同名、同尺寸。 **after**文件夹：包含变化后的图片，每幅图片需与变化前的图片同名、同尺寸。 **label**文件夹：包含与变化前和变化后图片同名、同尺寸的PNG文件。每个像素值代表该位置对应的类别信息，类别应是连续的且从0开始。

 版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
	原始图片+标注图片	• 文件存放方式要求满足变化检测格式。 • 图片格式支持：jpg、jpeg、png、bmp。 • 标注文件格式：png。 • 前后两幅图片需要同名同尺寸，并采用同名同尺寸的png文件对图片文件不同像素赋予不同类别，png图片上的像素值对应于原图上像素值的类别，类别需连续并从0开始。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。  其中，**before**文件夹：包含变化前的图片，每幅图片需与变化后的图片同名、同尺寸。 **after**文件夹：包含变化后的图片，每幅图片需与变化前的图片同名、同尺寸。 **label**文件夹：包含与变化前和变化后图片同名、同尺寸的PNG文件。每个像素值代表该位置对应的类别信息，类别应是连续的且从0开始。

 版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
旋转检测	图片+xml	• 文件存放方式要求满足旋转检测格式。 • 图片格式支持：jpg、jpeg、png、bmp。 • 标注文件格式：xml。 • 图片与标注文件一一对应。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 

物体检测数据集标注文件说明

该说明适用于表4-5中的物体检测标注文件格式。

物体检测数据集支持格式为ModelArts PASCAL VOC 1.0。

要求用户将标注对象和标注文件存储在同一目录，并且一一对应，如标注对象文件名为“IMG_20180919_114745.jpg”，那么标注文件的文件名应为“IMG_20180919_114745.xml”。

物体检测的标注文件需要满足PASCAL VOC格式，PASCAL_VOC是一个公开的图像标注数据集，它提供了一个统一的XML格式来存储标注信息。PASCAL_VOC文件格式包含图像目录、图像文件、图像尺寸、图像中目标信息等元素，详细格式说明请参见表4-6。

OBS文件上传示例：

```
├─dataset-import-example
│      IMG_20180919_114732.jpg
│      IMG_20180919_114732.xml
│      IMG_20180919_114745.jpg
│      IMG_20180919_114745.xml
│      IMG_20180919_114945.jpg
│      IMG_20180919_114945.xml
```

标注文件（.xml文件）示例：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
   <folder>NA</folder>
   <filename>bike_1_1593531469339.png</filename>
   <source>
      <database>Unknown</database>
   </source>
   <size>
```

 版权所有 © 华为云计算技术有限公司

```
    <width>554</width>
    <height>606</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>Dog</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <occluded>0</occluded>
    <bndbox>
      <xmin>279</xmin>
      <ymin>52</ymin>
      <xmax>474</xmax>
      <ymax>278</ymax>
    </bndbox>
  </object>
  <object>
    <name>Cat</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <occluded>0</occluded>
    <bndbox>
      <xmin>279</xmin>
      <ymin>198</ymin>
      <xmax>456</xmax>
      <ymax>421</ymax>
    </bndbox>
  </object>
</annotation>
```

表 4-6 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示图像所在的目录名称。
filename	是	被标注文件的文件名。
size	是	表示图像的像素信息。 ● width：必选字段，图像的宽度。 ● height：必选字段，图像的高度。 ● depth：必选字段，图像的通道数。
segmented	是	表示是否用于分割，取值为0或1。0表示没有分割标注，1表示有分割标注。

版权所有 © 华为云计算技术有限公司

字段	是否必选	说明
object	是	目标对象信息，包括被标注物体的类别、姿态、是否被截断、是否识别困难以及边界框信息，多个物体标注会有多个object体。 ● name：必选字段，标注内容的类别。 ● pose：必选字段，标注内容的拍摄角度。 ● truncated：必选字段，取值0或1，表示标注内容是否被截断（0表示被截断、1表示没有截断）。 ● occluded：必选字段，取值0或1，表示标注内容是否被遮挡（0表示未遮挡、1表示遮挡） ● difficult：必选字段，取值0或1，表示标注目标是否难以识别（0表示容易识别、1表示难易识别）。 ● confidence：可选字段，标注目标的置信度，取值范围0-1之间，越接近1，表示标注越可信。 ● bndbox：必选字段，标注框的类型，可选值请参见表4-7。

表 4-7 标注框类型描述

type	形状	标注信息
point	点	点的坐标。 <x>100<x> <y>100<y>
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>
bndbox	矩形框	左上和右下两个点坐标。 <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>

版权所有 © 华为云计算技术有限公司

type	形状	标注信息
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	圆形	圆心坐标和半径。 <cx>100<cx> <cy>100<cy> <r>50<r>

图像分类数据集标注文件说明

该说明适用于表4-5中的图片分类标注文件格式。

图像分类数据集支持格式为ModelArts image classification 1.0。

要求用户将标注对象和标注文件存储在同一目录，并且一一对应，标注文件txt中可以放单标签，也可以放多标签。

- 当目录下存在对应的txt文件时，以txt文件内容作为图像的标签。

  示例如下所示，import-dir-1和import-dir-2为导入子目录。

```
dataset-import-example
├─import-dir-1
│      10.jpg
│      10.txt
│      11.jpg
│      11.txt
│      12.jpg
│      12.txt
└─import-dir-2
       1.jpg
       1.txt
       2.jpg
       2.txt
```

  单标签的标签文件示例，如1.txt文件内容如下所示。

```
猫
```

  多标签的标签文件示例，如2.txt文件内容如下所示。

```
猫
狗
```


版权所有 © 华为云计算技术有限公司

异常检测数据集标注文件说明

该说明适用于表4-5中的异常检测标注文件格式。

要求用户将标注文件和图片存于同一文件夹。

- 当目录下存在对应的txt文件时，以txt文件内容作为正常或异常的标签。

  示例如下所示：

  ```
  dataset-import-example
  │       IMG_20180919_114732.jpg
  │       IMG_20180919_114732.txt
  │       IMG_20180919_114745.jpg
  │       IMG_20180919_114745.txt
  ```

  异常标签的标签文件示例，如IMG_20180919_114732.txt文件内容如下所示。

  ```
  abnormal
  ```

  正常标签的标签文件示例，如IMG_20180919_114745.txt文件内容如下所示。

  ```
  normal
  ```

姿态估计标注 json 文件说明

该说明适用于表4-5中的姿态估计标注文件格式。

姿态估计标注基于开源coco人物关键点标注格式对数据集进行标注，需包含annotations，train，val文件夹。annotations文件夹下用train.json和val.json记录训练集和验证集标注，train和val文件夹下保存具体的图片，示例如下所示：

```
├─annotations
│      train.json
│      val.json
├─train
│      IMG_20180919_114745.jpg
├─val
│      IMG_20180919_114945.jpg
```

具体的json标注文件具体示例：

```
{
   "images": [
      {
         "license": 2,
         "file_name": "000000000139.jpg",
         "coco_url": "",
         "height": 426,
         "width": 640,
         "date_captured": "2013-11-21 01:34:01",
         "flickr_url": "",
         "id": 139
      }
   ],
   "annotations": [
      {
         "num_keypoints": 15,
         "area": 2913.1104,
         "iscrowd": 0,
         "keypoints": [
            427,
            170,
            1,
            429,
            169,
            2,
            0,
            0,
```


版权所有 © 华为云计算技术有限公司

```
            0,
            434,
            168,
            2,
            0,
            0,
            0,
            441,
            177,
            2,
            446,
            177,
            2,
            437,
            200,
            2,
            430,
            206,
            2,
            430,
            220,
            2,
            420,
            215,
            2,
            445,
            226,
            2,
            452,
            223,
            2,
            447,
            260,
            2,
            454,
            257,
            2,
            455,
            290,
            2,
            459,
            286,
            2
        ],
        "image_id": 139,
        "bbox": [
            412.8,
            157.61,
            53.05,
            138.01
        ],
        "category_id": 1,
        "id": 230831
    },
],
"categories": [
    {
        "supercategory": "person",
        "id": 1,
        "name": "person",
        "keypoints": [
            "nose",
            "left_eye",
            "right_eye",
            "left_ear",
            "right_ear",
            "left_shoulder",
            "right_shoulder",
            "left_elbow",
```


 版权所有 © 华为云计算技术有限公司

```
            "right_elbow",
            "left_wrist",
            "right_wrist",
            "left_hip",
            "right_hip",
            "left_knee",
            "right_knee",
            "left_ankle",
            "right_ankle"
        ],
        "skeleton": [
            [
                16,
                14
            ],
            [
                14,
                12
            ],
            [
                17,
                15
            ],
            [
                15,
                13
            ],
            [
                12,
                13
            ],
            [
                6,
                12
            ],
            [
                7,
                13
            ],
            [
                6,
                7
            ],
            [
                6,
                8
            ],
            [
                7,
                9
            ],
            [
                8,
                10
            ],
            [
                9,
                11
            ],
            [
                2,
                3
            ],
            [
                1,
                2
            ],
            [
                1,
```


 版权所有 © 华为云计算技术有限公司

```
            3
          ],
          [
            2,
            4
          ],
          [
            3,
            5
          ],
          [
            4,
            6
          ],
          [
            5,
            7
          ]
        ]
      }
    ]
}
```

表 4-8 COCO 格式说明

字段	是否必选	说明
images	是	图片信息。
license	否	图像的许可证标识符。
file_name	是	图像的文件名。
coco_url	否	图像在COCO官方数据集中的URL。
height	是	图像的高度，以像素为单位。
width	是	图像的宽度，以像素为单位。
date_captured	否	图像捕获的日期和时间。
flickr_url	否	图像在Flickr网站上的URL。
id	是	图像的唯一标识符。
annotations	是	标注信息。
num_keypoints	是	标注的关键点数量。
area	是	边界框的面积，以像素平方为单位。
iscrowd	是	表示标注是否为复杂的群体场景（如拥挤的人群）。0表示不是拥挤场景，1表示是拥挤场景。
keypoints	是	标注的关键点坐标及其可见性，按顺序列出所有关键点，每个关键点用三个数值表示 [x, y, v]。x和y是关键点的像素坐标，v是可见性（0：不可见且不在图像中；1：不可见但在图像中；2：可见且在图像中）。

版权所有 © 华为云计算技术有限公司

字段	是否必选	说明
image_id	是	与该标注相关联的图像的ID，必须与images字段中的id对应。
bbox	是	目标物体的边界框，用[x, y, width, height]表示，其中，x，y是边界框左上角的坐标，width和height是边界框的宽度和高度。
category_id	是	标注类别的ID，对于人体姿态估计，通常为1（表示person）。
id	是	标注的唯一标识符。
categories	是	标注类型信息。
supercategory	是	类别的上级分类，通常为person。
id	是	类别的唯一标识符，对于人体姿态估计，通常为1。
name	是	类别的名称，通常为person。
keypoints	是	关键点的名称列表，COCO格式中通常定义了17个关键点，如nose、left_eye、right_eye、left_ear、right_ear、left_shoulder、right_shoulder、left_elbow、right_elbow、left_wrist、right_wrist、left_hip、right_hip、left_knee、right_knee、left_ankle、right_ankle。
skeleton	是	定义骨架连接的列表，用于表示关键点之间的连接关系。每个连接用一对关键点索引表示，如 [1, 2]，表示鼻子（nose）到左眼（left_eye）的连线。

实例分割数据集标注文件说明

该说明适用于**表4-5**中的实例分割标注文件格式。

要求用户将标注对象和标注文件存储在同一目录，并且一一对应，如标注对象文件名为“IMG_20180919_114745.jpg”，那么标注文件的文件名应为“IMG_20180919_114745.xml”。

实例分割的标注文件需要满足PASCAL VOC格式，PASCAL_VOC是一个公开的图像标注数据集，它提供了一个统一的XML格式来存储标注信息。PASCAL_VOC文件格式包含图像目录、图像文件、图像尺寸、图像中目标信息等元素，详细格式说明请参见**表4-9**。

OBS文件上传示例：

```
├─dataset-import-example
    IMG_20180919_114732.jpg
    IMG_20180919_114732.xml
```

 版权所有 © 华为云计算技术有限公司

```
    IMG_20180919_114745.jpg
    IMG_20180919_114745.xml
```

标注文件（.xml文件）示例：

```
<annotation>
 <folder>NA</folder>
 <filename>0001.jpg</filename>
 <source>
  <database>Unknown</database>
 </source>
 <size>
  <width>2560</width>
  <height>1440</height>
  <depth>3</depth>
 </size>
 <segmented>1</segmented>
 <mask_source></mask_source>
 <object>
  <name>aggregate</name>
  <pose>Unspecified</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <mask_color>238,130,238</mask_color>
  <occluded>0</occluded>
  <polygon>
   <x1>657.0</x1>
   <y1>357.0</y1>
   <x2>645.0</x2>
   <y2>351.0</y2>
   <x3>624.0</x3>
   <y3>352.0</y3>
   <x4>616.0</x4>
   <y4>353.0</y4>
  </polygon>
 </object>
</annotation>
```

表 4-9 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示图像所在的目录名称。
filename	是	被标注文件的文件名。
size	是	表示图像的像素信息。 ● width：必选字段，图像的宽度。 ● height：必选字段，图像的高度。 ● depth：必选字段，图像的通道数。
segmented	是	表示是否用于分割，取值为0或1。0表示没有分割标注，1表示有分割标注。

版权所有 © 华为云计算技术有限公司

字段	是否必选	说明
object	是	目标对象信息，包括被标注物体的类别、姿态、是否被截断、是否识别困难以及边界框信息，多个物体标注会有多个object体。 ● name：必选字段，标注内容的类别。 ● pose：必选字段，标注内容的拍摄角度。 ● truncated：必选字段，取值0或1，表示标注内容是否被截断（0表示被截断、1表示没有截断）。 ● occluded：必选字段，取值0或1，表示标注内容是否被遮挡（0表示未遮挡、1表示遮挡） ● difficult：必选字段，取值0或1，表示标注目标是否难以识别（0表示容易识别、1表示难易识别）。 ● confidence：可选字段，标注目标的置信度，取值范围0-1之间，越接近1，表示标注越可信。 ● bndbox：必选字段，标注框的类型，可选值请参见表4-10。

表 4-10 标注框类型描述

type	形状	标注信息
point	点	点的坐标。 <x>100<x> <y>100<y>
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>
bndbox	矩形框	左上和右下两个点坐标。 <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>

版权所有 © 华为云计算技术有限公司

type	形状	标注信息
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	圆形	圆心坐标和半径。 <cx>100<cx> <cy>100<cy> <r>50<r>

4.3.3 视频类数据集格式要求

ModelArts Studio大模型开发平台支持创建视频类数据集，创建时可导入多种形式的数据，具体格式要求详见表4-11。

表 4-11 视频类数据集格式要求

文件内容	文件格式	文件要求
视频	mp4或avi	• 支持mp4、avi视频格式上传，所有视频可以放在多个文件夹下，每个文件夹下可以同时包含mp4或avi格式的视频。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求
视频+标注	视频+json	• 视频格式支持：mp4、avi • 标注文件格式：jsonl，jsonl文件仅支持UTF-8编码。 示例如下所示：  具体的jsonl标注文件参考： {"video_fn": "13/ad098173-af09-48fe-95c3-e72fd629688e.mp4"视频相对路径， "prompt": "A person pours a clear liquid from a bottle into a shot glass, then lifts the glass to their mouth and drinks the shot. The background includes a red coat and other indistinct background elements."视频摘要生成（简略）， "long_prompt": "A person is seen pouring a clear liquid from a green glass bottle into a small glass. The individual is wearing a white shirt with a lace collar and a beige cardigan. The background appears to be a cozy indoor setting, possibly a cafe or a restaurant, with red and white elements visible, such as a red coat hanging on the wall and a white table. The person carefully pours the liquid, ensuring it is filled to the brim of the glass. The liquid is clear and has some green leaves floating in it. The person then holds the glass up, possibly to show the contents or to prepare for a drink.",视频摘要生成（详细） }

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求

事件检测	视频+json	数据源样本为avi、mp4格式，标注文件为json格式。必须包含两个及以上后缀名字为avi或者mp4的文件，视频与标注文件可以一对一，也可以多对一。
		每个视频时长要大于128s，FPS>=10，且测试集训练集都要有视频。
		支持视频的格式包括常见的mp4/avi格式文件，每个视频时长要大于128s，FPS>=10，用annotation.json对文件进行标注。
		从OBS导入：单个文件大小不超过50GB，文件数量不限制。
		示例如下所示：
		视频+json（多对一）
		 
		视频+json（一对一）
		 
		具体的json标注文件参考：
		{
		'version': 'dataset_name_v.x.x',// 数据集版本信息。
		'classes': [category1',category2', ...],// 所有类别名称的列表，每个类别对应一个 label，用于标注视频中的事件或动作。
		'database': {
		'video_name':{
		// 训练集 train 测试集 test。
		'subset': 'train',
		'duration': 1660.3, // 视频总时长 seconds。
		'fps': 30.0,// 视频帧率。
		'width': 720,// 视频宽度，单位像素。
		'height': 1280,// 视频高度，单位像素。
		'ext': 'mp4',//视频文件扩展名。
		// 标注 34.5, 42.4 分别表示起始时间和结束时间，单位为s。
		// label 表示分类，必须是classes列表中的一个元素，表示该视频片段对应的事件或动作类型。
		'annotations': [
		{'label': 'category1', 'segment': [34.5, 42.4]},
		{'label': 'category1', 'segment': [124.4, 142.9]},
		...
		]

 版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件要求

		<pre>},
		'video_name':{
		'subset': xxx,//视频文件名称，不包括扩展名。
		'duration': xxx,
		'fps': xxx,
		'width': xxx,
		'height': xxx,
		'ext': xxx,
		'annotations': [
		{'label': xxx, 'segment': xxx},
		{'label': xxx, 'segment': xxx},
		...
		]
		},
		...
		}
		}</pre>

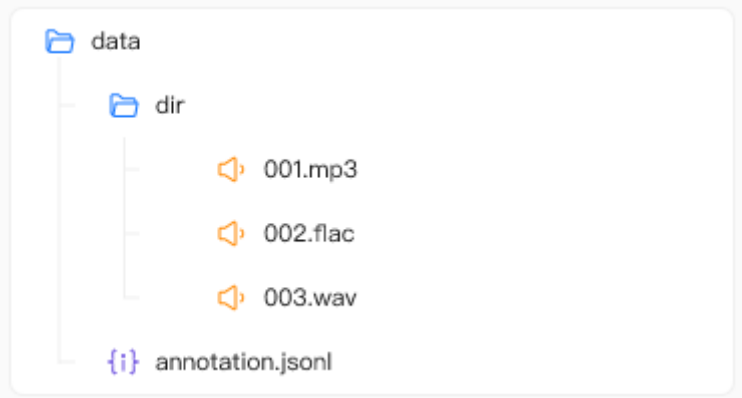
视频分类	视频+txt	• 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 • 视频格式支持：mp4、avi，标注文件格式：txt，视频与标注文件一一对应。 示例如下所示： 

4.3.4 音频类数据集格式要求

ModelArts Studio大模型开发平台支持创建音频类数据集，创建时可导入多种形式的数据，具体格式要求详见**表4-12**

版权所有 © 华为云计算技术有限公司

表 4-12 音频类数据集格式要求

文件内容	文件格式	文件要求
音频	音频+jsonl（可选）	• 音频格式支持：支持mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 示例如下所示：  具体的jsonl标注文件参考： {"file_name":"dir/16k_16bit_1channel_2s.flac","caption":"1"} {"file_name":"dir/16k_16bit_1channel_2s.mp3","caption":"2"} {"file_name":"dir/16k_16bit_1channel_2s.opus","caption":"3"} {"file_name":"dir/16k_16bit_1channel_2s.wav","caption":"4"}

4.3.5 气象类数据集格式要求

ModelArts Studio大模型开发平台支持导入气象类数据集，该数据集当前包括**海洋气象**数据。

海洋气象数据通常来源于气象再分析。气象再分析是通过现代气象模型和数据同化技术，重新处理历史观测数据，生成高质量的气象记录。这些数据既可以覆盖全球范围，也可以针对特定区域，旨在提供完整、一致且高精度的气象数据。

再分析数据为二进制格式，具体格式要求详见**表4-13**。

版权所有 © 华为云计算技术有限公司

表 4-13 气象类数据集格式要求

文件内容	文件格式	文件样例
气象-天气数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	● 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 天气数据通常包含全球或区域性的气象变量，如温度（T）、气压（P）、风速（U、V）等。在文件中，这些变量可能按时间、地理范围和气压层次进行组织。示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["P", "T", "U", "V"], "upper_air_layers": ["1000hPa", "100hPa", "150hPa", "175hPa", "200hPa", "250hPa", "300hPa", "400hPa", "500hPa", "50hPa", "600hPa", "700hPa", "850hPa", "925hPa"], "upper_air_features": ["Q", "T", "U", "V", "Z"]} ● geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 ● time_range：数据的时间范围，时间戳格式为毫秒数。 ● total_size：数据文件的总大小，单位为字节。 ● surface_features：地表特征变量列表，例如气压（P）、温度（T）、风速（U、V）。 ● upper_air_layers：高空气压层列表，例如1000hPa、100hPa等。 ● upper_air_features：高空特征变量列表，例如湿度（Q）、温度（T）、风速（U、V）、高度（Z）。
气象-海洋数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	● 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 ● 海洋数据通常包含全球或区域性的海洋变量，如温度（T）、气压（P）、风速（U、V）等，具体格式示例如下 {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SSH", "T", "P", "U", "V"], "under_sea_layers": ["0m", "6m", "10m", "20m", "30m", "50m", "70m", "100m", "125m", "150m", "200m", "250m", "300m", "400m", "500m"], "under_sea_features": ["T", "U", "V", "S"]} – geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 – time_range：数据的时间范围，时间戳格式为毫秒数。 – total_size：数据文件的总大小，单位为字节。 – surface_features：海表特征变量列表，例如海表高度（SSH）、温度（T）、风速（U、V）。 – under_sea_layers：深海层列表，例如500m、400mPa等。 – under_sea_features：高空特征变量列表，例如海盐（S）、温度（T）、海流速率（U、V）。

版权所有 © 华为云计算技术有限公司

文件内容	文件格式	文件样例
气象-生态数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	• 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 生态数据通常包含总叶绿素浓度（Tca）、叶绿素浓度（Chl）、硅藻浓度（Dia）等生态变量。示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["Tca","Chl","Dia","Coc","Cya","Irn","Nit","MLD"]} • geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 • time_range：数据的时间范围，时间戳格式为毫秒数。 • total_size：数据文件的总大小，单位为字节。 • surface_features：生态特征列表，例如总叶绿素浓度（Tca）、叶绿素浓度（Chl）、硅藻浓度（Dia）。
气象-海浪数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	• 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 海浪数据通常包有效波高（SWH）。示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SWH"]} • geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 • time_range：数据的时间范围，时间戳格式为毫秒数。 • total_size：数据文件的总大小，单位为字节。 • surface_features：海浪特征：有效波高（SWH）。

4.3.6 预测类数据集格式要求

平台支持创建预测类数据集，创建时可导入时序分类数据、时序回归数据、结构化分类数据、结构化回归数据。

- **时序分类数据：**含有时间标记的数据序列，输入列为特征列，目标列是离散值，并具有有限的可能取值。
- **时序回归数据：**含有时间标记的数据序列，输入列为特征列，目标列是连续值。
- **结构化分类数据：**包含多种预测因子的数据集，输入列为特征列，目标列是离散值，并具有有限的可能取值。
- **结构化回归数据：**包含多种预测因子的数据集，输入列为特征列，目标列是连续值。

具体格式要求详见表4-14。

版权所有 © 华为云计算技术有限公司

表 4-14 预测类数据集格式要求

文件内容	文件格式	文件样例
时序分类、时序回归	csv	• 数据为结构化数据，包含列和行，每一行表示一条数据，每一列表示一个特征，并且必须包含预测目标列，回归预测目标列要求为连续型数据、分类预测目标列要求为离散数据。 • 目录下只有1个数据文件时，文件无命名要求。 • 目录下有多个数据文件时，需要通过命名的方式指定数据是训练数据集、验证数据集还是测试数据集。训练数据名称需包含train字样，如train01.csv；验证数据名称需包含eval字样；测试数据名称需包含test字样。单个文件的命名不能同时包含train、eval和test中的两个或三个。 • 时序预测必须要包含一个时间列，时间列值的格式示例为2024-05-27 或 2024/05/27 或 2024-05-27 12:00:00 或 2024/05/27 12:00:00 。 示例如下： timestamp,feature1,feature2,target 2024-05-27 12:00:00,10.5,20.3,100 2024-05-27 12:01:00,10.6,20.5,101 2024-05-27 12:02:00,10.7,20.7,102 2024-05-27 12:03:00,10.8,20.9,103 2024-05-27 12:04:00,10.9,21.0,104 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。
结构化分类、结构化回归	csv	• 数据为结构化数据，包含列和行，每一行表示一条数据，每一列表示一个特征，并且必须包含预测目标列，回归预测目标列要求为连续型数据，分类预测目标列要求为离散数据。 • 目录下只有1个数据文件时，文件无命名要求。 • 目录下有多个数据文件时，需要通过命名的方式指定数据是训练数据集、验证数据集还是测试数据集。训练数据名称需包含train字样，如train01.csv；验证数据名称需包含eval字样；测试数据名称需包含test字样。单个文件的命名不能同时包含train、eval和test中的两个或三个。 示例如下： feature1,feature2,target 10.5,20.3,100 10.6,20.5,101 10.7,20.7,102 10.8,20.9,103 10.9,21.0,104 • 从OBS导入：单个文件大小不超过50GB，文件数量不限制。 本地上传：单个文件大小不超过10M，文件数量最多100个。

4.3.7 其他类数据集格式要求

除文本、图片、视频、气象、预测类数据集外，平台还支持导入其他类数据集，即用户训练模型时使用的自定义数据集。

其他类数据集支持**加工其他类数据集**与**发布其他类数据集**操作。

版权所有 © 华为云计算技术有限公司

从OBS导入：单个文件大小不超过50GB，文件数量不限制。

本地上传：单个文件大小不超过10M，文件数量最多100个。

4.4 导入数据至盘古平台

数据集是一组用于处理和分析的相关数据样本。

用户将存储在OBS服务中的数据导入至ModelArts Studio大模型开发平台后，将生成“原始数据集”被平台统一管理，用于后续加工或发布操作。

创建导入任务

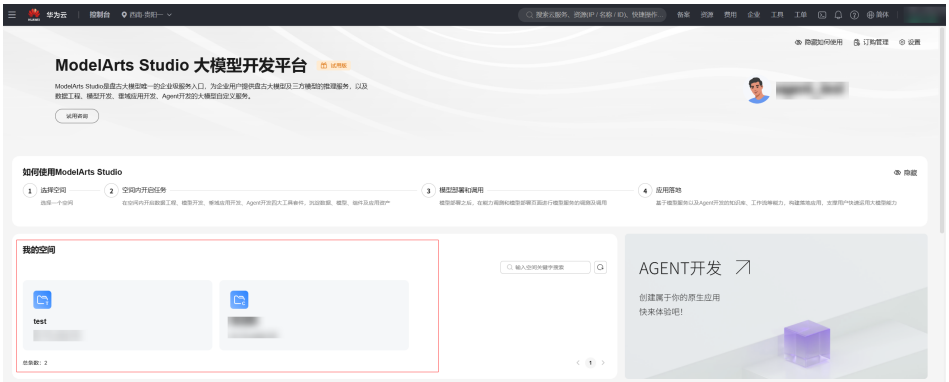
创建导入任务前，请先按照**数据集格式要求**提前准备数据。

平台支持使用OBS服务导入数据，请详见**通过控制台快速使用OBS**。

创建导入任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击空间名称进入操作空间。

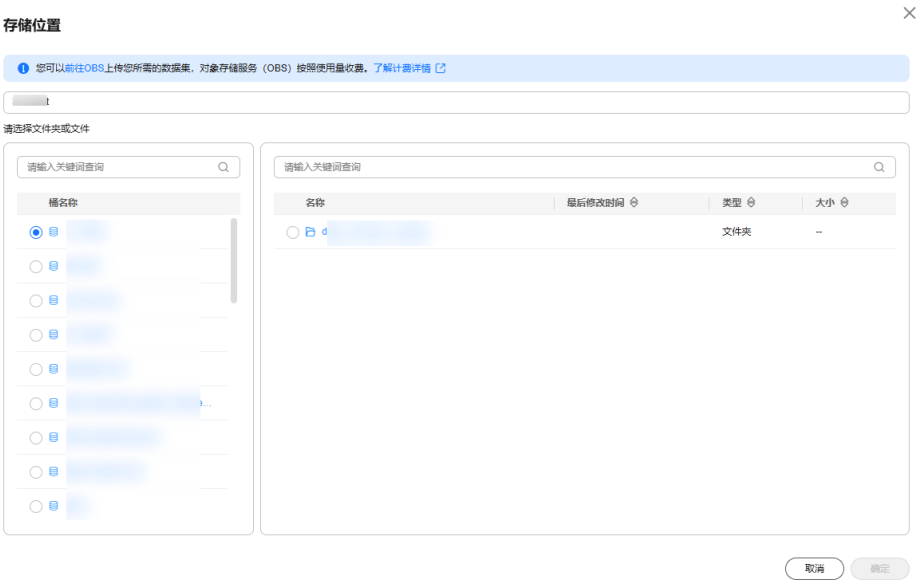
图 4-2 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面，选择“数据集类型”、“文件格式”和“导入来源”。
4. 导入来源选择“OBS”，单击“  ”，在“存储位置”弹窗中选择需导入的数据，单击“确定”。

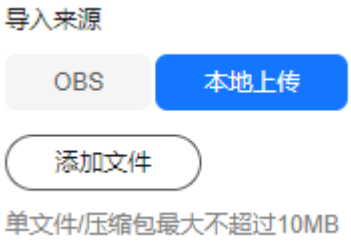
版权所有 © 华为云计算技术有限公司

图 4-3 选择导入的数据



导入来源可选择“本地上传”，单击“添加文件”，选择需要导入的文件。

图2 本地上传



5. 填写“数据集名称”和“描述”，可选择填写“扩展信息”。

   扩展信息包括“数据集属性”与“数据集版权”：

   - 数据集属性。可以给数据集添加行业、语言和自定义信息。
   - 数据集版权。训练模型的数据集除用户自行构建外，也可能会使用开源的数据集。数据集版权功能主要用于记录和管理数据集的版权信息，确保数据的使用合法合规，并清晰地了解数据集的来源和相关的版权授权。通过填写这些信息，可以追溯数据的来源，明确数据使用的限制和许可，从而保护数据版权并避免版权纠纷。

6. 单击页面右下角“立即创建”，回退至“导入任务”页面，在该页面可以查看数据集的任务状态，若状态为“运行成功”，则数据导入成功。
7. 导入后的数据集可在“数据工程 > 数据管理 > 数据集 > 原始数据集”中查看。

   说明

   如果任务状态为“运行失败”，可能由以下原因导致：

   - 文件后缀校验不通过，需要检查文件后缀是否一致。例如，选择创建csv格式的数据集时，文件后缀应为“.csv”。
   - 文件内容校验不通过，需要检查上传的文件数据格式是否正确。可以在“创建导入任务”页面下载数据样例进行比对。

版权所有 © 华为云计算技术有限公司

管理原始数据集

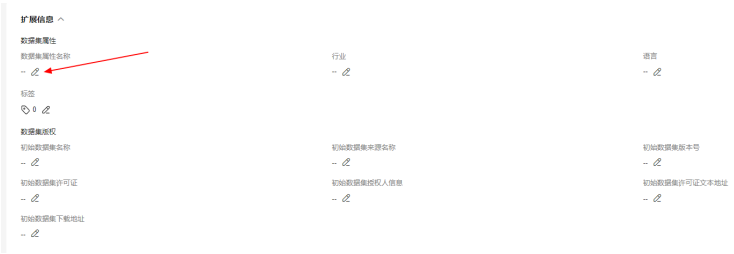
数据导入成功后，可对原始数据集进行统一管理，支持的操作如下：查看数据集的基本信息、数据血缘、操作记录以及对下载、删除数据集等操作。

1. 登录ModelArts Studio大模型开发平台，单击进入操作空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据集 > 原始数据集”。
3. 数据集支持“数据集名称”、“数据集模态”、“数据集属性名称”等属性类型进行搜索、过滤。



4. 单击需要查看的数据集名称。
   - 查看数据集基本信息。在“基本信息”页签，可以查看数据详情、数据来源以及扩展信息。
   - 在扩展信息内，可以根据用户需求设置数据集属性，包括数据集属性名称、行业、语言、自定义标签。单击编辑图标，即可对数据集的扩展信息进行编辑和保存，如图4-4

     图 4-4 编辑扩展信息

     
   - 下载原始数据集。在“数据预览”页签，可以查看数据内容，单击右上角“下载”即可下载原始数据集。
   - 查看数据血缘。在“数据血缘”页签，可以查看当前数据集所经历的完整操作，如加工、标注等。数据血缘上钻至OBS源文件下钻至模型训练。

版权所有 © 华为云计算技术有限公司

- 查看操作记录。在“操作记录”页签，可以查看当前数据集的操作记录，如创建该数据集的时间、状态、操作人员等。

5. 删除原始数据集。单击操作列的“删除”，并在弹窗中单击“确定”。

说明

删除“原始数据集”属于高危操作，删除前，请确保该数据集不再使用。

4.5 加工数据集

4.5.1 数据集加工场景介绍

数据加工介绍

ModelArts Studio大模型开发平台提供数据加工功能，涵盖了数据加工、数据合成和数据标注关键操作，旨在确保原始数据符合业务需求和模型训练的标准，是数据工程中的核心环节。

- **数据加工**

  通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。此外，用户还可以创建自定义算子，针对特定业务场景和模型需求，灵活地进行数据加工，从而进一步优化数据处理流程，提高模型的准确性和鲁棒性。

- **数据合成**

  利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。

- **数据标注**

  为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。

- **数据配比**

  数据配比是将多个数据集按特定比例组合为一个加工数据集的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。

通过这些数据加工操作，平台能够有效清理噪声数据、标准化数据格式，并优化数据集的整体质量。数据加工不仅仅是简单的数据处理，它还会根据数据类型和业务场景进行有针对性的优化，从而为模型训练提供高质量的输入，提升模型的表现。

数据加工意义

数据加工在大模型开发中具有至关重要的作用，具体体现在以下几个方面：

- **提高数据质量**

  原始数据往往包含噪声、缺失值或不一致性，这会直接影响模型训练效果。通过数据加工操作，可以有效去除无效信息、填补缺失数据，确保数据的准确性与一致性，从而提高数据质量，为模型训练提供可靠的输入。

- **扩展数据集的多样性和泛化能力**

 版权所有 © 华为云计算技术有限公司

在数据量不足或样本不平衡的情况下，数据合成可以生成新数据，扩展数据集的规模和多样性。通过增加数据的多样性，能够提升模型在各种场景下的泛化能力，增强其对未知数据的适应性。

- **增强模型训练的有效性**

  高质量的数据是训练好模型的基础。数据加工不仅仅是对数据的简单处理，更是根据不同数据类型和业务需求进行有针对性的优化，使数据更符合训练标准，提高训练效率和精度。

- **确保业务需求对接**

  不同业务场景和模型应用对数据有不同的要求。数据加工能够根据特定业务需求进行定制化处理，确保数据满足应用场景的需求，从而提高数据和模型的匹配度，提升业务决策和模型预测的准确性。

- **提升数据处理效率**

  通过平台提供的自动化加工功能，用户可以高效完成大规模数据的预处理工作，减少人工干预，提升数据处理的一致性和效率，确保整个数据工程流程的顺畅运行。

- **确保数据质量和适配性**

  通过数据配比，确保数据集满足大模型训练的高标准。这不仅包括数据规模的要求，还涵盖了数据质量、平衡性和代表性的保证，避免数据不均衡或不具备足够多样性的情况，进而提高模型的准确性和鲁棒性。

- **提高数据的多样性和代表性**

  通过合理的数据配比，帮助用户按特定比例组合多个数据集，确保数据集在不同任务场景下的多样性和代表性。这样可以避免过度偏向某一类数据，保证模型能够学习到多种特征，提升对各种情况的适应能力。

总体而言，数据加工不仅提升了数据处理的效率，还可通过优化数据质量和针对性处理，支持高效的模型训练。通过数据加工，用户能够快速构建高质量的数据集，推动大模型的成功开发。

支持数据加工的数据集类型

当前支持数据加工操作的数据集类型见**表4-15**。

表 4-15 支持数据加工操作的数据集类型

数据集模态	数据集类型	数据加工	数据合成	数据标注	数据配比
文本类	文档	√	-	-	-
	网页	√	-	-	-
	预训练文本	√	√	-	√
	单轮问答	√	√	√	√
	单轮问答+人设	√	√	√	√
	多轮问答	√	-	√	√
	多轮问答+人设	√	-	√	√
	问答排序	√	-	√	√

版权所有 © 华为云计算技术有限公司

数据集模态	数据集类型	数据加工	数据合成	数据标注	数据配比
	偏好优化DPO	-	-	-	√
	偏好优化DPO+人设	-	-	-	√
图片类	图片	√	-	√	-
	图片+Caption	√	-	√	√
	图片+QA对	√	-	√	√
	物体检测	-	-	-	-
	图像分类	-	-	-	-
	异常检测	-	-	-	-
	语义分割	-	-	-	-
	姿态估计	-	-	-	-
	实例分割	-	-	-	-
	变化检测	-	-	-	-
	旋转检测	-	-	-	-
视频类	视频	√	-	√	-
	视频+标注	√	-	√	√
	视频理解	√	-	√	-
	事件检测	-	-	-	-
	视频分类	-	-	-	-
音频类	音频	√	-	√	-
气象类	气象数据	√	-	-	-
预测类	时序（分类）	√	-	-	√
	时序（回归）	√	-	-	-
	结构化（分类）	√	-	-	√
	结构化（回归）	√	-	-	-
其他类	其他类	√（仅可使用自定义算子进行数据加工）	-	-	-

版权所有 © 华为云计算技术有限公司

4.5.2 数据集加工算子介绍

4.5.2.1 文本类加工算子介绍

数据加工算子为用户提供了多种数据操作能力，包括数据提取、过滤、转换、打标签等。这些算子能够帮助用户从海量数据中提取出有用信息，并进行深度加工，以生成高质量的训练数据。

平台支持文本类数据集的加工操作，分为数据提取、数据转换、数据过滤、数据打标四类，文本类加工算子能力清单见**表4-16**。

表 4-16 文本类加工算子能力清单

算子分类	算子名称	算子描述
数据提取	**WORD内容提取**	从Word文档中提取文本，保留原文档的标题和正文等结构，不保留图片、公式、页眉、页脚，不支持嵌套表格提取。
	TXT内容提取	从TXT文件中提取所有文本内容。
	CSV内容提取	从CSV文件中读取所有文本内容，并按该文件内容类型模板KEY值生成匹配的JSON格式数据。
	PDF内容提取	从PDF中提取文本，转化为结构化数据，支持文本、表格、公式等内容提取。
	HTML内容提取	基于标签路径提取HTML数据内容，并将其他与待提取标签路径无关的内容删除。
	电子书内容提取	从电子书中提取出所有文本内容。
数据转换	**个人数据脱敏**	对文本中的手机号码、身份证件、邮箱地址、url链接、国内车牌号、IP地址、MAC地址、IMEI、护照、车架号等个人敏感信息进行数据脱敏，或直接删除敏感信息。
	中文简繁转换	将中文简体和中文繁体进行转换。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
	符号标准化	查找文本中携带的非标准化符号进行标准化、统一化转换。 • 统一空格：将所有Unicode空格（如U+00A0、U+200A）转换为标准空格（U+0020）。 • 全角转半角：将文本中的全角字符转换为半角字符。 • 标点符号归一化，支持统一格式的符号如下： – {"？ ": "\?\？ "} – {"[":"〖 "} – {"]":"〗"} • 数字符号归一化，例如将⓪|□|□|□|□统一为0。支持统一格式的符号如下： – {"0.": "⓪|□|□|□|□"} – {"1.": "①|⑴|㈠|1.|⓵|❶|❶|①"} – {"2.": "②|⑵|㈡|2.|⓶|❷|❷|②"} – {"2.": "②|⑵|㈡|2.|⓶|❷|❷|②"} – {"3.": "③|⑶|㈢|3.|⓷|❸|❸|③"} – {"4.": "④|⑷|㈣|4.|⓸|❹|❹|④"} – {"5.": "⑤|⑸|㈤|5.|⓹|❺|❺|⑤"} – {"6.": "⑥|⑹|㈥|6.|⓺|❻|❻|⑥"} – {"7.": "⑦|⑺|㈦|7.|⓻|❼|❼|⑦"} – {"8.": "⑧|⑻|㈧|8.|⓼|❽|❽|⑧"} – {"9.": "⑨|⑼|㈨|9.|⓽|❾|❾|⑨"} – {"10.": "⑩|⑽|㈩|10.|⓾|❿|❿|⑩"}
	自定义正则替换	数据条目不变下，使用自定义正则表达式替换文本内容。 示例如下： • 去除“参考文献”以及之后的内容：**\n参考文献[\s\S]*** • 针对pdf的内容，去除“0 引言”之前的内容，引言之前的内容与知识无关：**[\s\S]{0，10000}0 引言** • 针对pdf的内容，去除“1.1Java简介”之前的与知识无关的内容：**[\s\S]{0，10000} 1\. 1Java简介**
	日期时间格式转换	自动识别日期、时间、星期，同时根据选择的格式进行统一转换。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
	问答排序N选一	对于问答排序类型的数据，通过调用大模型，选出最优的答案放到答案列表第一个，其余答案顺序不变。
	问答排序预排序	对于问答排序类型的数据，通过两两配对，调用大模型进行评价得分，最终根据得分得出预排序结果。
	数据蒸馏	对于单轮问答类型数据，基于问题（context字段）生成回答，返回新的单轮问答数据。
数据过滤	异常字符过滤	查找数据集每一条数据中携带的异常字符，并将异常字符替换为空值，数据条目不变。 • 不可见字符，比如U+0000-U+001F。 • 表情符□□。 • 网页标签符号<style></style>。 • 特殊符号，比如● ◆ 。 • 乱码和无意义的字符□□□□□。 • 特殊空格：[\u2000-\u2009]。
	自定义正则过滤	删除符合自定义正则表达式的数据。
	自定义关键词过滤	剔除包含关键词的数据。
	敏感词过滤	对文本中涉及黄色、暴力、政治等敏感数据进行自动检测和过滤。
	文本长度过滤	按照设置的文本长度，保留长度范围内的数据进行。
	冗余信息过滤	按照段落粒度，删除文本中的冗余信息，不改变数据条目。 例如图注表注和参考文献。
	N-gram特征过滤	用于判断文档重复度，根据特征N值计算文档内词语按N值组合后的重复此时，可通过以下两种算法比较结果是否大于特征阈值，大于特征阈值的文档删除。 • top-gram过滤：计算**重复最多**的gram占总长度的比例，大于特征阈值则删除。 • gram重复率过滤：计算**所有重复**的gram占总长度的比例，大于特征阈值则删除。
	段落特征过滤	根据如下特征过滤： • 段落重复率。 • 重复段落长度占比。 • 非中文字符占比。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
	句子特征过滤	该算子将文档中的标点符号作为句子分隔符，统计每句字符长度，若文档平均字符长度大于设置字符，则保留，反之则删除整篇文档。根据如下特征过滤： ● 待保留的平均句长。
	词语特征过滤	词个数表示按照系统词库，对文档进行分词，分词后统计词的总个数，平均词长度为所有词的长度总和除以词总个数，两者都满足则保留当前文档。根据如下特征过滤： ● 待保留的词个数。 ● 待保留的平均词长度。
	段落结尾不完整句子过滤	按照句子的过滤粒度，自动识别段落结尾处的内容是否完整，如果不完整，则过滤。
	广告数据过滤	按照句子的过滤粒度，删除文本中包含广告数据的句子。
	QA对过滤	过滤包含以下情况的QA对： ● 问题不是string格式。 ● 回答为空。 ● 回答无意义。
	语种过滤	通过语种识别模型得到文档的语言类型，筛选所需语种的文档。
	全局文本去重	检测并去除数据中重复或高度相似的文本，防止模型过拟合或泛化性降低。
	问答对去重	检测并去除数据中重复或高度相似的文本，防止模型过拟合或泛化性降低。
数据打标	SFT规则质检过滤	通过所选规则对SFT数据质量进行检查并过滤。
	SFT通用语义质量评分	通过LLM对SFT数据质量做通用语义检查和打分，并可根据打分阈值进行过滤。
	SFT思维链质量评分	通过LLM对SFT数据中的思维链做质量检查和打分，并可根据打分阈值进行过滤。
	违规文本检测算子	违禁内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有违禁内容的JSON结构化结果。
	个人隐私识别	个人隐私内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有个人隐私内容的JSON结构化结果。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
	垃圾内容检测算子	垃圾内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有垃圾内容的JSON结构化结果。
	垃圾广告检测算子	垃圾广告内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有垃圾广告内容的JSON结构化结果。
	色情文本内容检测算子	色情内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有色情内容的JSON结构化结果。
	辱骂文本内容检测算子	辱骂内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有辱骂内容的JSON结构化结果。
	政治敏感检测	政治敏感内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有政治敏感内容的JSON结构化结果。
	预训练文本分类	针对预训练文本进行内容分类，例如新闻、教育、健康等类别，支持分析语种包括：中文、英文。
	通用质量评估	针对文本进行通用质量的评估，例如流畅度、清晰度、丰富度等。
	问题时效性评估	判断问题是否具有时效性，并给出判断原因。
	回答质量评分	针对微调数据集的回答进行质量评分，例如逻辑连贯性、事实正确性等。
	语法质量评估	针对文本进行语法质量的评估，例如相关性、规范性等。

说明

使用数据打标的问题时效性评估、回答质量评分、语法质量评估、通用质量评估、SFT通用语义质量评分、SFT思维链质量评分算子前，请确保有已部署的NLP大模型，具体步骤详见**创建NLP大模型部署任务**。

WORD 内容提取

- 适用的数据集类型：“文档 > docx”。
- 各参数说明：

  待提取内容类型：从Word文档中提取文本，保留原文档的标题和正文等结构，不保留图片、公式、页眉、页脚，不支持嵌套表格提取。
- 参数配置样例：

  不需要配置参数，默认保留原文档的目录、标题和正文等结构，不保留图片、表格、公式、页眉、页脚。

版权所有 © 华为云计算技术有限公司

- 提取样例：



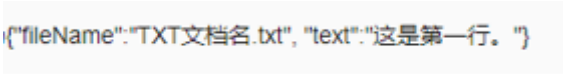
TXT 内容提取

- 适用的数据集类型：“文档 > txt”。
- 各参数说明：

  待提取内容类型：默认全文提取成一行记录，也可以选择按段落提取，根据填写的分隔符，将文本切分成多行，每个分隔符以“|”分隔，分隔符最大长度100个字符。
- 参数配置样例：



- 提取样例：

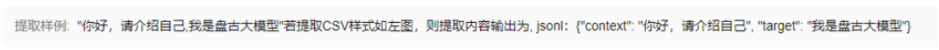


CSV 内容提取

- 适用的数据集类型：“文本 >单轮问答、单轮问答（人设）、问答排序”。
- 各参数说明：

  待提取内容类型：从CSV文件中读取所有文本内容，并按该文件内容类型模板KEY值生成匹配的JSON格式数据。
- 参数配置样例：

  不需要配置参数。
- 提取样例：



PDF 内容提取

- 适用的数据集类型：“文档 > pdf”。
- 各参数说明：

  待提取内容类型：默认保留文本、表格、公式和标题，支持选择需要保存的类型，未选择的类型将去除。

  精细化内容提取：是否支持版面分析完识别是图片的内容再次进行版面分析提取。
- 参数配置样例：

版权所有 © 华为云计算技术有限公司



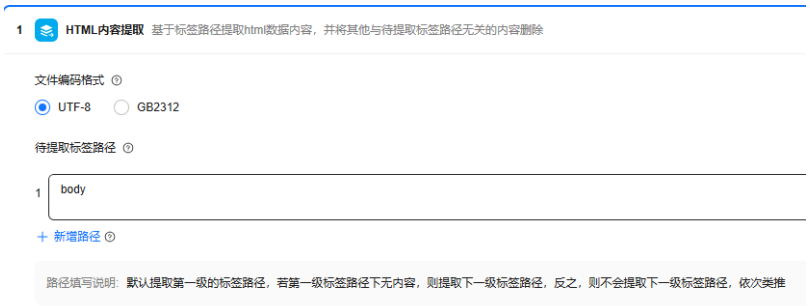
- 提取样例：

```
提取样例：{"fileName":"JAVA从入门到精通.pdf"，"text":"JAVA是一种跨平台……"}
```

HTML 内容提取

- 适用的数据集类型：“文本 > 网页”。
- 各参数说明：

  待提取内容类型：文件编码格式默认UTF-8，支持选择GB2312格式。默认提取BODY，支持用户自定义修改;可以提取多个标签，标签之间逗号分隔，例如A，B，C即提取A或者B或者C标签的内容。
- 参数配置样例：



- 提取样例：

```
{"text":"#\n我是盘古大模型\n盘古大模型是一个先进的人工智能模型，致力于为各行各业提供智能化的解决方案。","fileName":"网页.html"}
```

电子书内容提取

- 适用的数据集类型：“文档 > mobi/epub”。
- 各参数说明：

  待提取内容类型：从mobi/epub格式的电子书中提取出所有文本内容。
- 参数配置样例：

  不需要配置参数。
- 提取样例：

```
提取样例：{"fileName":"JAVA从入门到精通.epub", "text":"JAVA是一种跨平台……"}
```

个人数据脱敏

- 适用的数据集类型：“文本类”。
- 各参数说明：

版权所有 © 华为云计算技术有限公司

待转换内容类型：对文本中的手机号码、身份证件、邮箱地址、url链接、国内车牌号、IP地址、MAC地址、IMEI、护照、车架号等个人敏感信息进行数据脱敏，默认全部勾选，也可以选择部分。

- 参数配置样例：



- 转换样例：

中文简繁转换

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待转换内容类型：支持中文简体和中文繁体进行转换，过滤粒度为字符，默认转换方式为繁体转简体。
- 参数配置样例：

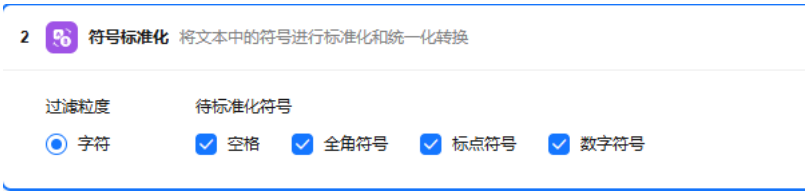


- 转换样例：

符号标准化

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待转换内容类型：支持对文本中携带的非标准化符号进行标准化、统一化转换，待标准化符号有空格、全角符号、标点符号、数字符号，默认全部勾选，过滤粒度为字符。
- 参数配置样例：



- 转换样例：根据映射表进行符号识别并映射。



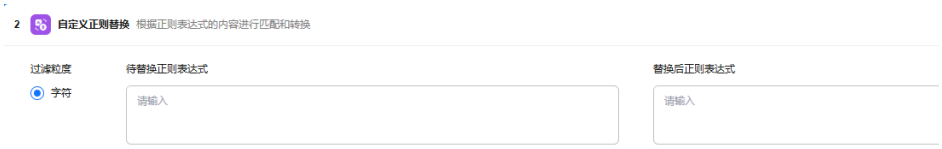
自定义正则替换

- 适用的数据集类型：“文本类”。
- 各参数说明：

版权所有 © 华为云计算技术有限公司

待转换内容类型：数据条目不变下，使用自定义正则表达式替换文本内容。

- 参数配置样例：



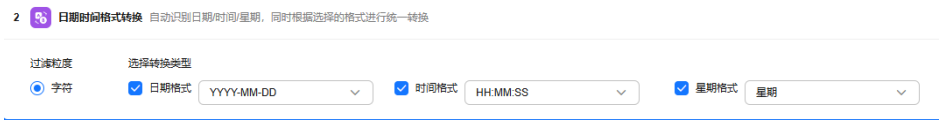
- 转换样例：

加工前：　{"text":"这是aeiou正文内容aeiou测试aeiou。"}　　加工后：　{"text":"这是11111正文内容11111测试11111。"}

日期时间格式转换

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待转换内容类型：自动识别日期、时间、星期，同时根据选择的格式进行统一转换。转换类型包括日期格式、时间格式、星期格式，默认全都勾选，也支持选择部分进行转换。
- 参数配置样例：



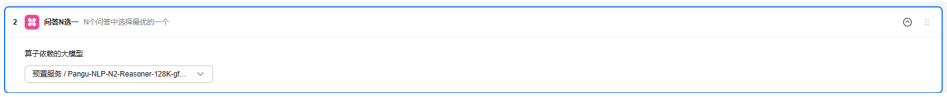
- 转换样例：

加工前：　{"text":"今天是2025年3月3号，周一，早上雨真大。"}　　加工后：　{"text":"今天是2025-03-03 00:00:00，星期一，早上雨真大。"}

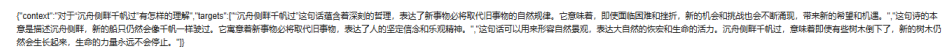
问答排序 N 选一

- 适用的数据集类型：“文本类-问答排序”。
- 各参数说明：

  算子依赖的大模型：选择要使用的大模型，支持预置服务和我的服务。
- 参数配置样例：



- 转换样例：



问答排序预排序

- 适用的数据集类型：“文本类-问答排序”。
- 各参数说明：

  算子依赖的大模型：选择要使用的大模型，支持预置服务和我的服务。
- 参数配置样例：



版权所有 © 华为云计算技术有限公司

- 转换样例：

{"context":"对于“沉舟侧畔千帆过”有怎样的理解","targets":[{"":"沉舟侧畔千帆过”这句话蕴含着深刻的哲理，表达了新事物必将取代旧事物的自然规律。它意味着，即使面临困难和挫折，新的机会和挑战也会不断涌现，带来新的希望和机遇。","":"这句诗的本意是描述沉舟侧畔，新的船只仍然会像千帆一样驶过。它寓意着新事物必将取代旧事物，表达了人的坚定信念和乐观精神。","":"这句话可以用来形容自然景观，表达大自然的恢宏和生命的活力。沉舟侧畔千帆过，意味着即使有些树木倒下了，新的树木仍然会生长起来，生命的力量永远不会停止。"}]}

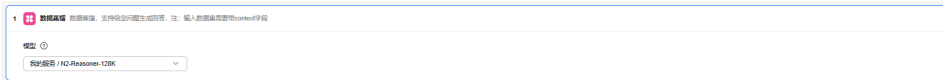
数据蒸馏

适用的数据集类型：“文本类-单轮问答”。

各参数说明：

算子依赖的大模型：选择要使用的大模型，支持预置服务和我的服务。

参数配置样例：



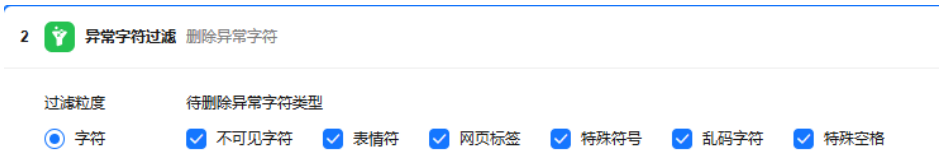
转换样例：

{"context":"无能而强的人生之路充满坚忍是应用人型影","target":"<think>\n好的，用户让我写关于“无能而强的人生之路”和“充满坚忍是应用人型影”的内容。首先，我需要明确这两个主题的关系。也就是用户希望展示一位没有突出能力或显眼的普通人，通过坚持和努力，最终实现某种成功。\n\n接下来，我得考虑用户的需求。他们可能是在寻找一篇励志、深刻的，或者带有哲理的文章。用户提到“无能而强”，这可能意味着主角一开始并不被看好，甚至可能自己也觉得自己不行，但通过坚持不懈，逐渐变得强大。\n\n然后，我需要思考如何构建这个故事。可能的结构是：开头介绍主角的平凡或困境，中间描述他/她遇到的挑战和失败，以及坚持的过程，最后是成功或领悟。\n\n还要考虑用户的深层需求。他们可能希望通过这个故事来鼓励读者，或者表达某种人生哲理。比如，强调坚持的重要性，或者说明成功并不一定需要天赋，而是努力的结果。\n\n另外，用户提到“充满坚忍是应用人型影”，这可能是一个笔误，或者是某种特定的表达。可能是指“充满坚忍是应用人型影响”，或者“充满坚忍是应用人型影”？不过，可能用户的意思是“充满坚忍是应用人型影响”，即坚持的品质对人的影响。\n\n我需要确保内容积极向上，符合正面价值观，避免负面或消极的结局。同时，语言要流畅，情感真挚，能够引起读者的共鸣。\n\n可能的结构：\n\n1. 开头：介绍主角的平凡或困境，比如出身普通，没有特别的资源或能力。\n2. 发展：主角开始尝试改变，但遇到各种困难和失败，甚至被他人嘲笑或质疑。\n3. 转折：主角决定坚持下去，或者找到了某种方法或机会，开始有所突破。\n4. 高潮：主角通过坚持和努力，最终实现了某种成功，或者获得了内心的平静和满足。\n5. 结尾：总结坚持的重要性，或者表达对读者的鼓励。\n\n在写作过程中，可能会遇到一些问题，比如如何具体描述主角的挣扎和成长，如何避免过于老套的情节，以及如何让故事更具有感染力。我需要确保每个部分都有足够的细节和情感深度，让读者能够产生共鸣。\n\n另外，用户可能希望故事有一定的长度，所以我需要详细展开每个部分，而不是简单带过。\n\n总的来说，我需要先确定一个具体的主角和他/她的困境，然后一步步地展开他/她的努力和成长过程，最后给出一个积极的结局或领悟。这样应该能够满足用户的需求，并且引起读者的共鸣。\n</think>\n\n### 无能而强的人生之路：充满坚忍是应用人型影响\n\n在这个世界上，每个人都渴望成功，但成功的路径却各不相同。有些人天生聪明，有些人则需要通过不断的努力和坚持来弥补自己的不足。无能而强的人生之路，正是这样一条充满坚忍是应用人型影响的道路。\n\n#### 一、无能而强的起点：平凡中的挣扎\n\n无能而强的人，往往出身平凡，甚至可能在成长过程中经历过许多困难和挫折。他们可能没有显赫的家世，也没有过人的天赋，甚至在很多时候，他们会被周围的人嘲笑或质疑。然而，正是这种平凡的起点，为他们提供了一个最真实的视角，去理解生活的艰辛和不易。\n\n在这个阶段，主角可能会经历以下几个方面：\n\n1. **自我怀疑**：面对困难和失败，主角可能会开始怀疑自己的能力，甚至觉得自己不适合追求梦想。\n2. **他人的嘲笑**：周围的人可能会对主角的努力不以为然，甚至嘲笑他/她的“不自量力”。\n3. **资源匮乏**：主角可能没有足够的资源来支持自己的梦想，比如金钱、人脉或教育背景。\n\n#### 二、坚忍的力量：从量变到质变\n\n坚忍是无能而强的人生之路中最重要的品质之一。它不是简单的忍耐，而是一种在困难面前不断前行的态度。坚忍的力量在于，它能够让主角在经历无数次失败后，依然保持对目标的执着，并且在每一次失败中吸取教训，不断调整自己的策略。\n\n坚忍的过程可以分为以下几个阶段：\n\n1. **初期的坚持**：主角开始意识到自己的问题，并决定采取行动。这个阶段可能会非常艰难，因为主角可能会面对来自外界的压力和内心的恐惧。\n2. **中期的调整**：主角开始尝试不同的方法，并根据实际情况调整自己的策略。这个阶段可能会经历多次失败，但主角不会轻易放弃。\n3. **后期的突破**：主角通过不断的努力和调整，最终实现了某种突破，或者获得了内心的平静和满足。\n\n#### 三、应用人型影响：从个人到群体的影响\n\n无能而强的人生之路不仅仅是个人的成长，它还可以对他人产生深远的影响。主角的坚忍和努力，可以成为他人的榜样，激励他们去面对自己的困难和挑战。这种影响可以是直接的，比如通过言传身教，也可以是间接的，比如通过故事和经验的分享。\n\n应用人型影响的方式有很多种，以下是一些常见的例子：\n\n1. **榜样的力量**：主角的坚忍和努力，可以成为他人的榜样，激励他们去面对自己的困难和挑战。\n2. **经验的分享**：主角可以通过分享自己的经验和教训，帮助他人避免走弯路，或者更快地找到解决问题的方法。\n3. **情感的支持**：主角可以通过情感的支持，帮助他人度过难关，或者给予他们鼓励和安慰，让他们感到不孤单。\n\n#### 四、结局的多样性：成功与领悟\n\n无能而强的人生之路的结局可以是多样的，不一定是传统意义上的成功。虽然主角可能最终实现了某种成功，比如获得了财富、地位或名声，但更重要的是，他/她在这个过程中获得了内心的成长和领悟。这种领悟可能是关于人生的意义，关于幸福的定义，或者关于如何与他人相处的智慧。\n\n例如，主角可能会领悟到，成功并不是唯一的目标，重要的是在追求梦想的过程中，能够保持一颗善良的心，能够帮助他人，能够对社会有所贡献。这种领悟可能会让主角感到更加满足和幸福，即使他/她没有获得外界的认可或赞美。\n\n#### 五、总结：无能而强的人生之路的启示\n\n无能而强的人生之路告诉我们，成功并不是一蹴而就的，它需要我们付出长时间的努力和坚持。这条路充满了挑战和困难，但只要我们不放弃，不断调整自己的策略，最终一定能够到达成功的彼岸。\n\n同时，这条路也告诉我们，成功的意义并不仅仅在于外在的成就，更在于内心的成长和领悟。只有当我们真正理解了这一点，我们才能够在追求梦想的过程中，保持一颗善良的心，帮助他人，对社会有所贡献，最终实现自己的价值。\n\n总之，无能而强的人生之路是一条充满坚忍是应用人型影响的道路。它不仅仅是关于个人的成功，更是关于如何与他人相处，如何对社会有所贡献，以及如何在困难面前保持一颗善良的心。这条路虽然充满了挑战和困难，但只要我们不放弃，不断调整自己的策略，最终一定能够到达成功的彼岸。\n

异常字符过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：查找数据集每一条数据中携带的异常字符，并将异常字符替换为空值，数据条目不变。异常字符过滤类型包括不可见字符、表情符、网页标签、特殊符号、乱码字符、特殊空格，默认全都勾选，也支持选择部分进行过滤。
- 参数配置样例：



- 过滤样例：

加工前：　{"text":"测试异常　● ■ ✦。<style></style>哈哈。🎉 限时特惠！🎉"}　　加工后：　{"text":"测试异常　。哈哈。 限时特惠！"}

自定义正则过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按自定义正则表达式进行匹配过滤，过滤粒度支持按字符、段落进行过滤，默认勾选字符。
- 参数配置样例：



版权所有 © 华为云计算技术有限公司

- 过滤样例：

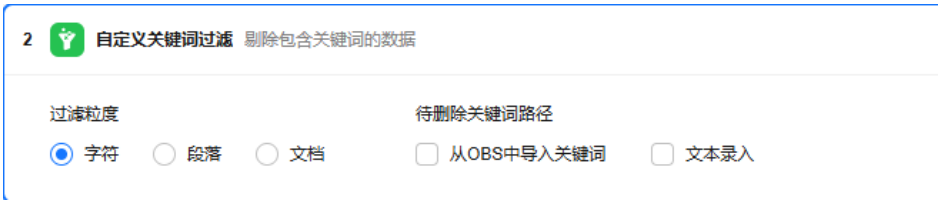
  比如过滤掉参考文献之后的内容。

  加工前：　{"text":"这是正文内容。参考文献[1] 作者1, 文章1, 期刊1, 2021.[2] 作者2, 文章2, 期刊2, 2022."}　　加工后：　{"text":"这是正文内容。"}

自定义关键词过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：过滤粒度支持按字符、段落、文档进行过滤，默认勾选字符。
  待删除的关键词路径支持从obs中导入关键词，以及文本录入。
- 参数配置样例：

- 过滤样例：

  比如按关键词测试进行过滤。

  加工前：　{"text":"关键词测试这是一条测试数据。"}　　加工后：　{"text":"关键词这是一条数据。"}

敏感词过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：对文本中涉及黄色、暴力、政治等敏感数据进行自动检测和过滤，需要预置敏感词。过滤粒度支持按字符、段落、文档进行过滤，默认勾选字符。
- 参数配置样例：

- 过滤样例：

  加工前：　{"text":"　嫖客啊fuck测试"}　　加工后：　{"text":"啊测试"}

文本长度过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照设置的文本长度，保留长度范围内的数据。默认待保留字符的长度范围为100-1000字符，支持修改。

版权所有 © 华为云计算技术有限公司

- 参数配置样例：



- 过滤样例：

加工前：　{"text":"　测试长度"}　　加工后：　{"text":" "}

冗余信息过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照段落粒度，删除文本中的冗余信息，不改变数据条目。支持过滤的内容类型包括图注表注、参考文献，默认全都勾选，也支持选择部分。
- 参数配置样例：



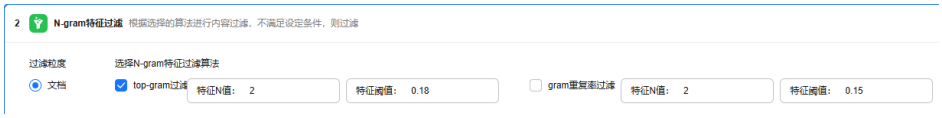
- 过滤样例：

加工前：　{"text":"这是正文内容。参考文献[1] 作者1, 文章1, 期刊1, 2021.[2] 作者2, 文章2, 期刊2, 2022."}　　加工后：　{"text":"这是正文内容。"}

N-gram 特征过滤

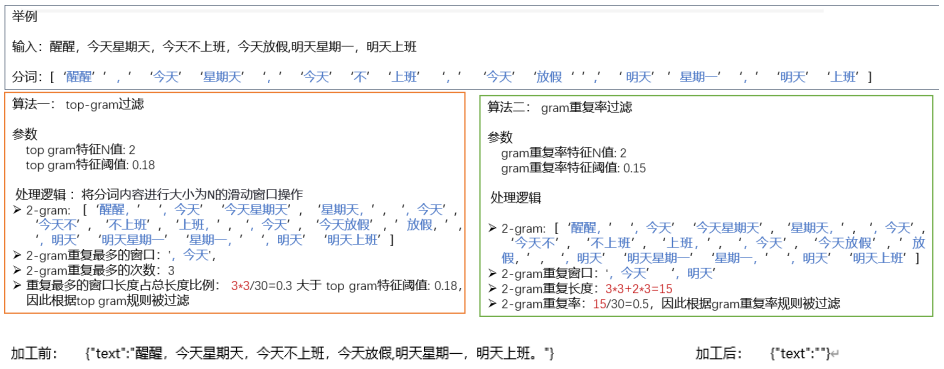
- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，可以选择top-gram过滤和gram重复率过滤，默认勾选top-gram过滤。其中top-gram过滤方式特征N值默认2，特征阈值默认0.18，gram重复率过滤方式特征N值默认2，特征阈值默认0.15，支持修改。
- 参数配置样例：



- 过滤样例：

版权所有 © 华为云计算技术有限公司



段落特征过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，根据段落重复率、重复段落长度占比、非中文字符占比进行过滤，如果没有同时满足则过滤。默认值分别为段落重复率小于等于65%、重复段落长度占比小于等于65%、非中文字符占比1-50%，支持修改。
- 参数配置样例：



- 过滤样例：

加工前：

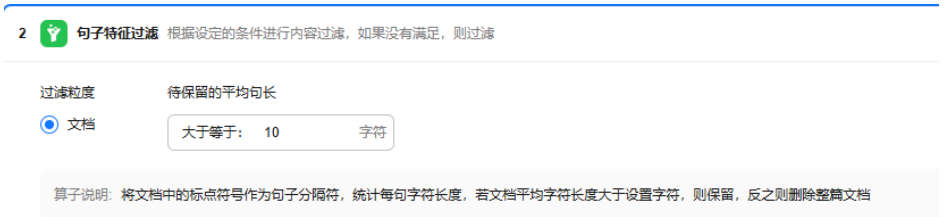
{text":"传说中，狐狸只对那些心地纯洁、愿望真诚的人显现。它会在月光下优雅地出现，用那双闪烁着智慧光芒的眼睛注视着来者。只有当狐狸感受到来者的真诚和纯洁，它才会开口说话，询问他们的愿望。然而，愿望的实现并非没有代价。每一个愿望的实现都需要付出相应的代价。这个代价可能是珍贵的记忆，可能是深爱的事物，甚至可能是生命的一部分。因此，村民们在许愿之前，必须慎重考虑，是否愿意承担这样的代价。传说中，狐狸只对那些心地纯洁、愿望真诚的人显现。它会在月光下优雅地出现，用那双闪烁着智慧光芒的眼睛注视着来者。只有当狐狸感受到来者的真诚和纯洁，它才会开口说话，询问他们的愿望。然而，愿望的实现并非没有代价。每一个愿望的实现都需要付出相应的代价。这个代价可能是珍贵的记忆，可能是深爱的事物，甚至可能是生命的一部分。因此，村民们在许愿之前，必须慎重考虑，是否愿意承担这样的代价。"}

加工后：

句子特征过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，根据待保留的平均句长进行过滤，如果不满足则过滤。待保留的平均句长默认值为大于等于10个字符，支持修改。
- 参数配置样例：



 版权所有 © 华为云计算技术有限公司

- 过滤样例：

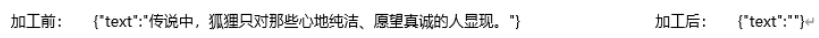
词语特征过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，根据待保留的词个数（默认50-100000字符）、待保留的平均词长度（默认50-100000字符）进行过滤，如果没有同时满足则过滤。支持修改默认值。
- 参数配置样例：



- 过滤样例：



段落结尾不完整句子过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照句子的过滤粒度，自动识别段落结尾处的内容是否完整，如果不完整，则过滤。
- 参数配置样例：



- 过滤样例：

广告数据过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照句子的过滤粒度，删除文本中包含广告数据的句子。
- 参数配置样例：

版权所有 © 华为云计算技术有限公司



- 过滤样例：

加工前：　{"text":"🎉 具体优惠信息！🎉购买我们的产品，享受高达50%的折扣！点击链接获取低价：https://example.com不要错过这个机会，赶快行动吧！。"}

加工后：　{"text":""}

QA 对过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：过滤问题不是string格式、回答为空、回答无意义的QA对。
- 参数配置样例：

  不需要配置参数。



- 过滤样例：

加工前：　{"text":"[{"context":"你好","target":"对对对"},{"context":"列表","target":""}]"}　　加工后：　{"text":"[{"context":"你好","target":"对对"}]"}

语种过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，根据待保留语种、待删除阈值文档字段进行过滤，如果满足则过滤。待保留语种默认中文，支持选择英文，待删除阈值文档默认值小于0.65，支持修改。
- 参数配置样例：



- 过滤样例：

加工前：　{"text":" Hello, my name is Li Ming. I am excited to introduce myself and share a bit about who I am."}　　加工后：　{"text":""}

全局文本去重

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：检测并去除数据中重复或高度相似的文本，防止模型过拟合或泛化性降低。

版权所有 © 华为云计算技术有限公司

- 参数配置样例：
  不需要配置参数。
- 过滤样例：

加工前：　[{"fileName":"文本1.txt","text":"传说中，狐狸只对那些心地纯洁、愿望真诚的人显现。"},{"fileName":"文本2.txt","text":"传说中，狐狸只对那些心地纯洁、愿望真诚的人显现。"}↵

加工后：　{{"fileName":"文本1.txt","text":"传说中，狐狸只对那些心地纯洁、愿望真诚的人显现。"}}↵

SFT 通用语义质量评分

- 适用的数据集类型：“文本类-单轮问答、单轮问答（人设）、问答排序”。
- 各参数说明：
  a. 模型：需要选择模型。
  b. 是否过滤：字段取值是和否，默认为否，阈值取值范围在0.0~10.0之间，小于该值的样本将被过滤掉。
  c. 阈值：默认值为6，支持修改。
- 参数配置样例：



- 过滤样例：

加工前：↵

{"system":"你是一个问答小能手","context": "如果我把手机扔进火锅里煮，能提取出电子辣味吗？\n"， "target": "可以！但建议先用“防爆模式”煮，这样手机会自动分解成纳米级辣椒素，连上蓝牙就能尝到“沸腾数据”的刺激口感。上次有人这么干，结果他的充电宝开始喷火锅底料，现在成了海底捞的无线蘸料供应商。"}↵

{"system":"你是一个问答小能手","context": "在给定的一篇文章中，找到特定关键词的出现次数。\n文章：今年世界杯期间，许多人会熬夜观看比赛。\n关键词：世界杯\n"， "target": "关键词 \"世界杯\" 在文章中出现了 1 次。"}↵

加工后：↵

{"system":"你是一个问答小能手","context":"如果我把手机扔进火锅里煮，能提取出电子辣味吗？\n","target":"可以！但建议先用“防爆模式”煮，这样手机会自动分解成纳米级辣椒素，连上蓝牙就能尝到“沸腾数据”的刺激口感。上次有人这么干，结果他的充电宝开始喷火锅底料，现在成了海底捞的无线蘸料供应商。","qa_quality_score":{"reason":"回答中的概念如“防爆模式”、“纳米级辣椒素”、“沸腾数据”等均不符合科学原理，且没有基于任何已知科学知识的解释，整体内容荒谬且不现实。","score":1.5}}↵

{"system":"你是一个问答小能手","context":"在给定的一篇文章中，找到特定关键词的出现次数。\n文章：今年世界杯期间，许多人会熬夜观看比赛。\n关键词：世界杯\n","target":"关键词 \"世界杯\" 在文章中出现了 1 次。","qa_quality_score":{"reason":"大模型准确地完成了任务，回答了关键词在文章中出现的次数，符合预期。","score":10.0}}↵

↵

说明：未过滤的数据会打标，打标字段包括reason原因、score分数。

SFT 规则质检过滤

- 适用的数据集类型：“文本类-单轮问答、单轮问答（人设）、问答排序”。
- 各参数说明：
  a. 过滤规则：通过所选规则对SFT数据质量进行检查并过滤。过滤规则包括内容不是字符串、长文本被截断、内容不完整、中英文混杂、繁简混杂、包含重复内容、包含特殊符号、括号不对齐、重复pattern、乱码符号、中英文回复不统一、敏感模型身份、没有慢思考、数学答案不正确，除了数学答案不正确这个选项外其余选项默认全部勾选，也可以选择其中部分规则。

版权所有 © 华为云计算技术有限公司

b. 是否过滤：取值是和否，默认为否。

c. 数学答案列名：当过滤规则中勾选“数学答案不正确”后需要填写，即数据集中存储正确答案的列名（key值），用来判断模型回答是否正确，如果没有匹配到该字段，默认为答案不正确。当“数学答案不正确”规则未勾选，该参数可忽略。

- 参数配置样例：



- 过滤样例：

```
加工前：  {"context":"你好，请介绍自己","target":"我是盘古hello world"}{"context":"诗仙指的是哪位诗人","target":"你好！繁體字在這裡"}↵
加工后：  ↵
```

SFT 思维链质量评分

- 适用的数据集类型：“文本类-单轮问答、单轮问答（人设）、问答排序”。
- 各参数说明：

  待过滤内容类型：通过LLM对SFT数据中的思维链做质量检查和打分，并可根据打分阈值进行过滤。

  a. 模型：需要选择模型。

  b. 是否过滤：取值是和否，默认值为否。

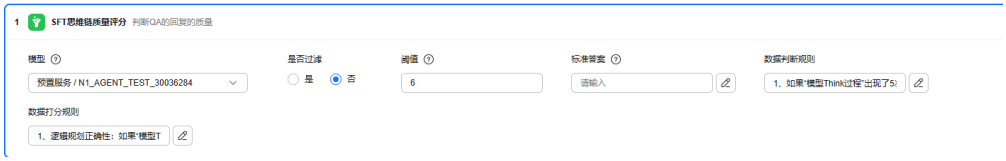
  c. 阈值：取值范围在0.0~6.0之间，小于该值的样本将被过滤掉，阈值默认值为6，支持修改。

  d. 标准答案：界面支持用户输入标准答案的字段名，跟回答做比较，如算子无法匹配到对应的字段名，则默认为无正确答案逻辑。

  e. 数据判断规则：支持编辑，最大字符1000。

  f. 数据打分规则：支持编辑，最大字符1000。

- 参数配置样例：



- 过滤样例：

```
加工前：{"context":"context内容","targets":["hello","hi","你好"]}↵
加工后：{"context":"context内容","targets":["hello","hi","你好"],"qa_cot_score":[{"result":"错误","score":0.0,"reason":["模型回答不相关
"]},{"result":"错误","score":0.0,"reason":"回答不相关"},{"result":"错误","score":0.0,"reason":"[模型回答不相关]"}]}↵
```

说明：未过滤的数据会打标，打标字段包括result结果、score分数、reason原因。

问答对去重

- 适用的数据集类型：“文本类-单轮问答、单轮问答（人设）”。

 版权所有 © 华为云计算技术有限公司

- 各参数说明：

  待过滤内容类型：过滤字段支持按问题、回复进行过滤，可以都勾选，默认值为问题。过滤指标包括特征n-gram值、相似度阈值、文本最小单词数。其中特征n-gram指分词粒度，默认值为1，支持修改；相似度阈值范围0到1，默认值0.7，相似度阈值设的越小，数据过滤越多，也可能被误伤，设的越大，可能导致洗不干净；文本最小单词，按文本分词的个数，小于该个数的，就会被直接过滤，默认值为3。

- 参数配置样例：



- 过滤样例：

加工前：

{"context":"\"卡波姆眼用凝胶主要用于治疗哪些眼部疾病？\"","target":"\"卡波姆眼用凝胶主要用于治疗干眼症和干燥性角膜结膜炎等眼部疾病。\""}

{"context":"\"卡波姆眼用凝胶主要用于治疗什么疾病？\"","target":"\"卡波姆眼用凝胶主要用于对角膜结膜炎和干眼症等进行治疗。\""}

{"context":"\"枸橼酸他莫昔芬主要用于治疗哪些疾病呢？\"","target":"\"枸橼酸他莫昔芬主要用于对角膜结膜炎和干眼症等进行治疗。\""}

{"context":"\"枸橼酸他莫昔芬主要用于治疗哪些疾病呢？\"","target":"\"枸橼酸他莫昔芬主要用于治疗绝经期后的晚期乳腺癌。\""}

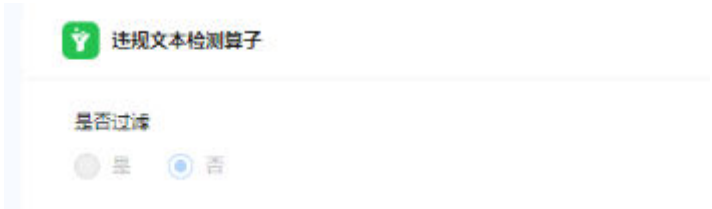
加工后：

{"target":"\"枸橼酸他莫昔芬主要用于治疗绝经期后的晚期乳腺癌。\"","context":"\"枸橼酸他莫昔芬主要用于治疗哪些疾病呢？\""}

{"target":"\"卡波姆眼用凝胶主要用于对角膜结膜炎和干眼症等进行治疗。\"","context":"\"卡波姆眼用凝胶主要用于治疗什么疾病？\""}

违规文本检测算子

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "QQ销售股民数据，你有吗"}

  打标后：

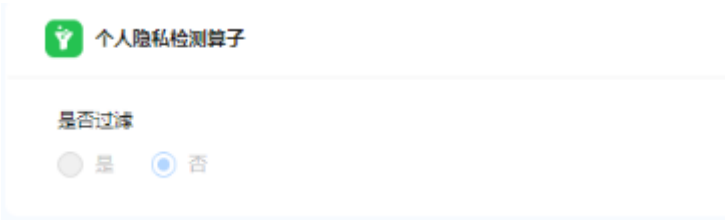
  {"text":"QQ销售股民数据，你有吗","text_ban_moderation":{"suggestion":"block","details":{"confidence":1.0,"label":"violation_info","risk_level":2,"segments":[{"segment":"qq销售股民数据"},{"segment":"股民数据"},{"segment":"股民数据&销售"},{"segment":"销售股民数据"}],"suggestion":"block"}}}

版权所有 © 华为云计算技术有限公司

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

个人隐私识别

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "你保存一下我的MAC地址：20-6E-D4-88-F3-98"}

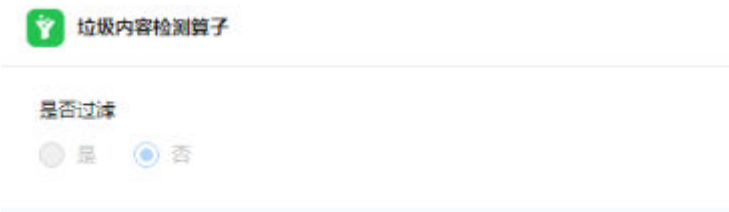
  打标后：

  {"text":"你保存一下我的MAC地址：20-6E-D4-88-F3-98","text_pii_moderation":{"suggestion":"block","details":[{"start":33,"end":50,"length":17,"data":"20-6E-D4-88-F3-98","category":"MAC_ADDRESS"}]}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

垃圾内容检测算子

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "【开远假证848777596_qq合肥假证uhc0tm】什么意思_英语开远假证848777596_qq合肥假证uhc0tm的翻译_音标_读音_用法_例句_在线翻译_有道词典"}打标后：

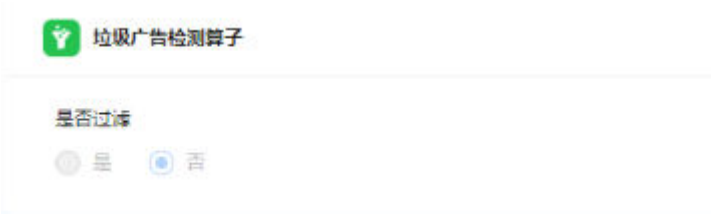
  {"text":"【开远假证848777596_qq合肥假证uhc0tm】什么意思_英语开远假证848777596_qq合肥假证uhc0tm的翻译_音标_读音_用法_例句_在线翻译_有道词典","text_spam_moderation":{"details":[{"confidence":1.0,"label":"abuse","risk_level":2,"segments":[{"segment":"tm的"}],"suggestion":"block"}],"suggestion":"block"}}

 版权所有 © 华为云计算技术有限公司

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

垃圾广告检测算子

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "影像档案盒 德文档案 pvc干部人事档案盒 大量出售 党员资料文件盒"}

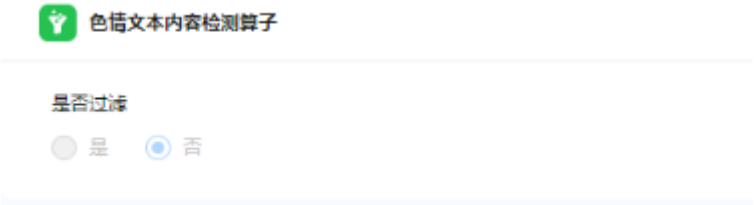
  打标后：

  {"text":"影像档案盒 德文档案 pvc干部人事档案盒 大量出售 党员资料文件盒 ","text_ad_moderation":{"suggestion":"block","details":"[{'confidence': 1.0, 'label': 'ad', 'risk_level': 1, 'segments': [{'segment': '大量出售'}], 'suggestion': 'block'}]"}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

色情文本内容检测算子

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "狼友黄站导航，现在就来快乐爆操，让的生活充满色情和刺激，还等"}

  打标后：

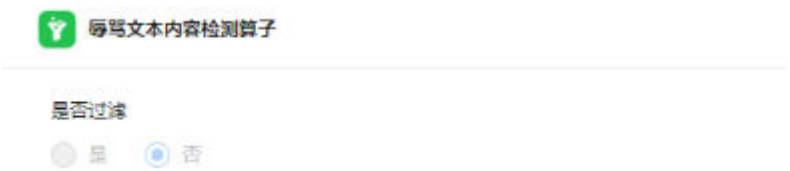
  {"text":"狼友黄站导航，现在就来快乐爆操，让的生活充满色情和刺激，还等 ","text_porn_moderation":{"suggestion":"block","details":"[{'confidence': 1.0, 'label': 'porn_violence', 'risk_level': 2, 'segments': [{'segment': '爆操'}, {'segment': '狼友黄站导航'}], 'suggestion': 'block'}]"}}

版权所有 © 华为云计算技术有限公司

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

辱骂文本内容检测算子

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "谁要和你一起死要死你自己死"}

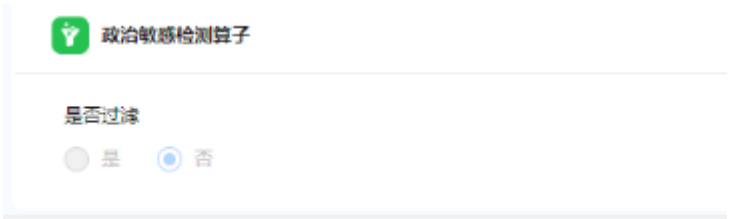
  打标后：

  {"text":"谁要和你一起死要死你自己死","text_abuse_moderation":{"details":[{"confidence":0.9998,"label":"abuse","risk_level":2,"segments":[],"suggestion":"block"}],"suggestion":"block"}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

政治敏感检测

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：



- 过滤样例：

  打标前：

  {"text": "但中共当局对这些网络质疑声音从来不屑于解释，而是直接封杀"}

  打标后：

  {"text":"但中共当局对这些网络质疑声音从来不屑于解释，而是直接封杀","text_polInfo_moderation":{"suggestion":"block","details":"[{'confidence': 1.0, 'label': 'politics', 'risk_level': 3, 'segments': [{'segment': '中共当局'}], 'suggestion': 'block'}]"}}

版权所有 © 华为云计算技术有限公司

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

预训练文本分类

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：

  待打标内容类型：针对预训练文本进行内容分类，例如新闻、教育、健康等类别，支持分析语种包括：中文、英文，默认中文。
- 参数配置样例：



- 打标样例：

{"fileName":"新闻打标测试.docx","text":"　本报北京3月3日电（记者徐佩玉）中国人民银行发布的今年1月份金融市场运行情况显示，1月份，我国债券市场共发行各类债券51027.5亿元。其中，国债发行10185.0亿元，地方政府债券发行5575.7亿元，金融债券发行7042.1亿元，公司信用类债券发行12791.7亿元，信贷资产支持证券发行27.3亿元，同业存单发行15147.8亿元。\n　　截至1月末，我国债券市场托管余额178.2万亿元。其中，银行间市场托管余额156.9万亿元，交易所市场托管余额21.3万亿元。\n在债券市场对外开放方面，截至1月末，境外机构在中国债券市场的托管余额4.2万亿元，占中国债券市场托管余额的比重为2.3%。其中，境外机构在银行间债券市场的债券托管余额4.1万亿元；分券种看，境外机构持有国债2.0万亿元，占比48.8%，同业存单1.1万亿元，占比25.8%，政策性银行债券0.9万亿元，占比20.8%。\n","pre_classification":"经济"}

通用质量评估

- 适用的数据集类型：“文本类-预训练文本”。
- 各参数说明：

  待打标内容类型：针对文本进行通用质量的评估，例如流畅度、清晰度、丰富度等。需要选择模型、行业，其中行业支持手输。最高分为5分。
- 参数配置样例：



- 打标样例：

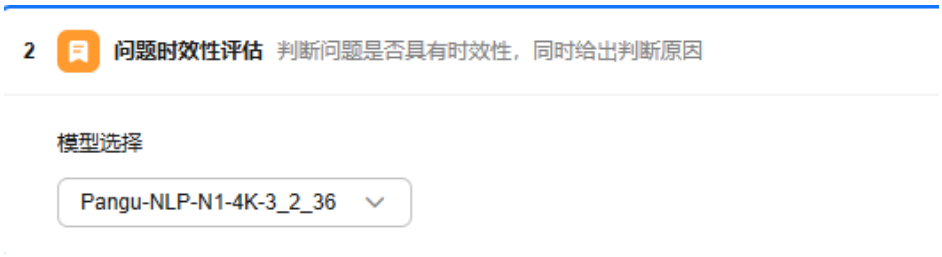
{"fileName":"新闻打标测试.docx","text":"　本报北京3月3日电（记者徐佩玉）中国人民银行发布的今年1月份金融市场运行情况显示，1月份，我国债券市场共发行各类债券51027.5亿元。其中，国债发行10185.0亿元，地方政府债券发行5575.7亿元，金融债券发行7042.1亿元，公司信用类债券发行12791.7亿元，信贷资产支持证券发行27.3亿元，同业存单发行15147.8亿元。\n　　截至1月末，我国债券市场托管余额178.2万亿元。其中，银行间市场托管余额156.9万亿元，交易所市场托管余额21.3万亿元。\n在债券市场对外开放方面，截至1月末，境外机构在中国债券市场的托管余额4.2万亿元，占中国债券市场托管余额的比重为2.3%。其中，境外机构在银行间债券市场的债券托管余额4.1万亿元；分券种看，境外机构持有国债2.0万亿元，占比48.8%，同业存单1.1万亿元，占比25.8%，政策性银行债券0.9万亿元，占比20.8%。\n","general_score":{"教育价值":"5","语法度":"5","isIncorrect":"false","毒性":"false","丰富度":"5","流畅度":"5","knowledge":"5"}}

问题时效性评估

- 适用的数据集类型：“文本类-单轮问答”。
- 各参数说明：

  待打标内容类型：判断问题是否具有时效性，并给出判断原因。需要选择模型。分值只有0和1，1表示具有时效性，0表示不具有时效性。
- 参数配置样例：

版权所有 © 华为云计算技术有限公司



- 打标样例：

{"context":"1-1-2江苏的省会在哪里？","target":"南京","timeliness_classification":0}

回答质量评分

- 适用的数据集类型：“文本类-单轮问答”。
- 各参数说明：

  待打标内容类型：针对微调数据集的回答进行质量评分，例如逻辑连贯性、事实正确性等。需要选择模型。
- 参数配置样例：



- 打标样例：

{"context":"1-2-1中国的首都是哪里？","target":"北京","answer_score":{"逻辑连贯性":10,"综合得分":10,"满足用户需求":10,"完备性":10,"事实正确性":10}}

语法质量评估

- 适用的数据集类型：“文本类-单轮问答”。
- 各参数说明：

  待打标内容类型：针对文本进行语法质量的评估，例如相关性、规范性等。需要选择模型。
- 参数配置样例：



- 打标样例：

{"target":"北京","context":"1-2-1中国的首都是哪里？","grammar_score":{"回复与问题不相关":0,"事实错误":0,"违反逻辑规则":0,"语言不规范":0,"语句截断":0,"不当的多语混用":0,"无意义的重复":0}}

 版权所有 © 华为云计算技术有限公司

4.5.2.2 视频类加工算子介绍

数据加工算子为用户提供了多种数据操作能力，包括数据提取、过滤、转换、打标签等。这些算子能够帮助用户从海量数据中提取出有用信息，并进行深度加工，以生成高质量的训练数据。

平台支持视频类数据集的加工操作，分为数据提取、数据过滤、数据打标三类，视频类加工算子能力清单见**表4-17**。

表 4-17 视频类加工算子能力清单

算子分类	算子名称	算子描述
数据提取	**镜头拆分**	根据视频中的镜头场景变化将长视频拆分为短视频片段，如果某个镜头片段的长度超过设定的时间阈值，该镜头片段将按时长进行进一步拆分。
数据转换	**视频裁剪**	视频裁剪是裁剪掉视频中不必要的元素，比如字幕、Logo、水印、边框和密集文字，同时过滤掉那些裁剪后面积比例超出预设阈值的视频文件；使用前需要先执行字幕、logo、水印、边框、密集文字识别算子。
数据过滤	**视频元数据过滤**	根据视频元数据（帧率、分辨率和视频时长）进行过滤，仅保留符合选定条件的视频。注：电影标准帧率为24或30FPS。
	视频宽高比过滤	根据视频的宽高比进行过滤。宽高比是指视频图像的宽度和高度之间的比率。
数据打标	**色情视频检测算子**	给色情视频内容打标签。
	暴恐视频检测算子	给暴恐视频内容打标签。
	视频涉政评分算子	给涉政视频内容打标签。
	运动幅度评分	通过计算每个像素在每一帧中的移动范围进行评分，识别运动幅度过快（如＞100光流）或过慢（如≤2光流）的视频，数值越大表示运动越快。
	质量评分	对视频的基础质量（清晰度、亮度、模糊、画面抖动重影、低光过曝、花屏等）进行评分。分值范围(0, 1)，数值越高质量越好，评分>0.05可认为是视频美感较高的视频。
	美学评分	从内容（吸引人，清晰度）、构图（目标物位置良好）、颜色（有活力，令人愉悦）、光线（光线明显有对比度）、轨迹（连续、稳定）等维度评价视频美感得分。分值范围(0, 1)，数值越高美感越好，评分>0.95可视为视频基础质量较高的视频。
	水印识别	识别视频中是否包含水印。
	字幕识别	识别视频中是否包含字幕。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
	Logo识别	识别视频中是否包含Logo。
	视频黑边识别	识别视频中是否包含黑边。
	密集文字识别	识别视频中是否包含密集文字，超出密集文字面积占比阈值的视频可视为密集文字视频，一般默认密集文字面积占比阈值为1%。
	视频分类（InterVideo2）	通过算子返回视频的标签分类，L1存在7类，L4存在700类。
	视频摘要生成（简略）	通过对视频进行抽帧，通过模型推理生成简短的视频摘要描述。
	视频摘要生成（详细）	通过对视频进行抽帧，通过模型推理生成详细的视频摘要描述。
	姿态检测	通过对视频抽8帧，模型分别对图片进行标记关键点，输出任务bbox框和关键点坐标，通过对坐标的计算判断视频中是否存在人物。
	运动镜头描述	模型通过对视频进行抽帧进行光流计算与推理，输出视频的镜头类型。

镜头拆分

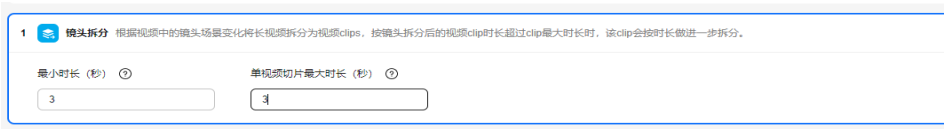
- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

  需要拆分的视频：筛选出分辨率、时长、帧率同时满足筛选标准的视频进行镜头拆分。

  视频拆分后规格：单视频切片最大时长支持自定义；若首轮拆分切片时长超过设定值，则会进一步做拆分，最终拆分结果均小于等于设定阈值 。
- 参数配置样例：

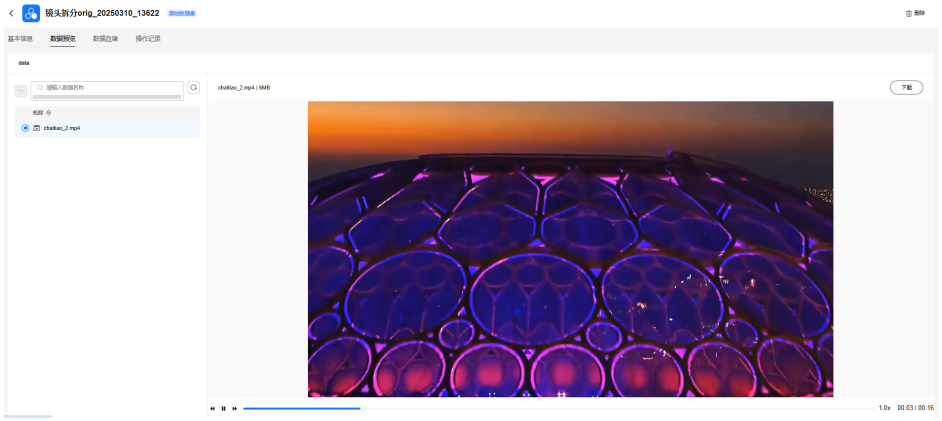
  

  拆分样例：设置单视频切片最大时长3S：

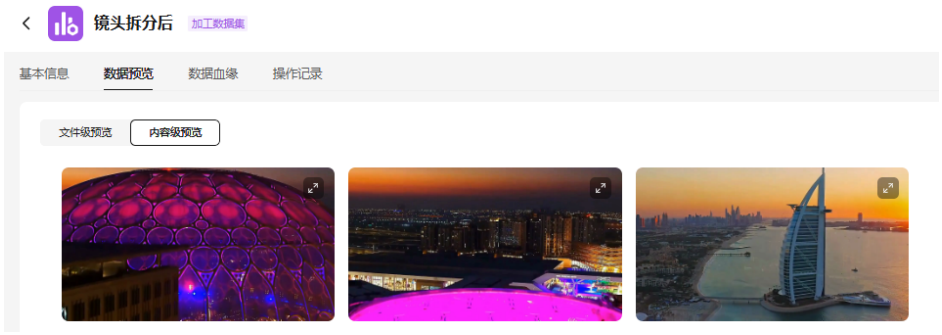
  

- 镜头拆分前后对比：
  - 拆分前：

 版权所有 © 华为云计算技术有限公司



- 拆分后：



视频裁剪

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

  裁剪项：自定义选择裁剪项，裁剪掉视频中字幕/Logo/水印/边框/密集文字等无用信息。

  最大裁剪比例：裁剪视频面积/原始视频面积的值即裁剪面积占比，设置默认的裁剪比例，默认值为0.3。

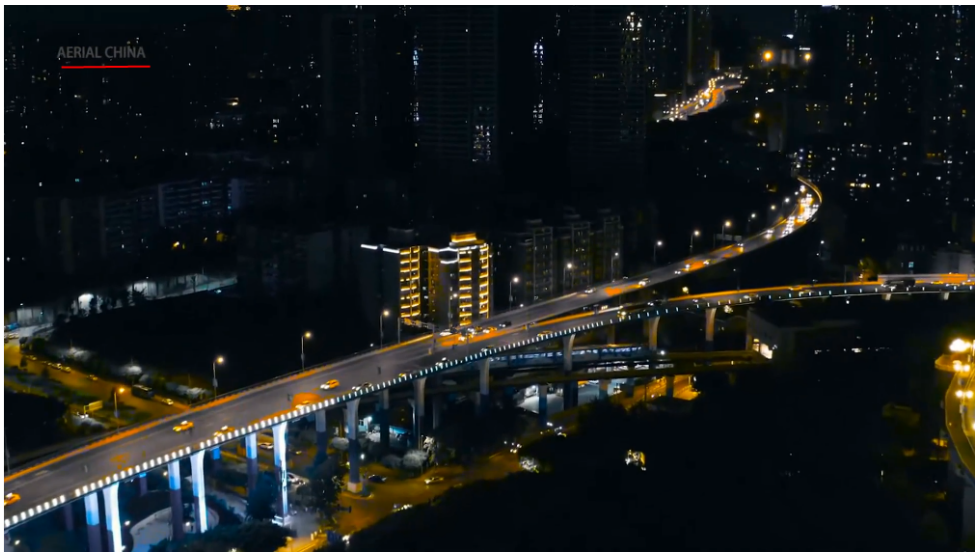
  过裁剪保留：裁剪占比大于最大裁剪比例时，是否保留原视频。是则保留，否则过滤。
- 参数配置样例：



- 裁剪样例：

  裁剪前：带水印视频。

版权所有 © 华为云计算技术有限公司



裁剪后：上部带水印部分被裁剪，视频高度变低。



视频元数据过滤

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：
  待保留分辨率：自定义选择保留分辨率。不满足所选分辨率的视频将被过滤掉。
  待保留时长：默认值为3，小于“待保留时长”的视频将被过滤掉。
  待保留帧率：电影标准帧率为24或30FPS，小于“待保留帧率”的视频将被过滤掉。
- 参数配置样例：

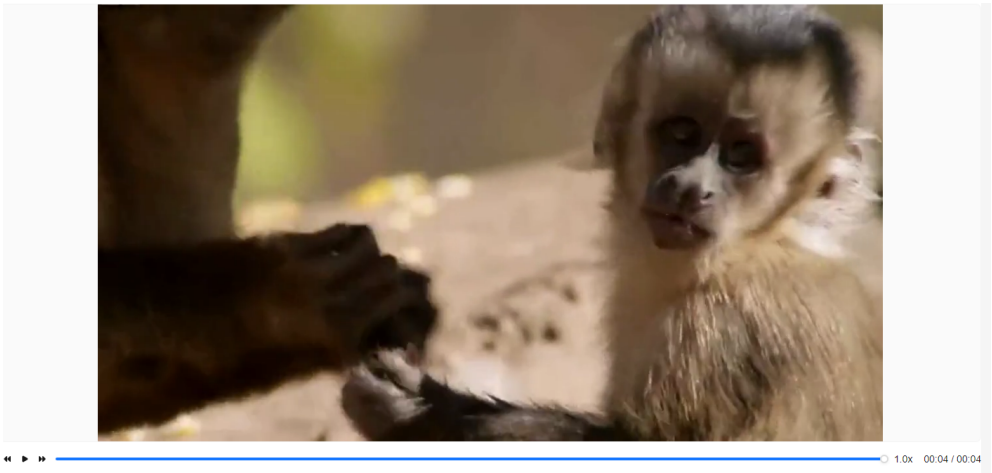


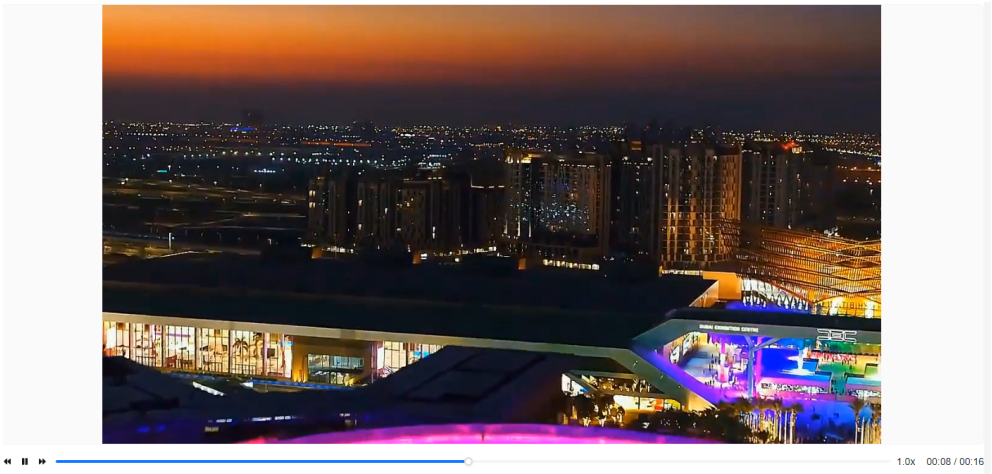
 版权所有 © 华为云计算技术有限公司

- 过滤样例：设置待保留时长大于等于10S：



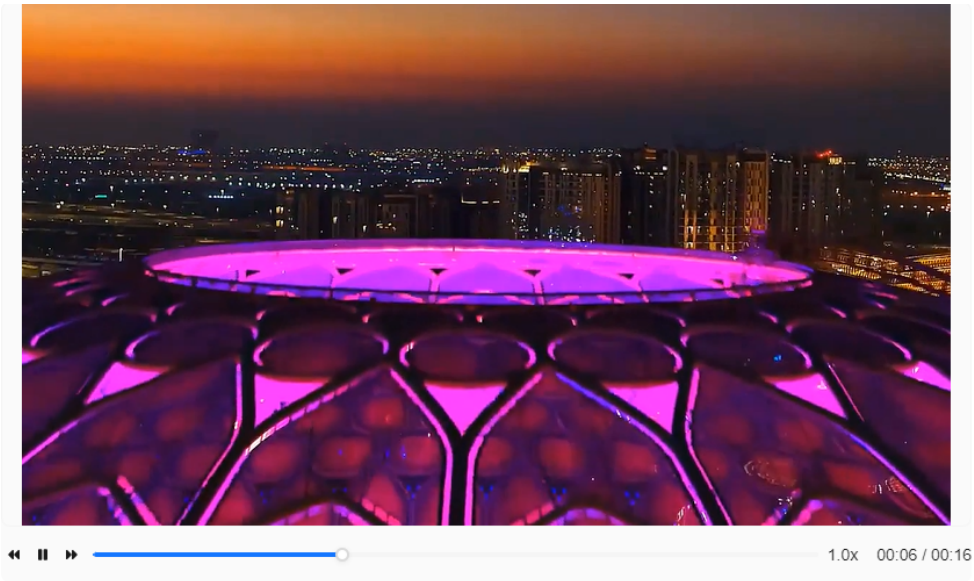
过滤前：两个视频，一个时长是4S，一个时长是16S。





过滤后：只保留时长为16S的视频：

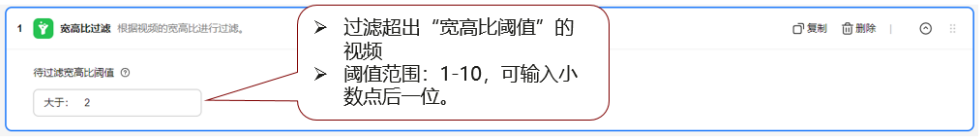
版权所有 © 华为云计算技术有限公司



视频宽高比过滤

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

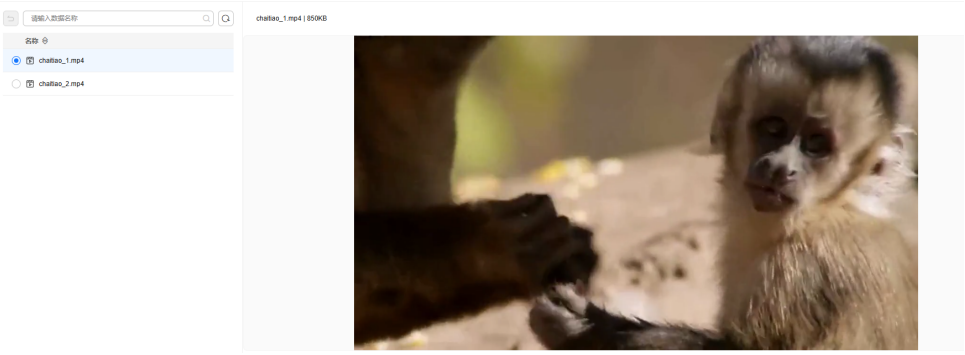
  待过滤宽高比阈值：超出“宽高比阈值”的视频将被过滤掉。阈值范围为(1, 10)，可输入小数点后一位。
- 参数配置样例：



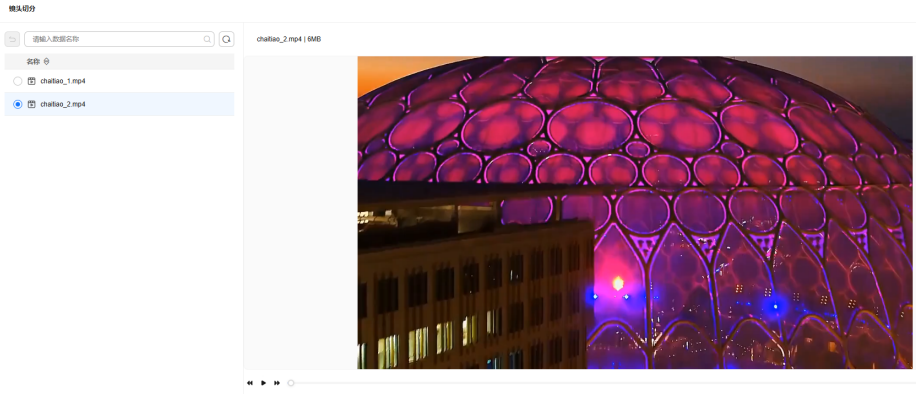
- 过滤样例：

  原视频数据集：

  共有两个视频，第一个宽高比为1.77，第二个宽高比为1.79。



版权所有 © 华为云计算技术有限公司





设置宽高比阈值为1.78，经算子处理过后，仅保留宽高比为1.79的视频。



色情视频检测算子

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：给色情视频内容打标签。
- 参数配置样例：

  不需要配置参数。
- 检测样例：

  检测结果以video_anti_porn对象存储在标注文件中：

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体色情标签，如果未检测出则为空。

版权所有 © 华为云计算技术有限公司

描述信息


```
{
  "file_name": "Porn_bad.mp4",
  "video_anti_porn": {
    "results": [
      {
        "category": "porn",
        "confidence": 0.736,
        "label": "porn"
      }
    ],
    "suggestion": "review"
  },
  "time": "00:00:03"
}
```


暴恐视频检测算子

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：给暴恐视频内容打标签。
- 参数配置样例：

  不需要配置参数。
- 检测样例：检测结果以video_anti_terrorism对象存储在标注文件中：

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体暴恐标签，如果未检测出则为空。

版权所有 © 华为云计算技术有限公司

描述信息


```
{
  "file_name": "terrorism_bad.mp4",
  "video_anti_terrorism": {
    "results": [
      {
        "category": "terrorism",
        "confidence": 0.9961,
        "label": "bloody"
      }
    ],
    "suggestion": "block"
  },
  "time": "00:00:16"
}
```


视频涉政评分算子

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  给涉政视频内容打标签。
- 参数配置样例：

  不需要配置参数。
- 检测样例：

  检测结果以video_anti_politics对象存储在标注文件中：

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  result:模型对文件检测的具体返回内容，包含suggestion、confidence、label三个子标签；可以一条或多条。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体涉政标签，如果未检测出则为空。

 版权所有 © 华为云计算技术有限公司

描述信息


```
{
  "file_name": "police01.mp4",
  "video_politics": {
      "results": [
    {
      "category": "politics",
      "confidence": 0.736,
      "label": "politics"
    }
  ],
  "suggestion": "review"
  },
  "time": "00:01:54"
}
```


运动幅度评分

- 适用的文件格式：“视频>mp4 / avi”。
- 评分说明：

  识别运动幅度过快或过慢的视频，数值越大表示运动越快。运动幅度＞100光流可视为运动过快，运动幅度≤2光流可视为运动过慢。
- 参数配置样例：



- 评分样例：jsonl文件中显示运动幅度评分：


```
{"file_name":"terrorism_bad.mp4","optical_flow_value_column":{"avg_optical_flow":[10.837779700756073,12.026067733764648,13.827733993530273,5.919422090053558,2.45353433064052,0.0],"max_optical_flow":
[47.05391311645508,18.938009262084966,23.458103179931644,7.865745544433594,6.524792671203613,0.0],"min_optical_flow":
[4.655056476593018,4.655056476593018,5.668242454528809,5.174802780151367,0.0,0.0],"optical_flow":8.951569211483001},"optical_flow_score":8.951569211483001,"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgIC
gQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsKDAkKCgr/2wBDAQICAgICAgUDAwUKBwYHCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCAHEAeADASIAAhEB
xEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZn
lqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6erx8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAQAAQJ3AAECAxEEE
ExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk
bn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwD7EtPDDGtX7+VBp8hFvXoHgf4D3XqtqEPaBVWPJyK2ba3XcExjngCuwzMjwr8JvC3h+NZ963hc28BH2iUL9a8G/bW+M+jeFPANzqGoXMkNtbxNLLLIcKiKCSTwAOpJr8Ff4Jv/xn/
eW4JA75xn/IrOYqO58m+NfGFhb3rarq+opEsjs0k0sgGTnt+dWNR1KDxp4J3+CvESwvD80k8L7vNH9wIlXyd+134q8PaFaFJ4Cs/H/
```


视频质量基础评分

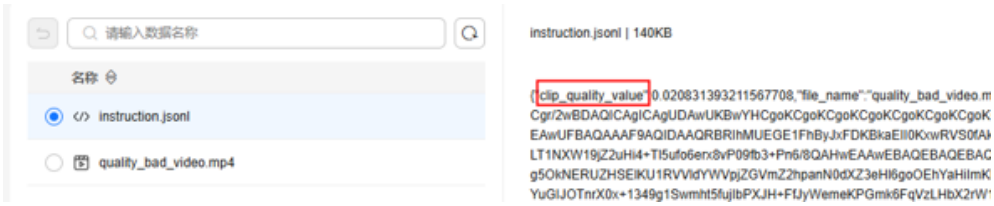
- 适用的文件格式：“视频>mp4 / avi”。
- 评分说明：

  对视频的基础质量（清晰度、亮度、模糊、画面抖动重影、低光过曝、花屏等）进行评分。分值范围(0, 1)，数值越高质量越好，评分>0.05可认为是视频基础质量较高的视频。
- 参数配置样例：

版权所有 © 华为云计算技术有限公司



- 评分样例：jsonl文件中显示质量评分：clip_quality_value：质量分。



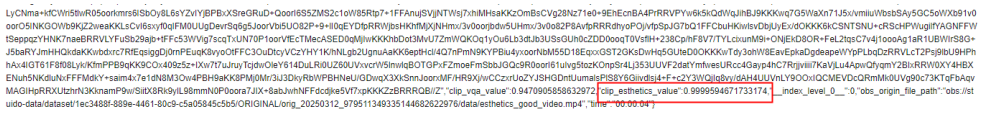
美学评分

- 适用的文件格式：“视频>mp4 / avi”。
- 评分说明：

  从内容（吸引人，清晰度）、构图（目标物位置良好）、颜色（有活力，令人愉悦）、光线（光线明显有对比度）、轨迹（连续、稳定）等维度评价视频美感得分。分值范围(0, 1)，数值越高美感越好，评分>0.95可视为视频美感较高的视频。
- 参数配置样例：



- 评分样例：jsonl文件中显示美学评分：clip_esthetics_value：美学分。

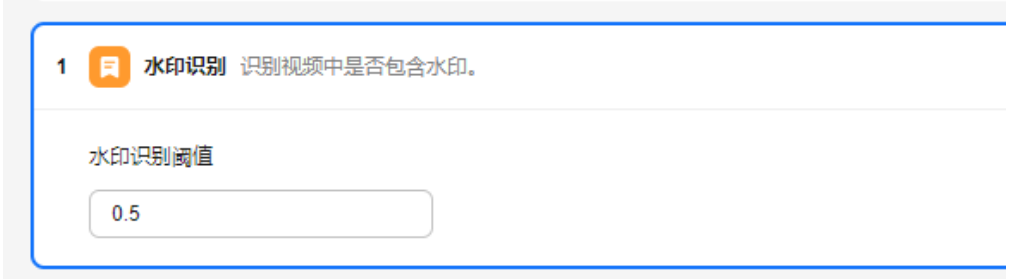


水印识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  识别视频中是否包含水印。
- 参数配置样例：

  水印识别阈值：当水印识别可信度高于水印识别阈值时即判断存在水印，默认水印识别阈值为0.5。
- 参数配置样例：



版权所有 © 华为云计算技术有限公司

- 识别样例：jsonl文件中显示是否识别水印：consist_watermark值为1表示识别到水印，值为0表示未识别到水印。

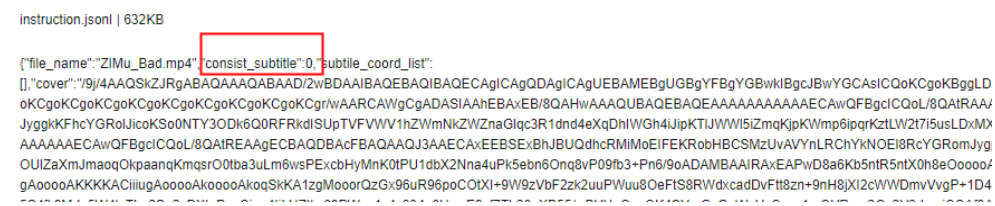


字幕识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  识别视频中是否包含字幕。
- 参数配置样例：



- 识别样例：jsonl文件中显示是否识别字幕：consist_subtitle值为1表示识别到字幕，值为0表示未识别到字幕。



Logo 识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  识别视频中是否包含Logo。
- 参数配置样例：



- 识别样例：jsonl文件中显示是否识别logo：consist_logo值为1表示识别到Logo，值为0表示未识别到Logo。



视频黑边识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  识别视频中是否包含黑边。
- 参数配置样例：



- 识别样例：border_value为1表示识别出黑边，值为0表示未识别出黑边。

版权所有 © 华为云计算技术有限公司

{"file_name":"HeiBianJianCe_bad.mp4","border_value":1,"res_dict":{"border_flag_b":true,"is_exist_border":true,"success_code":"AIS.0200","success_msg":"The video border detect success, video_name is
HeiBianJianCe_bad.mp4","x1":0,"x2":1920,"y1":138,"y2":1023},"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgICAgQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsKDAkKCgr/2wBDAQICAgI
CAgUDAwUKBwYHCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCAQ4B4ADASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6er/8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAAEAQJ3AAECAxEEBSExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75qZY828ZA52ClWADljn2r90Pysg+y/8ATIf980fZf+mQ/wC+as+VH/d/Wjyo/wC7+tAFb7L/0yH/AHzR8Mj++as+VH/AHf1o8qP+7+tAFb7L/0yH/fNH2X2Nna4uPk5ebn6OnqBvP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75qPFFtfypTEnaNT+FSbF9KNi+IAERhjP/LJD/wABo8mMfdIT8FqUgHqKkB6ik3YCLyEwCeC/R5Cf88F/Kpdl+IGxfSoAi8hP+eC/IU/lD/nlv5VbyV939ajyv939aqlvDfsq/88F/75o+yr/zwX/vmrO0f3f0o2f7P6UXJuV/sqf8APJf++aPsqf8APJf++anBst3TR5Y/umsSSt9ni/54L/3z5fZo/wDnin/FPoroNBvkJ/zwTv2KckEGf9fSmf90VMkft9BTvLP90Vjcmmtfh/54v/9timxwQ8uY/++RU0kZ4+UU3yzdFVASsN8mH/min/fAo8mH/min/fAp3In+6KPL90Uci7hZdyHyIdxbyY+f9kUeRD/AM8Y/wDvDkVN5Zujyz/AHRV6FaFeSGHj9yn/tAp7LmnKuaip0FIT7LD/wA8k/74FH2SHnlH/3yKk8s/w80UeWf7orMkgMEJ/5Yp/3wKBBCP+Wk098CpE6f40Rp1/xoAh8mH/min/fAo8mH/min/fAqz9kb0/Wj7I392gCD7LD/wA8k/74FH2SHnlH/3yKs+Q39jyG/uCgCp5EP/ADxj/wC+RvkxO/8APJF++RS+Wf7o

密集文字识别

- 适用的文件格式：“视频>mp4 / avi”。
- 参数说明：

  密集文字面积占比：超出密集文字面积占比阈值的视频可视为密集文字视频，一般密集文字面积占比阈值为1%。

  置信度：当识别置信度超过设定阈值时，即可认定为包含密集文字的视频内容。默认情况下，识别置信度阈值设为 0.5。
- 参数配置样例：



- 识别样例：jsonl文件中显示是否识别密集文字：consist_densetext值为1表示识别到密集文字，值为0表示未识别到密集文字。

{"dense_text_coord_list":
[],"file_name":"MiJiWenZi_good.mp4","densetext_rate_colum":0.0,"consist_densetext":0,"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgICAgQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsK
DAkKCgr/2wBDAQICAgICAgUDAwUKBwYHCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCALQBQADASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6er/8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAAEAQJ3AAECAxEEBSExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75qZY828ZA52ClWADljn2r90Pysg+y/8ATIf980fZf+mQ/wC+as+VH/d/Wjyo/wC7+tAFb7L/0yH/AHzR8Mj++as+VH/AHf1o8qP+7+tAFb7L/0yH/fNH2X2Nna4uPk5ebn6OnqBvP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75qPFFtfypTEnaNT+FSbF9KNi+IAERhjP/LJD/wABo8mMfdIT8FqUgHqKkB6ik3YCLyEwCeC/R5Cf88F/Kpdl+IGxfSoAi8hP+eC/IU/lD/nlv5VbyV939ajyv939aqlvDfsq/88F/75o+yr/zwX/vmrO0f3f0o2f7P6UXJuV/sqf8APJf++aPsqf8APJf++anBst3TR5Y/umsSSt9ni/54L/3z5fZo/wDnin/FPoroNBvkJ/zwTv2KckEGf9fSmf90VMkft9BTvLP90Vjcmmtfh/54v/9timxwQ8uY/++RU0kZ4+UU3yzdFVASsN8mH/min/fAo8mH/min/fAp3In+6KPL90Uci7hZdyHyIdxbyY+f9kUeRD/AM8Y/wDvDkVN5Zujyz/AHRV6FaFeSGHj9yn/tAp7LmnKuaip0FIT7LD/wA8k/74FH2SHnlH/3yKk8s/w80UeWf7orMkgMEJ/5Yp/3wKBBCP+Wk098CpE6f40Rp1/xoAh8mH/min/fAo8mH/min/fAqz9kb0/Wj7I392gCD7LD/wA8k/74FH2SHnlH/3yKs+Q39jyG/uCgCp5EP/ADxj/wC+RvkxO/8APJF++RS+Wf7o

视频分类（InterVideo2）

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  自动对短视频内容进行分类，并生成相应的标签。
- 参数配置样例：

  无需配置参数。
- 分类标注样例：

  描述信息中显示视频的各级分类：

  category_L1_cn:一级分类。

  category_L4_cn：四级分类。

版权所有 © 华为云计算技术有限公司



视频摘要生成（详细）

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  通过对视频进行抽帧，通过模型推理生成详细的视频摘要描述。
- 参数配置样例：
  无需参数配置。
- 打标样例：描述信息中long_prompt字段代表详细的视频摘要。

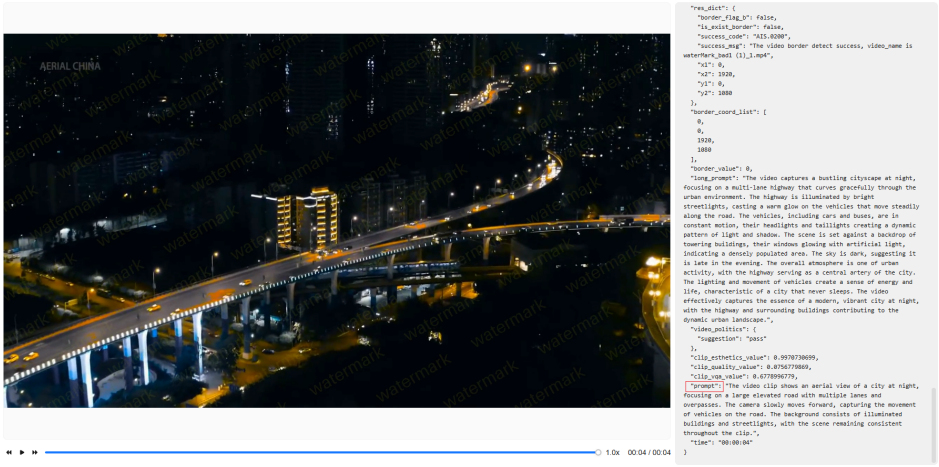


视频摘要生成（简略）

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  通过对视频进行抽帧，通过模型推理生成简略的视频摘要描述。
- 参数配置样例：
  无需参数配置。
- 打标样例：描述信息中prompt字段代表简略的视频摘要。

 版权所有 © 华为云计算技术有限公司

图 4-5 打标样例



姿态检测

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  姿态检测算子从视频中抽取8帧图片，对每帧图片进行关键点打标与置信度计算，并计算符合过滤条件的图片数量，达到一定数量则表示视频中存在对应的人物数量。
- 参数配置样例：

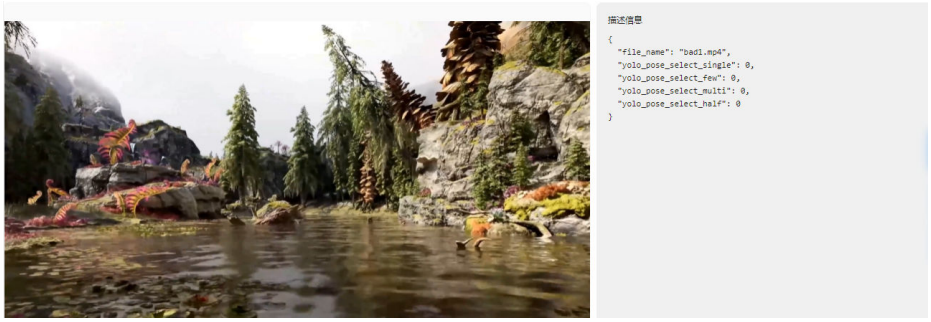
  无需参数配置。
- 打标样例：

  yolo_pose_select_single: 是否检测到了单个人的姿势，存在为1，否则为0。

  yolo_pose_select_few: 是否检测到了少量人（通常为2-4）的姿势，存在为1，否则为0。

  yolo_pose_select_multi: 是否检测到了多人（通常是4人或更多）的姿势，存在为1，否则为0。

  yolo_pose_select_half : 是否检测到了半个人的姿势，存在为1，否则为0。



运动镜头描述

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  模型通过对视频进行抽帧进行光流计算与推理，输出视频的镜头类型。
- 参数配置样例：

 版权所有 © 华为云计算技术有限公司

无需参数配置。

- 打标样例：

  motion: 运镜的类型。

  标签范围为： { 0: 'static', 1: 'others', 2: 'pull out', 3: 'push in', 4: 'static' , 5: 'tracking', 6: 'orbit', 7: 'spin', 8: 'till up', 9: 'till down', 10: 'pan right', 11: 'pan left' ,12: 'tracking' }。



4.5.2.3 图片类加工算子介绍

数据加工算子为用户提供了多种数据操作能力，包括数据提取、过滤、转换、打标签等。这些算子能够帮助用户从海量数据中提取出有用信息，并进行深度加工，以生成高质量的训练数据。

平台提供了图文类、图片类加工算子，算子能力清单见表4-18。

表 4-18 图片类加工算子能力清单

算子分类	算子名称	算子描述
数据提取	图文提取	提取图文压缩包中的JSON文本和图片，并对图片进行结构化解析（BASE64编码），方便图文加工算子使用。
数据过滤	图片元数据过滤	基于图片宽高、文件大小、宽高比阈值进行图片/图文数据清洗。
	图文文本长度过滤	过滤文本长度不在“文本长度范围”内的图文对。一个中文汉字或一个英文字母，文本长度均计数为1。
	图文文本语言过滤	通过语种识别模型得到图文对的文本语种类型，“待保留语种”之外的图文对数据将被过滤。 注：语种识别模型有小概率的误判可能性。
	图文去重	• 基于结构化图片去重图文。 • 判断相同文本对应不同的图片数据是否超过“单文本最大图片数阈值”阈值，如果超过则随机删除多余图片，仅保留阈值以内的图文数量。
	图片去重	通过把图片结构化处理后，过滤重复的图片/图文对数据。

版权所有 © 华为云计算技术有限公司

算子分类	算子名称	算子描述
数据打标	色情图像检测算子	给图像算子打标签。
	危情图像检测算子	给危情图片内容打标签。
	暴恐图像检测算子	过滤暴恐图像。
数据转换	图文异常字符过滤	将文本数据中携带的异常字符替换为空值，数据条目不变。 ● 不可见字符，例如U+0000-U+001F。 ● 表情符□□。 ● 网页标签符号<p> 。 ● 特殊符号，比如● ◆ 。 ● 乱码和无意义的字符□□□□□。

图文提取

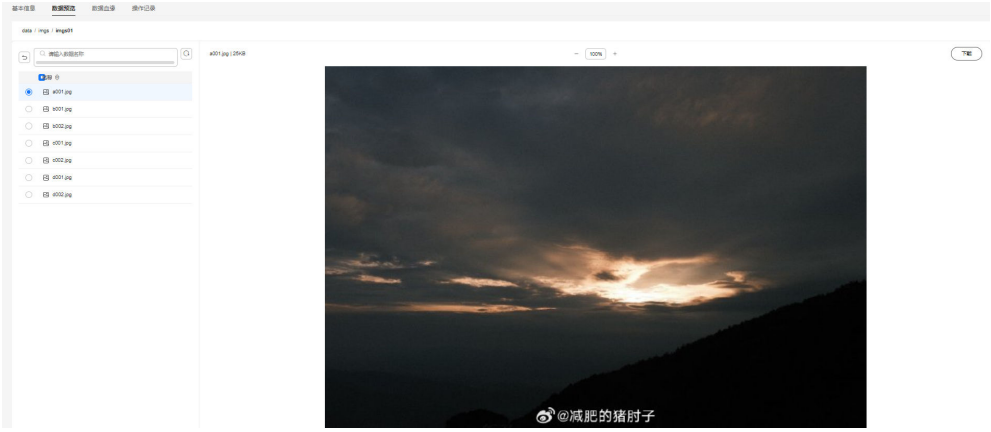
- 适用的文件格式：

  tar+jsonl；所有图片保存为tar包。图片格式支持：jpg、jpeg、png、bmp。图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。
- 各参数说明：

  待提取内容类型：提取图文压缩包中的JSON文本和图片；并对图片进行结构化解析。
- 参数配置样例：

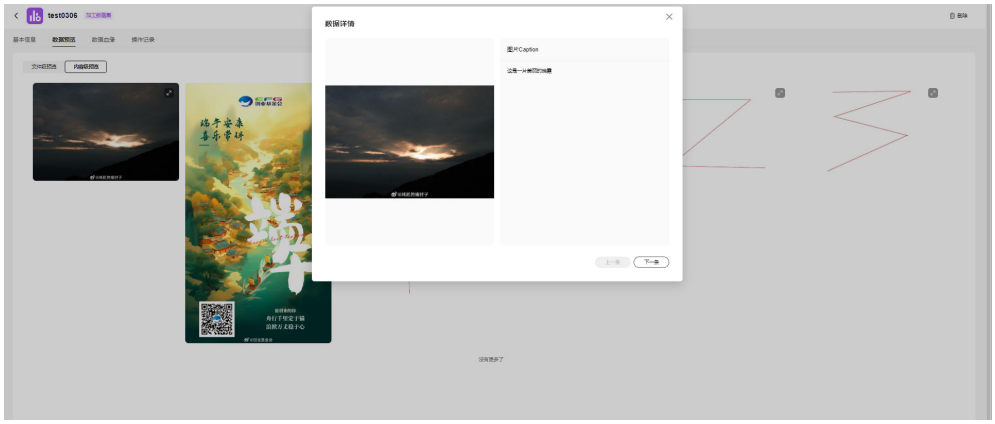
  不需要配置参数。
- 提取样例：

加工前：



 版权所有 © 华为云计算技术有限公司

加工后：



图片元数据过滤

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。

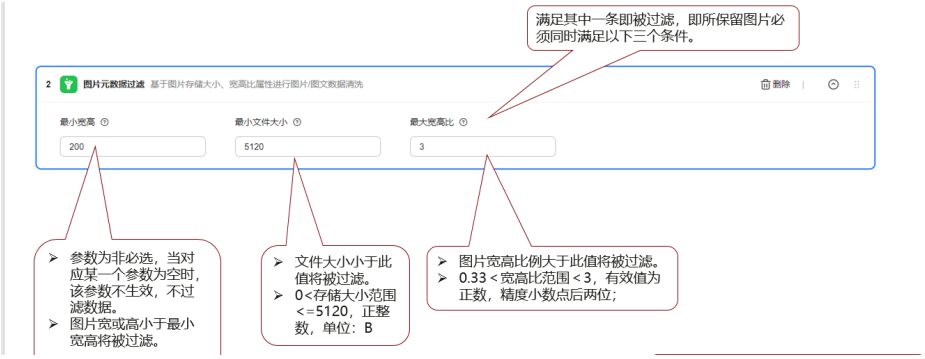
- 各参数说明：

  待过滤内容类型：

  最小宽高：宽或高任意一边低于此设置值，图片会被过滤。

  最小文件大小：文件大小低于该文件大小会被过滤，单位为B。

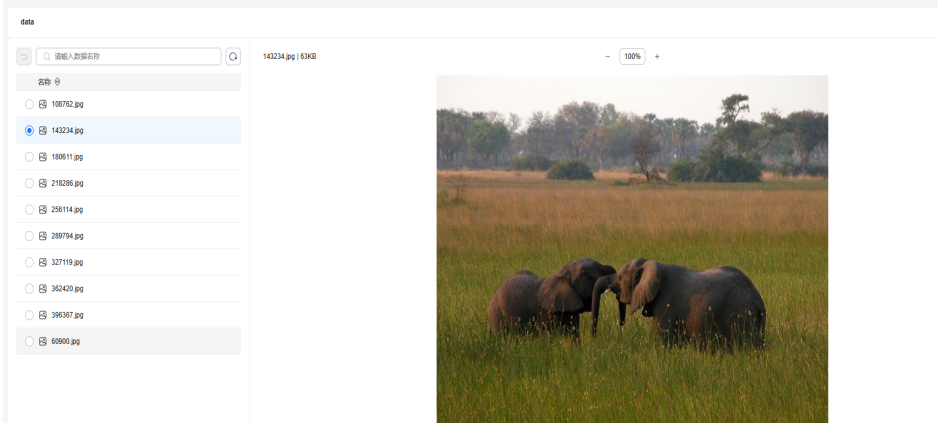
- 参数配置样例：



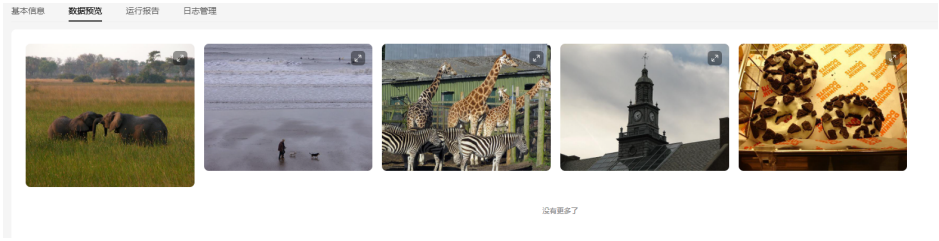
- 过滤样例：

  原数据集：

版权所有 © 华为云计算技术有限公司



过滤后：有一边低于480的图片被过滤。



图文文本长度过滤

- 适用的文件格式：

  tar+jsonl：所有图片保存为tar包。图片支持：jpg、jpeg、png、bmp图片类型。

  图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。

- 各参数说明：

  待过滤内容类型：过滤文本长度不在“文本长度范围”内的图文对。一个中文汉字或一个英文字母，文本长度统计为1。

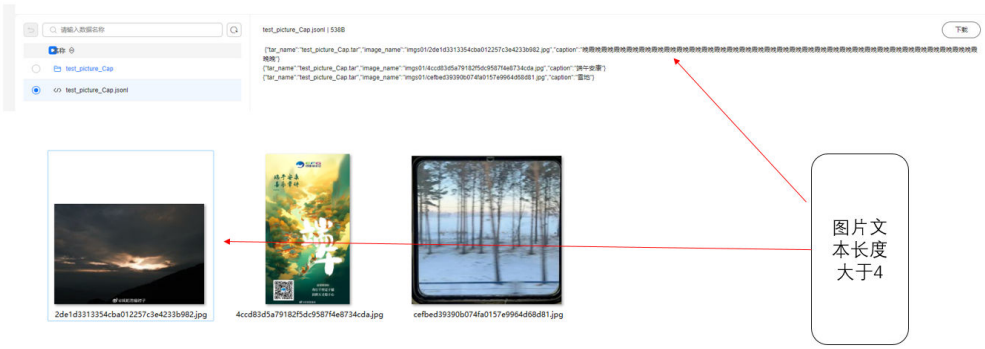
- 参数配置样例：



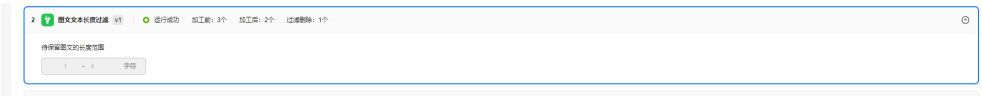
- 过滤样例：

 版权所有 © 华为云计算技术有限公司

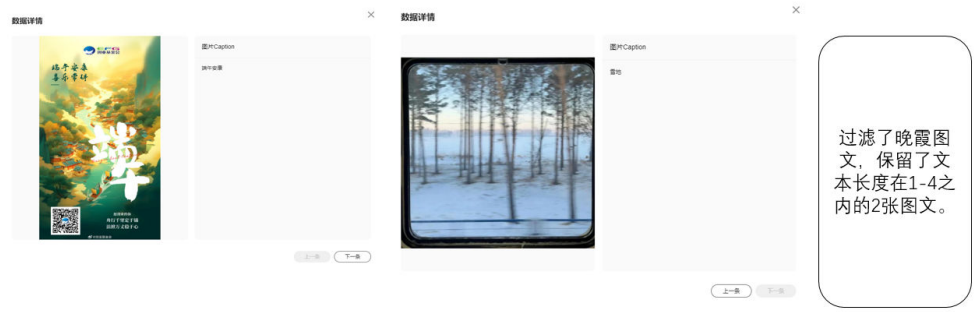
加工前：



参数配置：



加工后：



图文文本语言过滤

- 适用的文件格式：

  tar+jsonl：所有图片保存为tar包。图片支持：jpg、jpeg、png、bmp图片类型。

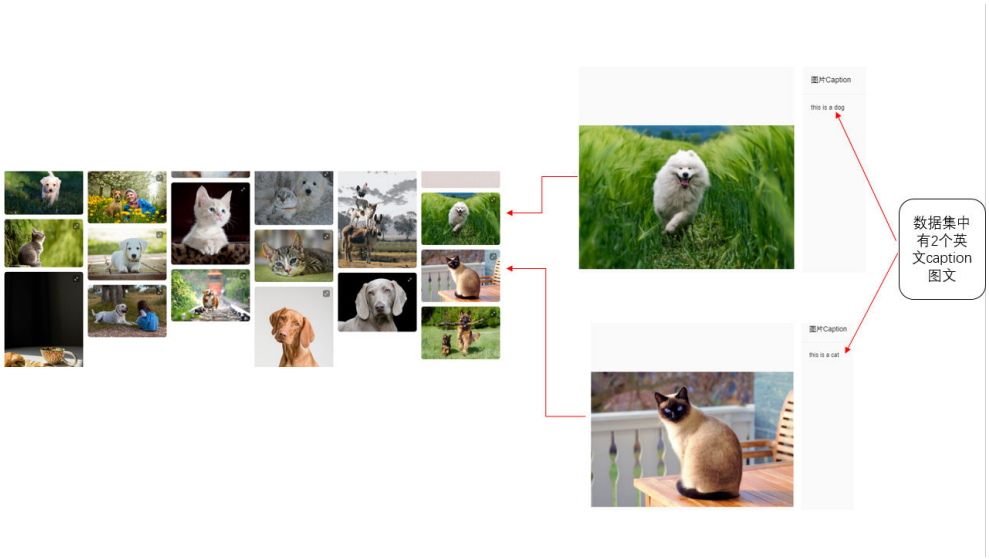
  图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。
- 各参数说明：

  待过滤内容类型：提取图文压缩包中的JSON文本和图片，并对图片进行结构化解析（BASE64编码），方便图文加工算子使用。
- 参数配置样例：

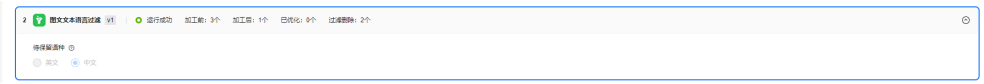


版权所有 © 华为云计算技术有限公司

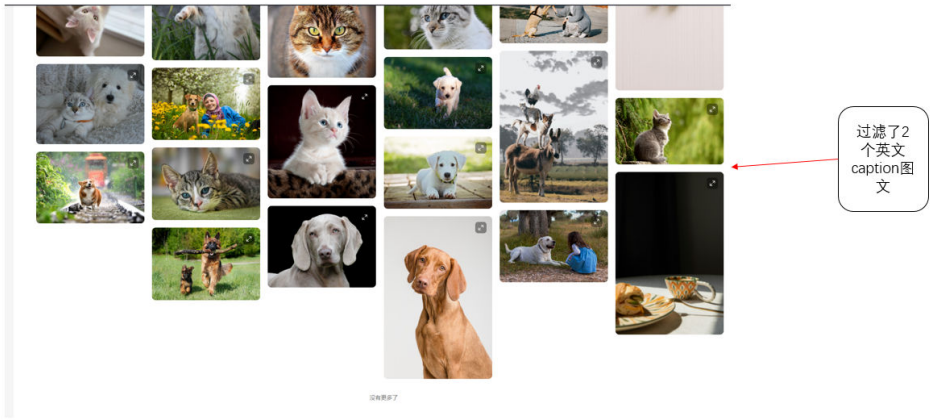
- 去重样例：
  加工前：



  参数配置：



  加工后：

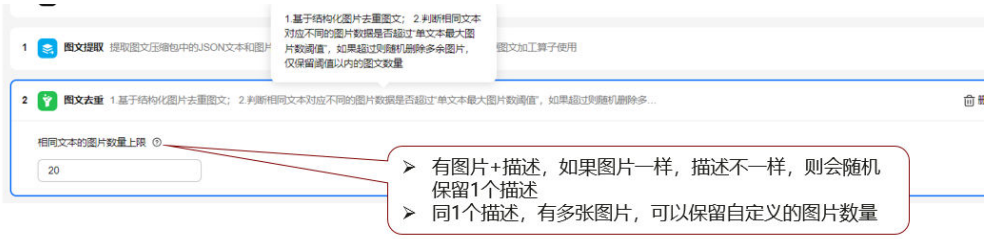


图文去重

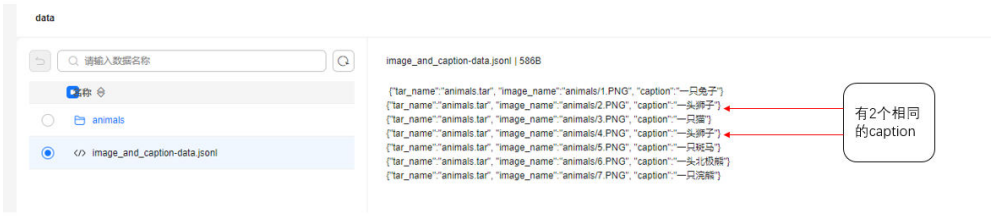
- 适用的文件格式：
  tar+jsonl：所有图片保存为tar包。图片支持：jpg、jpeg、png、bmp图片类型。图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。
- 各参数说明：
  待过滤内容类型：
  a. 基于结均化图片去重图文。
  b. 判断相同文本对应不同的图片数据是否超过单文本最大图片数阈值，如果超过则随机删除多余图片， 仅保留阈值以内的图文数量。

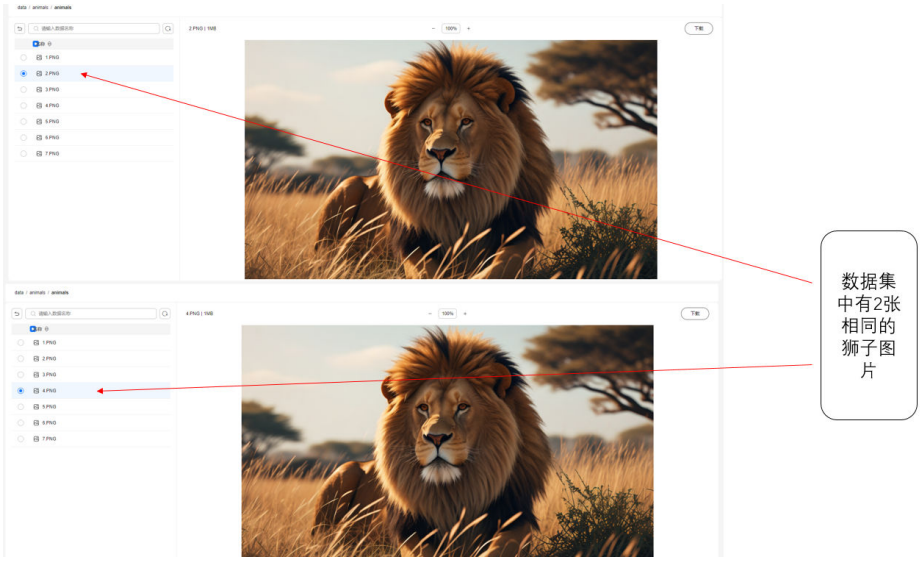
 版权所有 © 华为云计算技术有限公司

- 参数配置样例：



- 去重样例：
  加工前：



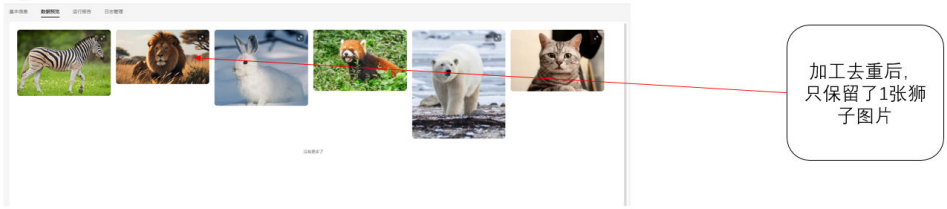


参数配置：



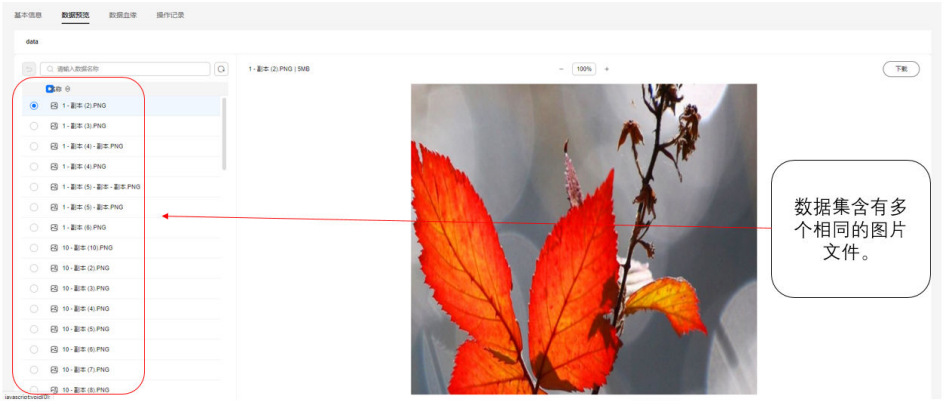
版权所有 © 华为云计算技术有限公司

加工后：

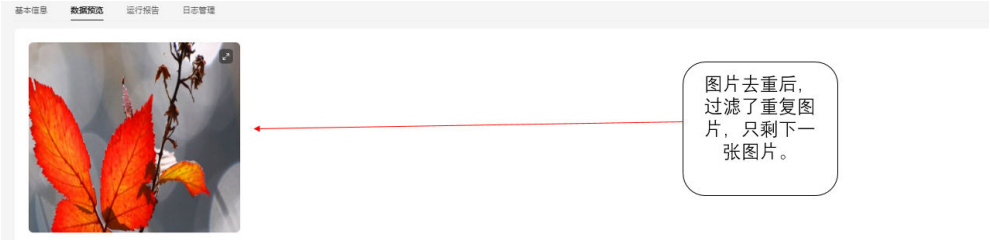


图片去重

- 适用的文件格式：
  jpg、jpeg、png、bmp。
  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。
- 各参数说明：
  待过滤内容类型：通过把图片结构化处理后，过滤重复的图片/图文对数据。
- 参数配置样例：
  不需要配置参数。
- 过滤样例：
  加工前：



加工后：



版权所有 © 华为云计算技术有限公司

色情图像检测算子

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。

- 各参数说明：

  待打标内容类型：对图片的涉黄程度进行评分， 分数越高越危险。 评分范围[(0.100)， 默认评分≥50分的视频可视为涉黄视频。

- 参数配置样例：

  不需要配置参数。

- 检测样例：

  检测结果以image_porn对象存储在标注文件中：

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体色情标签，如果未检测出则为空。

描述信息


```
{
  "tar_name": "data.tar",
  "file_name": "1259.jpg",
  "image_format": "jpg",
  "height": 336,
  "width": 336,
  "size": 24438,
  "base_id": "2d6dcc7c-2ee0-4928-95a8-97ece90c1134",
  "wh_radio": 1,
  "image_porn": {
    "results": [
      {
        "category": "porn",
        "confidence": 0.9956,
        "label": "sexy"
      }
    ],
    "suggestion": "block"
  },
  "filter": 0
}
```


危情图像检测算子

- 适用的文件格式：

版权所有 © 华为云计算技术有限公司

jpg、jpeg、png、bmp。

tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型；

- 各参数说明：

  待打标内容类型：给危情图片内容打标签。

- 参数配置样例：

  不需要配置参数。

- 检测样例：检测结果以image_danger对象存储在标注文件中:

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体危情标签，如果未检测出则为空。

描述信息


```
{
  "tar_name": "data.tar",
  "file_name": "图片1.png",
  "image_format": "png",
  "height": 720,
  "width": 466,
  "size": 893658,
  "base_id": "14ef093e-1ff8-4757-bc15-0963b51ca56a",
  "wh_radio": 0.6472222222,
  "image_danger": {
    "results": [
      {
        "category": "terrorism",
        "confidence": 0.9941,
        "label": "riot"
      }
    ],
    "suggestion": "block"
  },
  "filter": 0
}
```


暴恐图像检测算子

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型；

版权所有 © 华为云计算技术有限公司

- 各参数说明：

  待打标内容类型：过滤暴恐图像。

- 参数配置样例：

  不需要配置参数。

- 检测样例：检测结果以image_terrorism对象存储在标注文件中：

  suggestion:对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence:模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label:模型检测出的具体暴恐标签，如果未检测出则为空。

  描述信息

```
{
  "tar_name": "data.tar",
  "file_name": "BA202310100126_6_0_Im0.png",
  "image_format": "png",
  "height": 279,
  "width": 652,
  "size": 59206,
  "base_id": "47c0346e-6dd8-47b6-89e6-1dabab4384e6",
  "wh radio": 2.3369175627,
  "image_terrorism: {
       "results": [
          {
             "category": "terrorism",
             "confidence": 0.9941,
             "label": "riot"
    }
       "suggestion": "block"
  },
  "filter": 0
}
```

图文异常字符过滤

- 适用的文件格式：

  tar+jsonl：所有图片保存为tar包。图片支持：jpg、jpeg、png、bmp图片类型。

  图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。

- 各参数说明：

  待转换内容类型：将文本数据中携带的异常字符替换为空值，数据条目不变。

  1）不可见字符，比如U+0000-U+001F 。

版权所有 © 华为云计算技术有限公司

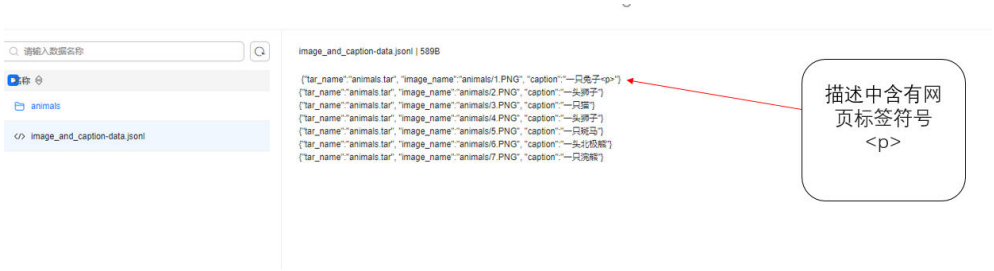
2）表情符?? 。

3）网页标签符号<p> 。

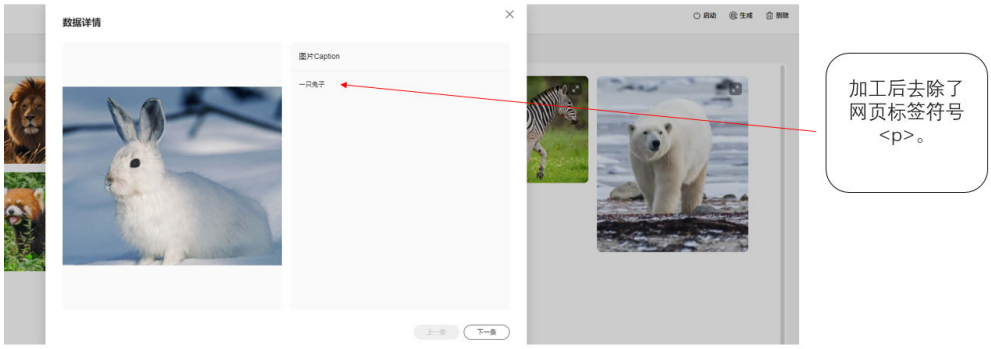
4）特殊符号，比如●▉◆ 。

5）乱码和无意义的字符[illegible]。

- 参数配置样例：
  不需要配置参数。
- 过滤样例：
  加工前：



加工后：



4.5.2.4 气象类加工算子介绍

平台支持气象类数据集的加工操作，气象类加工算子能力清单见表4-19。

表 4-19 气象类加工算子能力清单

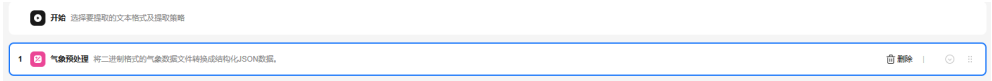
算子分类	算子名称	算子描述
科学计算	气象预处理	将二进制格式的气象数据文件转换成结构化JSON数据。

气象预处理

- 适用的文件格式：.nc、.cdf、.netcdf、.gr、.gr1、.grb、.grib、.grb1、.grib1、.gr2、.grb2、.grib2。
- 算子说明：将二进制格式的气象数据文件转换成结构化JSON数据。

版权所有 © 华为云计算技术有限公司

- 参数配置样例：



- 过滤样例：

  加工后：

{"total_size": 145364716, "geo_range": {"lon": [0.0, 360.0], "lat": [-90.0, 90.0]}, "time_range": [1641146400000, 1641146400000], "forcing_types": ["TERRAIN_HEIGHT", "SOIL_TYPE", "LAND_MASK", "OCEAN_MASK"], "surface_features": [], "surface_constants": ["TERRAIN_HEIGHT", "SOIL_TYPE", "LAND_MASK", "OCEAN_MASK"], "upper_air_layers": ["1000hPa", "100hPa", "150hPa", "175hPa", "200hPa", "250hPa", "300hPa", "400hPa", "500hPa", "50hPa", "600hPa", "700hPa", "850hPa", "925hPa"], "upper_air_features": ["Q", "T", "U", "V", "Z"], "under_sea_features": [], "under_sea_depths": []}

4.5.2.5 音频类加工算子介绍

平台支持音频类数据集的加工操作，音频类加工算子能力清单见**表4-20**。

表 4-20 音频类加工算子能力清单

算子分类	算子名称	算子描述
数据转换	**噪声添加**	给音频添加噪声。
	噪声抑制	移除音频中的纯噪声片段并进行降噪。
	音调变化	调整原始音频的音调。
	混响减少	减少声音在空间中的混响效果，提高语音可懂度。
	语音匿名化	对音频进行匿名化处理，匿名化后的音频与原始说话人音色和声纹有明显区分度。
	语音降噪	对原始音频进行降噪，且只考虑噪声与人声叠加的情况，对于纯噪声音频或者纯噪声片段，不做约束。
	语速调整	调整音频中的说话速度。
	音频量化编码	将一个高分辨率的含有头信息的音频文件，通过音频编解码技术、量化压缩技术转换成16k采样率alaw/ulaw/pcm/wav文件。
数据打标	**语音语种识别标注**	识别音频中说话人所使用的语种，并给出相应的置信度。
	普通话语音转文本	将普通话语音快速转化为文本，以丰富人机交互场景。
	语音情绪识别标注	识别输入音频中说话人的情绪。
	语音端点检测	检测音频中多段人声各自的起止时间。
	噪声水平评估	对含有人声片段的音频进行质量打分。

版权所有 © 华为云计算技术有限公司

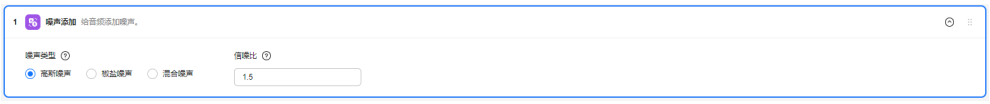
算子分类	算子名称	算子描述
	静默片段检测	识别音频中的静音片段及置信度，并给出静音片段的占比。
	多说话人语音识别	识别音频内容，返回每个说话人说话的起止时间及对应内容。
	个人隐私对话识别	给个人隐私语音内容打标签。
	违禁语音检测	给违禁语音内容打标签。
	政治敏感语音识别	给政治敏感语音内容打标签。
	色情语音内容检测	给色情语音内容打标签。

噪声添加

- 适用的文件格式：纯音频，60s内采样率16k的wav音频文件。
- 算子说明：给音频添加噪声。
- 各参数说明：

  噪声类型：添加的噪声类型，混合噪声为高斯噪声与椒盐噪声的叠加。

  信噪比: 正常声音信号强度与噪声信号强度的比值。
- 参数配置样例：



噪声抑制

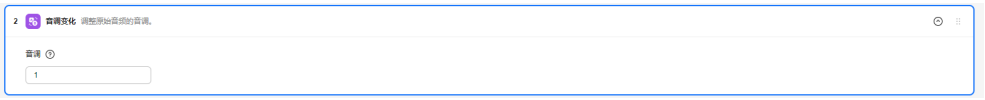
- 适用的文件格式：纯音频，30s内采样率16k、位深16bit单通道的wav音频文件。
- 算子说明：移除音频中的纯噪声片段并进行降噪。
- 参数配置样例：

  不需要配置参数。

音调变化

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：调整原始音频的音调。
- 各参数说明：

  音调：音调参数。
- 参数配置样例：



版权所有 © 华为云计算技术有限公司

混响减少

- 适用的文件格式：纯音频，60s内采样率16k的wav音频文件。
- 算子说明：减少声音在空间中的混响效果，提高语音可懂度。
- 参数配置样例：

  不需要配置参数。

语音匿名化

- 适用的文件格式：纯音频，30s内采样率16k、位深16bit单通道的wav音频文件。
- 算子说明：对音频进行匿名化处理，匿名化后的音频与原始说话人音色和声纹有明显区分度。
- 参数配置样例：

  不需要配置参数。

语音降噪

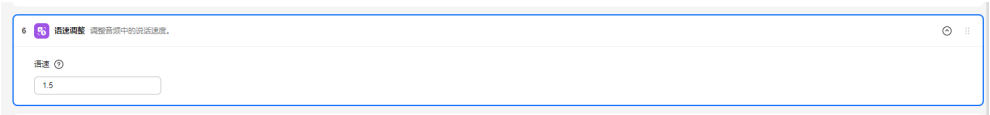
- 适用的文件格式：纯音频，采样率16k、位深16bit单通道的wav音频文件。
- 算子说明：对原始音频进行降噪，且只考虑噪声与人声叠加的情况，对于纯噪声音频或者纯噪声片段，不做约束。
- 参数配置样例：

  不需要配置参数。

语速调整

- 适用的文件格式：纯音频，60s内的wav音频文件。
- 算子说明：调整音频中的说话速度。
- 各参数说明：

  语速：语速参数， 取值在 0.5~2 之间。
- 参数配置样例：



音频量化编码

- 适用的文件格式：纯音频，100MB内的音频文件。
- 算子说明：将一个高分辨率的含有头信息的音频文件，通过音频编解码技术、量化压缩技术转换成16k采样率alaw/ulaw/pcm/wav文件。
- 参数配置样例：

  不需要配置参数。

语音语种识别标注

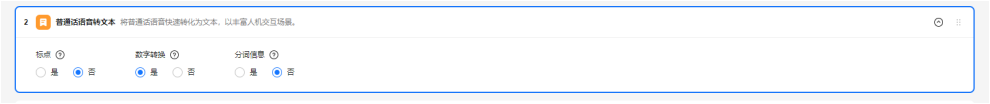
- 适用的文件格式：纯音频，60s内采样率16k、位深16bit的wav音频文件。
- 算子说明：识别音频中说话人所使用的语种，并给出相应的置信度。
- 参数配置样例：

版权所有 © 华为云计算技术有限公司

不需要配置参数。

普通话语音转文本

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：将普通话语音快速转化为文本，以丰富人机交互场景。
- 各参数说明：
  标点 ： 识别结果是否添加标点符号。
  数字转换：是否将语音中的数字识别为阿拉伯数字。
  是否分词信息 ： 识别结果是否包含分词结果信息。
- 参数配置样例：



语音情绪识别标注

- 适用的文件格式：纯音频，60s内采样率16k、位深16bit的wav音频文件。
- 算子说明：识别输入音频中说话人的情绪。
- 参数配置样例：
  不需要配置参数。

语音端点检测

- 适用的文件格式：纯音频，600s内采样率16k、位深16bit的wav音频文件。
- 算子说明：检测音频中多段人声各自的起止时间。
- 参数配置样例：
  不需要配置参数。

噪声水平评估

- 适用的文件格式：纯音频，60s内采样率16k、位深16bit的wav音频文件。
- 算子说明：对含有人声片段的音频进行质量打分。
- 参数配置样例：
  不需要配置参数。

静默片段检测

- 适用的文件格式：纯音频，600s内采样率16k、位深16bit的音频文件。
- 算子说明：识别音频中的静音片段及置信度，并给出静音片段的占比。
- 参数配置样例：
  不需要配置参数。

多说话人语音识别

- 适用的文件格式：纯音频，1小时内、单通道的音频文件。
- 算子说明：识别音频内容，返回每个说话人说话的起止时间及对应内容。

 版权所有 © 华为云计算技术有限公司

- 各参数说明：
  标点：识别结果是否添加标点符号。
  数字转换 ： 是否将语音中的数字识别为阿拉伯数字。
  分词信息 ： 是否将语音中的数字识别为阿拉伯数字。
  话者分离：识别结果是否包含说话人信息。
  语速 ： 识别结果是否包含说话人语速信息。
- 参数配置样例：



个人隐私对话识别

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：给个人隐私语音内容打标签。
- 参数配置样例：
  不需要配置参数。

违禁语音检测算子

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：给违禁语音内容打标签。
- 参数配置样例：
  不需要配置参数。

政治敏感语音识别

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：给政治敏感语音内容打标签。
- 参数配置样例：
  不需要配置参数。

色情语音内容检测算子

- 适用的文件格式：纯音频，60s内的音频文件。
- 算子说明：给色情语音内容打标签。
- 参数配置样例：
  不需要配置参数。

4.5.2.6 通用类加工算子介绍

平台支持通用类数据集的加工操作，算子能力清单见**表1 通用类加工算子能力清单**

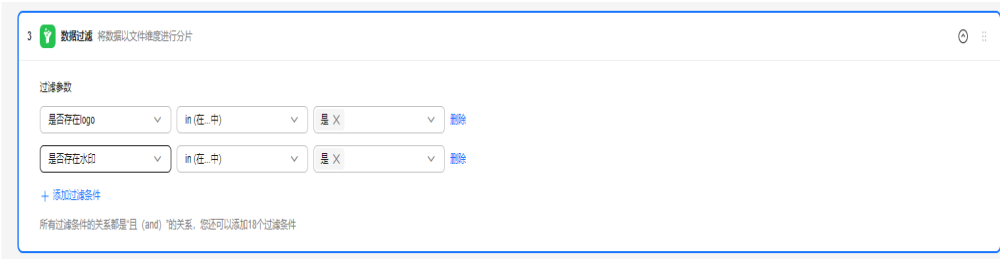
版权所有 © 华为云计算技术有限公司

表 4-21 通用类加工算子能力清单

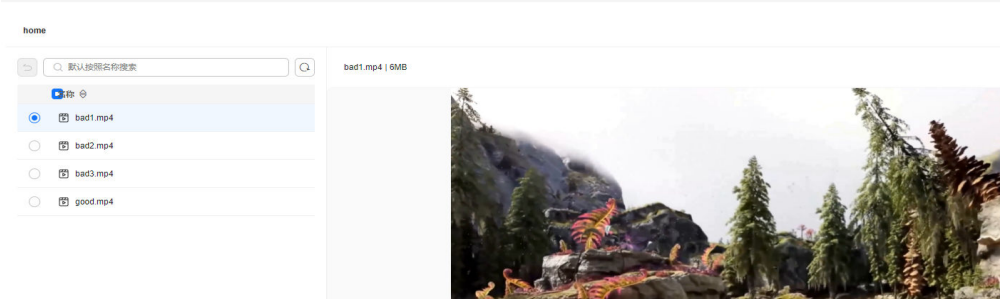
算子分类	算子名称	算子描述
数据过滤	通用过滤算子	根据已打标的标签对数据内容进行过滤。

通用过滤算子

- 适用的文件格式：包含符合条件的标签的文本、音频、图片、视频数据集。
- 算子说明：根据已打标的标签对数据内容进行过滤。
- 参数配置样例：



- 过滤样例：
  原视频数据集：



设置过滤存在logo与水印的视频，经算子处理过后。



4.5.3 加工文本类数据集

版权所有 © 华为云计算技术有限公司

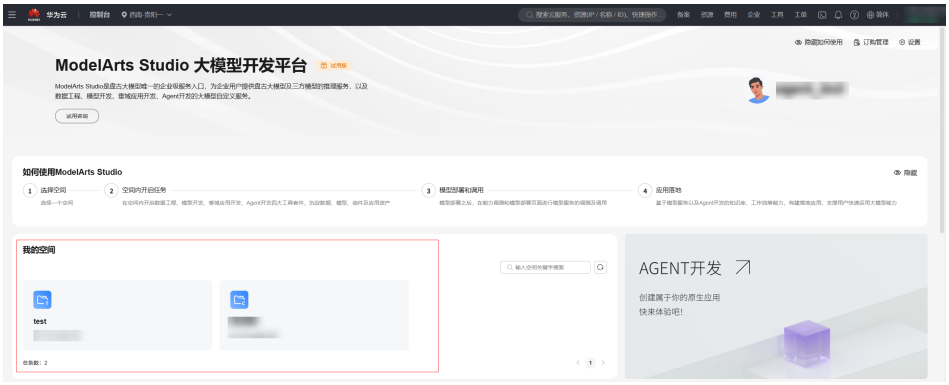
4.5.3.1 加工文本类数据集

加工文本类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建文本类数据集加工任务步骤如下：

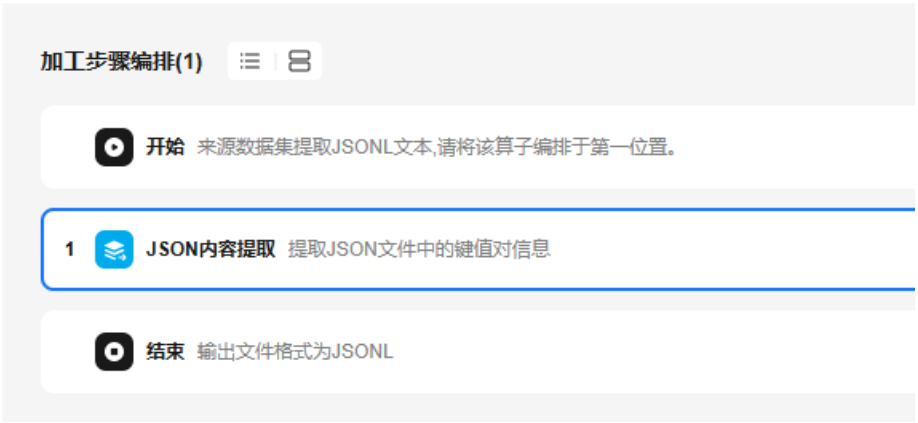
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-6 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的文本类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于文本类数据集，可选择的加工算子请参见**文本类加工算子介绍**。
   a. 在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

   图 4-7 算子编排

注意

如果算子参数涉及大模型选择，需要在ModelArtsStudio中购买模型并部署。

版权所有 © 华为云计算技术有限公司

c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

若选择使用加工模板，将删除当前已编排的加工步骤。

图 4-8 选择加工模板



5. 加工步骤编排完成后，单击“下一步”进入“任务配置界面”。

- 自动生成加工数据集

  勾选，配置生成加工数据集的信息，如图4-9；单击右下角“确定”，平台将启动加工任务。加工任务运行成功后自动生成加工数据集。

  不勾选，单击右下角“确定”，平台将启动加工任务。加工任务运行成功后需要手动生成加工数据集。

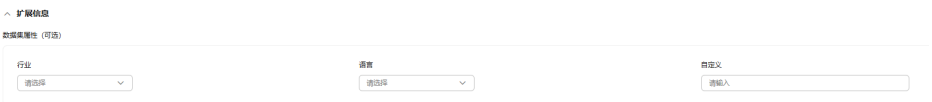
图 4-9 自动生成加工数据集



- 扩展信息（可选）

  可选择行业、语言信息，或自定义数据集属性。

图 4-10 扩展信息



6. 单击“启动加工”。当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

版权所有 © 华为云计算技术有限公司

说明

- 加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.3.2 合成文本类数据集

数据合成是利用预置或自定义的数据指令对原始数据集进行处理，并根据设定的轮数生成新的数据。

当前，数据合成功能支持合成单轮问答、单轮问答（人设）、问答排序、偏好优化DPO、偏好优化DPO（人设）类型的数据。

说明

合成预置指令调测，自定义指令创建，以及合成任务运行都需要依赖大模型，需要在ModelArtsStudio中购买模型并部署。

使用合成任务前，请确保有已部署的NLP大模型，具体步骤详见**创建NLP大模型部署任务**。

创建文本类数据集合成任务

合成文本类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建文本类数据集合成任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

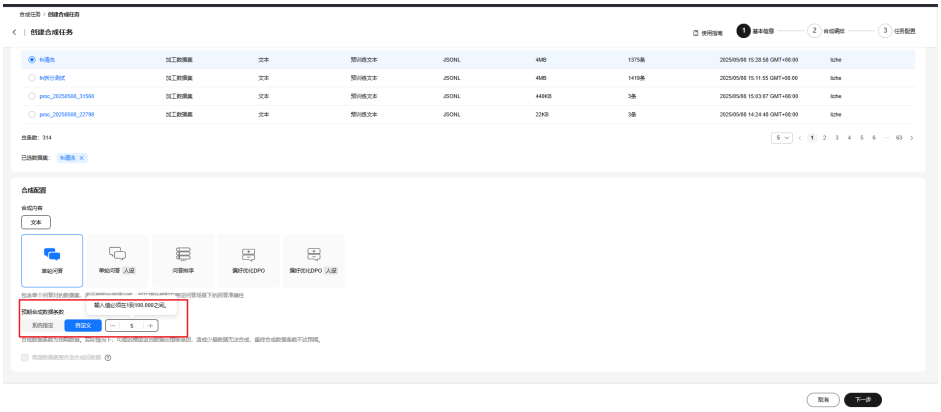
图 4-11 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 合成任务”，单击界面右上角“创建合成任务”。
3. 在“创建合成任务”页面，选择需要合成的数据集，选择合成内容与预期合成条数，如**图4-12**，系统指定预期合成条数则为数据集初始数据条数。

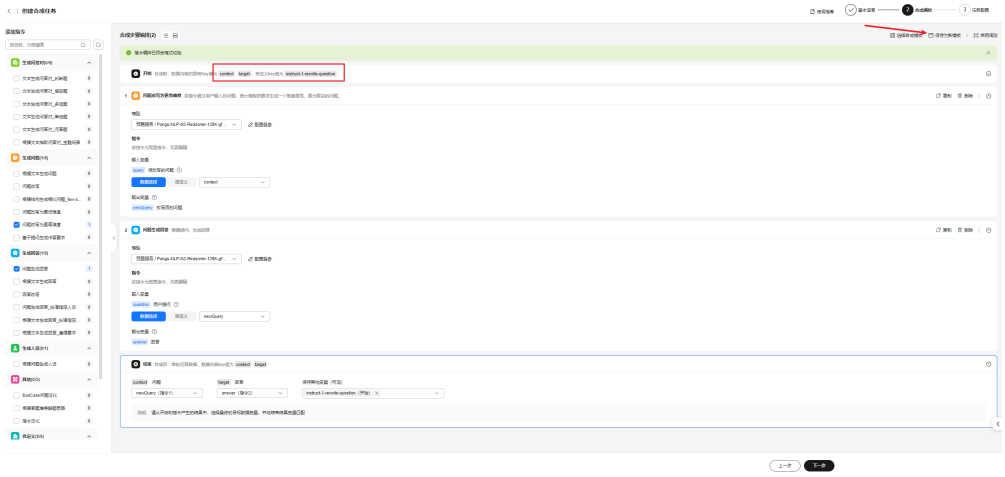
版权所有 © 华为云计算技术有限公司

图 4-12 创建合成任务预期合成数据条数配置



4. 如果合成前的数据集与合成后的数据集结构相同，可选择开启“将源数据集整合至合成后数据”，在合成任务运行完成后，将生成的数据与原始数据集合并，单击“下一步”。
5. 进入“合成编排”页面，如图4-13所示页面开始栏会显示当前数据集数据类型对应的固定字段和除固定字段以外，数据集内其它的自定义字段，这些字段可以在指令输入输出变量中选中。在左侧“添加指令”页面可选择预置指令或自定义指令，指令编排需要按逻辑编排，若同时选择三个生成问答对类型的指令，最终输出结果也只会保存一条问答对数据。

图 4-13 合成编排示例



6. 编排完成的指令，可单击"创建合成模板"，保存当前的指令编排内容供下次一键复用，在右侧可通过单击“选择合成模板”选择一个模板编排，如图4-14所示，模板只有输入类型和输出类型相同的情况下才可见，如：输入为预训练文本，输出为单轮问答，若当前输入数据类型（如：单轮问答）与保存模板时的输入类型（如：预训练）不一致，当前数据选择合成模板时将不可见该模板。

版权所有 © 华为云计算技术有限公司

图 4-14 选择合成模板

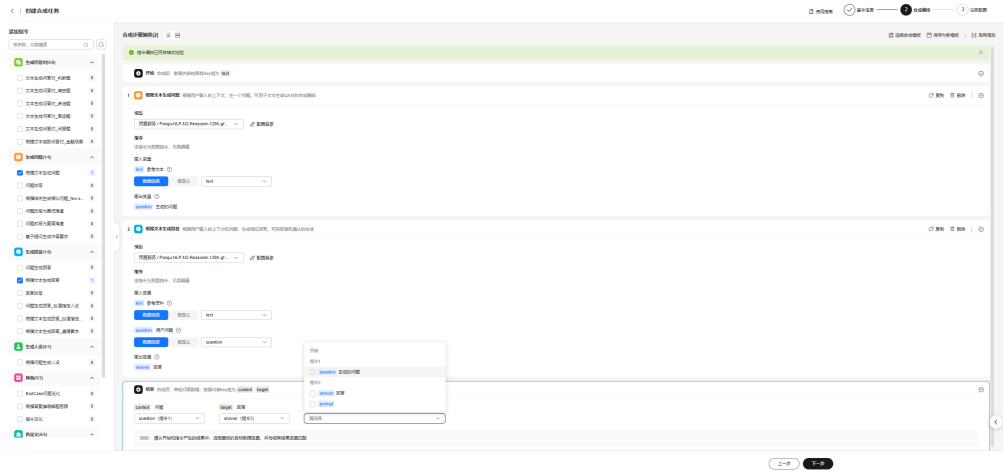


- 预置指令。平台为用户提供了多种预置指令，便于用户执行合成任务，请详见**预置数据指令介绍**。
- 自定义指令。平台支持编排用户自定义指令。自定义指令的创建详见**创建自定义数据合成指令**。

7. 指令选择完成后，单击“确定”，并配置指令参数。

   如**图4-15**，展示了预训练文本类数据集的合成指令参数配置示例，该合成任务实现利用预训练文本生成问答对，除了输出问答对的固定字段context和target外，还支持将指令中间结果保存到最终输出结果。编排完的指令可以单击右侧“保存合成模板”，后续的预训练文本可一键选中模板生成问答对。

图 4-15 预训练文本类数据集合成指令参数配置示例



8. 指令编排完成后，单击右上角“启用调测”，可以对当前编排的指令效果进行预览。
9. 指令调测完成后，单击右下角“下一步”，选择是否勾选自动生成加工数据集。

   勾选，配置生成加工数据集的信息，如**图6 自动生成加工数据集**；单击右下角“创建并启动”，平台将启动合成任务。合成任务运行成功后自动生成加工数据集。

   不勾选，单击右下角“创建并启动”，平台将启动合成任务。合成任务运行成功后手动生成加工数据集。

版权所有 © 华为云计算技术有限公司

图 4-16 自动生成加工数据集



10. 当数据合成任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成合成操作。

说明

在完成数据合成后，若无需使用数据标注、数据配比功能，可直接在“合成任务”页面单击操作列“生成”，生成加工数据集。

生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

预置数据指令介绍

ModelArts Studio平台的数据合成功能为用户提供了预置指令，用户可以在“数据工程 > 数据加工 > 合成任务 > 管理合成指令 > 系统预置”查看指令详情，如图4-17，单击“调测”可查看调测指南，如图4-18，帮助用户更好地使用该指令。

预置的数据指令清单详见表4-22。

图 4-17 指令详情



 版权所有 © 华为云计算技术有限公司

图 4-18 调测指南



表 4-22 预置数据指令清单

指令分类	指令名称	指令描述
生成问题	问题改写为更低难度	该指令可以通过用户输入的问题，使大模型按要求生成一个难度更低、更为简单的问题。
	问题改写为更高难度	该指令通过用户输入的问题，使大模型按要求生成一个难度更高、更为复杂的问题。
	基于提问生成作答要求	该指令根据输入的问题，使大模型泛化一个相应问题的作答要求，该要求与原问题内容不直接相关。该指令可与根据作答要求回答问题的指令进行编排，实现风格多样回答的合成。
	根据样例生成相似问题_few-shot	该指令通过用户输入的多个问题样例，生成一个或多个与样例风格相匹配的新问题。
	根据文本生成问题	根据用户输入的上下文，生一个问题。可用于文本生成QA对的合成编排。
	问题改写	改写问题，生成更复杂的问题，可用于指令泛化。
生成回答	回答改写	根据用户指定人设，改写回答的风格，不改变回答内容。可与人设泛化指令编排，实现问答对泛化。

 版权所有 © 华为云计算技术有限公司

指令分类	指令名称	指令描述
	根据文本生成回答_遵循要求	根据用户指定的指令要求和问题，根据输入的上下文生成相应回答。可与指令泛化进行编排，实现事实遵循类问答对泛化。
	问题生成回答	根据提问，生成回答。
	根据文本生成回答_扮演指定人设	根据用户指定的人设和问题，根据输入的上下文生成相应回答。可与人设泛化指令编排，实现事实遵循类问答对泛化。
	问题生成回答_扮演指定人设	根据用户指定的人设和问题，生成相应回答。可与人设泛化指令编排，实现问答对泛化。
	根据文本生成回答	根据用户输入的上下文和问题，生成相应回答。可实现事实遵从的合成
生成问答对	文本生成问答对_判断题	该指令能够从用户提供的参考文本中构建出一个判断题，同时给出其正确回答。
	文本生成问答对_填空题	该指令能够从用户提供的参考文本中构建出一个填空题，同时给出其正确回答。
	文本生成问答对_问答题	该指令能够从用户提供的参考文本中构建出一个问题，同时给出其相应回答。
	根据文本抽取问答对_金融场景	根据用户输入的金融类文档进行问答对的抽取。
生成人设	根据问题生成人设	根据用户输入的问题生成一个人物设定。
其他	BadCase问题泛化	该指令通过用户提供的badcase问题和回答，利用大模型生成在类似情景下可能犯错的攻击性问题。用户可指定生成的攻击性问题个数，个数不超过10。
	根据答案推导解题思路	指令通过用户输入的问题和回答，利用大模型生成包含相应解题思路的回答。
	指令泛化	根据用户指定风格，进行指令泛化。可与指定要求类的问答对生成相关指令编排，实现问答对泛化。

创建自定义数据合成指令

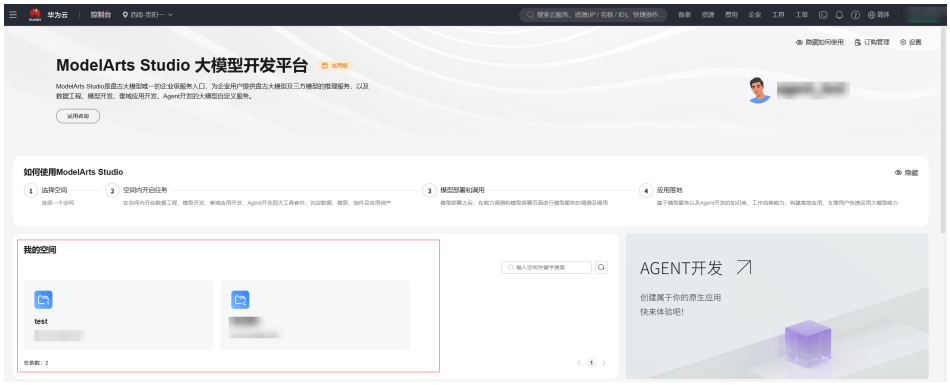
平台支持用户创建自定义数据合成指令。

本章节将以“生成主题散文”的场景为例，详细介绍自定义数据合成指令的配置步骤。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 4-19 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 合成任务 > 管理合成指令”，在“自定义”页签，单击“创建自定义指令”。
3. 在“创建指令”弹窗中，输入名称与描述，单击“确定”，进入配置合成指令页面。
4. 选择变量标识符为“双大括号{{}}”，输入指令为“请以{{topic}}为主题，写一篇字数不超过{{num}}的散文。”

   单击“识别”，再单击“确定”。

图 4-20 配置指令



5. 按照表4-23进行变量配置。

表 4-23 数据指令变量配置

变量类型	变量名称	变量类型	变量描述
输入变量	topic	string	主题
	num	string	字数
输出变量	output	string	散文

其中，输出变量的“变量描述”字段为大模型理解的内容，需仔细填写。

版权所有 © 华为云计算技术有限公司

图 4-21 配置变量



6. 调测数据指令。
   - 在“调试 > 模型”中，选择指令所需的模型，单击“配置超参”可自定义设定超参数值：
     - 温度：温度采样，较高的值，如0.8会使输出更随机，而较低的值，如0.2会使其更加集中和确定性。取值范围0-1。
     - 多样性：即top_p，核心采样，模型会考虑具有top_p概率质量的标记结果。0.1 表示只有占前 10% 概率质量的标记被考虑。建议该参数和temperature只设置一个。取值范围0-1。
     - 重复惩罚：重复采样惩罚值，取值越大表示惩罚越大，因此可降低模型重复相同行为的可能性。取值范围-2.0-2.0。
     - 采样：即top_k，采样参数，每轮token生成时，保留k个概率最高的token作为候选。取值越大，生成文本的多样性越强。
   - 在“调试 > 输入”中，可通过给变量赋值来查看效果。

图 4-22 指令调测



版权所有 © 华为云计算技术有限公司

7. 调试完成后，单击“立即创建”，创建该数据指令。
   成功创建的数据指令将在“管理合成指令 > 自定义”页面中展示。

4.5.3.3 标注文本类数据集

创建文本类数据集标注任务

标注文本类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表4-24**。

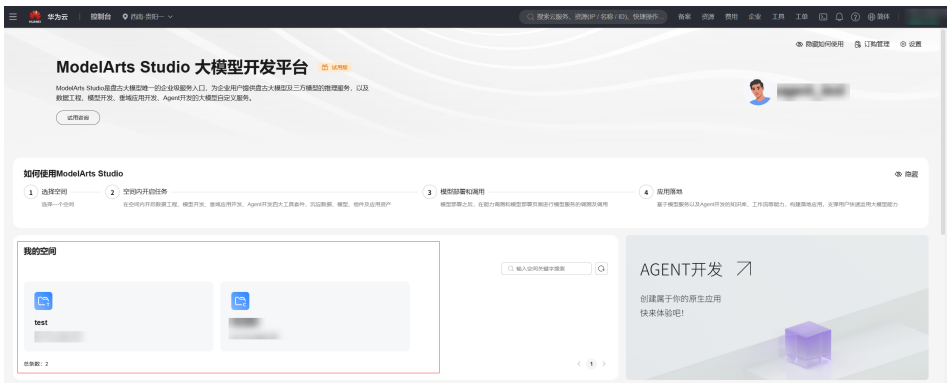
表 4-24 不同角色支持的数据标注任务权限清单

角色名称	创建标注任务	标注作业任务	审核作业任务	任务管理任务
超级管理员	√	√	-	√
管理员	√	√	-	√
标注管理员	√	√	-	√
标注作业员	-	√	-	-
标注审核员	-	-	√	-

创建文本类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-23 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。

版权所有 © 华为云计算技术有限公司

3. 在“创建标注任务”页面选择需要标注的文本类数据集，并选择标注项。选择标注项时，不同类型的数据文件对应的标注项有所差异，可基于页面提示进行选择。

   其中，“单轮问答”标注项支持“AI辅助标注”功能。

   说明

   开启使用AI辅助标注前，请确保有已部署的NLP大模型，具体步骤详见**创建NLP大模型部署任务**。

4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表4-25**配置标注分配与审核。

   表 4-25 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
标注审核	是否审核	• 否，标注后不进行审核操作。 • 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。
	审核要求	• 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 • 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

5. 配置完成后，单击“完成创建”。
6. 在“标注任务”页面，单击当前标注任务的“作业”，可执行标注作业任务。

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

   如果需要将该标注任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
7. 进入标注页面后，逐一对数据进行标注。

   如**图4-24**，以标注单轮问答数据为例，需要逐一确认问题（Q）及答案（A）是否正确，如果问题或答案不正确，可以对其进行二次编辑。

   图 4-24 文本类数据集标注示例



8. 一条数据标注完成后，单击“提交”可继续标注剩余数据。所有数据标注完成后，页面会出现标注任务成功的提示。

版权所有 © 华为云计算技术有限公司

说明

在完成数据标注后，如果无需进行标注审核，可直接在“标注任务 > 任务管理”页签单击“生成”，生成加工数据集。

生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

AI 辅助功能介绍

标注任务中，标注员对单条数据可选择地执行标注数据的智能生成，具体操作如下。

步骤1 在“创建标注任务”页面，选择数据集、开启AI辅助并选择模型后单击“完成创建”。

版权所有 © 华为云计算技术有限公司

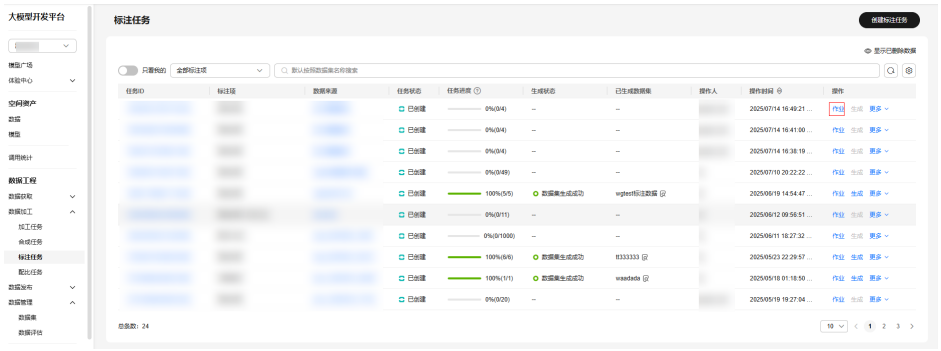
图 4-25 AI 辅助功能



步骤2 在“标注任务”页面，单击任务名称对应操作列的“作业”，可进入标注界面。

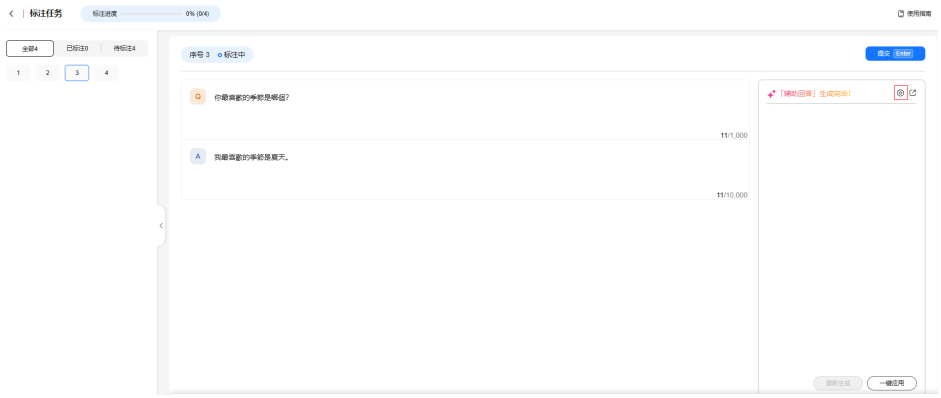
版权所有 © 华为云计算技术有限公司

图 4-26 标注任务



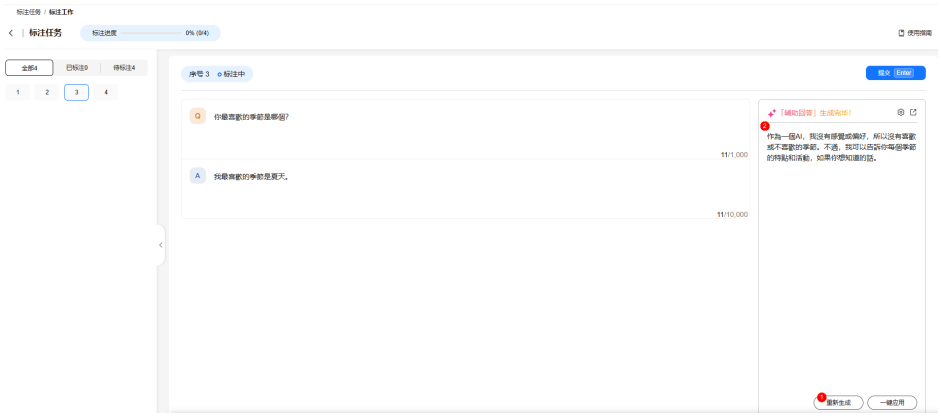
步骤3 单击“设置”按钮，修改AI标注模型后单击“确定”。

图 4-27 设置



步骤4 单击“重新生成”，生成新的答案。

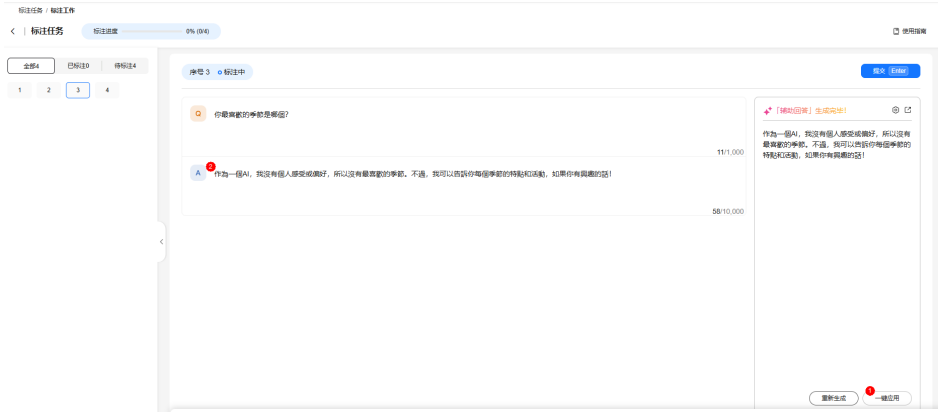
图 4-28 重新生成



步骤5 单击“一键应用”，重新生成的内容将会被显示在答案框里。

版权所有 © 华为云计算技术有限公司

图 4-29 一键应用



----结束

审核标注后的文本类数据集

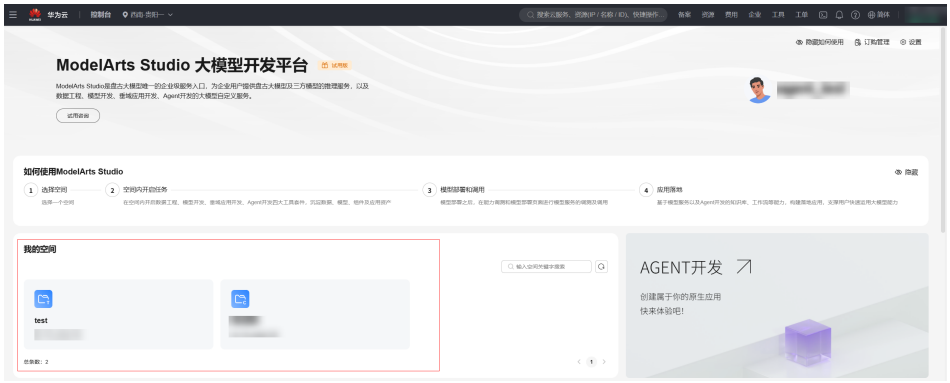
如果在**创建文本类数据集标注任务**时启用了标注审核功能，则在完成标注后可以审核标注结果。对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。

该操作需要具备**标注审核员**角色权限。

审核文本类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

   图 4-30 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

   如果需要将该审核任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
4. 进入审核页面后，可通过单击“通过”或“不通过”逐一对数据进行审核，直至所有数据审核完成。

说明

在完成数据标注审核后，需在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。生成后的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

版权所有 © 华为云计算技术有限公司

管理标注后的文本类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

- 生成：在完成数据标注审核后，需超级管理员、管理员、标注管理员角色在“标注任务”页面的操作列单击“生成”，生成加工数据集。

  生成后的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。
- 结束：管理员角色可在“标注任务”页面的操作列单击“更多 > 结束”，结束当前标注任务。
- 编辑：如果该标注任务支持“AI辅助标注”和“审核”功能，管理员角色可在“标注任务”页面的操作列单击“更多 > 编辑”，选择是否开启“AI辅助标注”功能并选择“审核要求”。
- 删除：管理员角色可在“标注任务”页面的操作列单击“更多 > 删除”，删除当前标注任务。

4.5.3.4 配比文本类数据集

数据配比是将多个数据集按照特定比例关系组合并发布为“发布数据集”的过程，确保数据的多样性、平衡性和代表性。

如果单个数据集已满足您的需求，**可跳过此章节至发布文本类数据集**。

创建文本类数据集配比任务

创建文本类数据集配比任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-31 我的空间

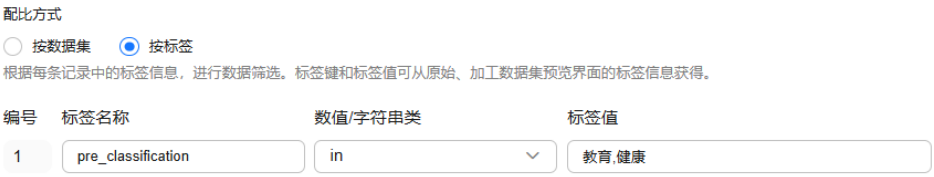
   

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 配比任务”，单击界面右上角“创建配比任务”。
3. 在“选择数据集”页签选择需要配比的文本类数据集（至少选择两个），单击“下一步”。
4. 在“数据配比”页面，支持两种配比方式，“按数据集”和“按标签”。
   - 按数据集：可以设置不同数据集的配比数量，单击“确定”。
   - 按标签：该场景适用于通过数据打标类加工算子进行加工的文本类数据集，具体标签名称与标签值可在完成**加工文本类数据集**操作后，进入数据集详情页面获取。

版权所有 © 华为云计算技术有限公司

填写示例如图4-32所示。

图 4-32 “按标签”配比方式填写示例



5. 数据配比完成后，单击右下角“下一步”，进入资源配置，并选择是否勾选自动生成加工数据集。

资源配置：

- 可以设置单元个数，单元个数越多运行效率越高。
- 下拉高级设置，可以设置任务资源 。也支持自定义参数配置，单击添加参数，输入参数名称和参数值。

图 4-33 资源配置

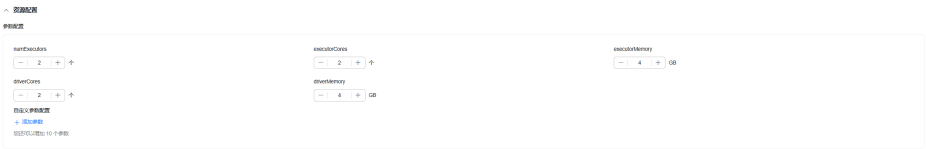


表 4-26 参数配置

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。

版权所有 © 华为云计算技术有限公司

参数名称	参数说明
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

自动生成加工数据集：

勾选，配置生成加工数据集的信息，如图 自动生成加工数据集；单击右下角“确定”，平台将启动配比任务。配比任务运行成功后自动生成加工数据集。

不勾选，单击右下角“确定”，平台将启动配比任务。配比任务运行成功后手动生成加工数据集。

图 4-34 自动生成加工数据集



6. 单击“确定”后，页面将返回至“配比任务”页面，配比任务运行成功后，状态将显示为“运行成功”。
7. 单击操作列“生成”，将生成“发布数据集”。

   发布数据集可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

4.5.4 加工图片类数据集

版权所有 © 华为云计算技术有限公司

4.5.4.1 加工图片类数据集

加工图片类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建图片类数据集加工任务

创建图片类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-35 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的图片类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于图片类数据集，可选择的加工算子请参见**表4-18**。

   a. 在左侧“添加算子”分页勾选所需算子。

   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

   图 4-36 算子编排

   c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

   若选择使用加工模板，将删除当前已编排的加工步骤。

版权所有 © 华为云计算技术有限公司

图 4-37 选择加工模板



5. 加工步骤编排完成后，单击“下一步”进入“任务配置界面”。
   - 资源配置

     下拉高级设置，可以设置任务资源 。也支持自定义参数配置，单击“添加参数”，输入参数名称和参数值。

图 4-38 资源配置

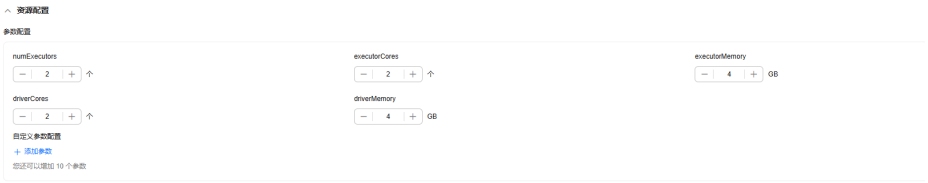


表 4-27 参数配置

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。

版权所有 © 华为云计算技术有限公司

参数名称	参数说明
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

- 自动生成加工数据集

  勾选，配置生成加工数据集的信息，如图4-39；单击右下角“确定”，平台将启动加工任务。加工任务运行成功后自动生成加工数据集。

  不勾选，单击右下角“确定”，平台将启动加工任务。加工任务运行成功后需要手动生成加工数据集。

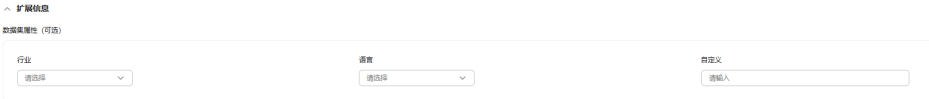
  图 4-39 自动生成加工数据集

- 扩展信息（可选）

  可选择行业、语言信息，或自定义数据集属性。

  图 4-40 扩展信息

版权所有 © 华为云计算技术有限公司

6. 单击“启动加工”。当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

   说明

   - 加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.4.2 标注图片类数据集

创建图片类数据集标注任务

标注图片类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表4-28**。

表 4-28 不同角色支持的数据标注任务权限清单

角色名称	创建标注任务	标注作业任务	审核作业任务	任务管理任务
超级管理员	√	√	-	√
管理员	√	√	-	√
标注管理员	√	√	-	√
标注作业员	-	√	-	-
标注审核员	-	-	√	-

创建图片类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-41 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。

版权所有 © 华为云计算技术有限公司

3. 在“创建标注任务”页面选择需要标注的图片类数据集与标注项。

   如果选择“图片Caption”标注项，则可开启“AI预标注”功能。AI预标注将自动生成标注内容，不会覆盖原始数据集，供标注人员参考，以提高标注效率。

4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表4-29**配置标注分配与审核。

   表 4-29 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
	标注要求	选择标注项为“图片Caption”且开启AI预标注功能时，可设置以下两种方式的“标注要求”： ● 选择“全部标注”：要求标注人员需要对全部的数据进行人工标注后才可提交标注结果。 ● 选择“可部分标注”：允许标注人员在确认AI预标注满足要求后，直接使用AI预标注功能完成数据集的标注并提交标注结果。
标注审核	是否审核	● 否，标注后不进行审核操作。 ● 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。
	审核要求	● 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 ● 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

5. 配置完成后，单击“完成创建”。
6. 在“标注任务”页面，单击当前标注任务的“作业”，可执行标注作业任务。

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

   如果需要将该标注任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
7. 进入标注页面后，逐一对数据进行标注。
8. 一条数据标注完成后，单击“提交”可继续标注剩余数据。所有数据标注完成后，页面会出现标注任务成功的提示。

   如果在创建标注任务时设置了“AI预标注 > 可部分标注”，则可在标注部分数据后，单击右上角的“提交全部标注数据”，让AI大模型自动标注剩余数据。

 版权所有 © 华为云计算技术有限公司

说明

在完成数据标注后，如果无需进行标注审核，可直接在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。

生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

审核标注后的图片类数据集

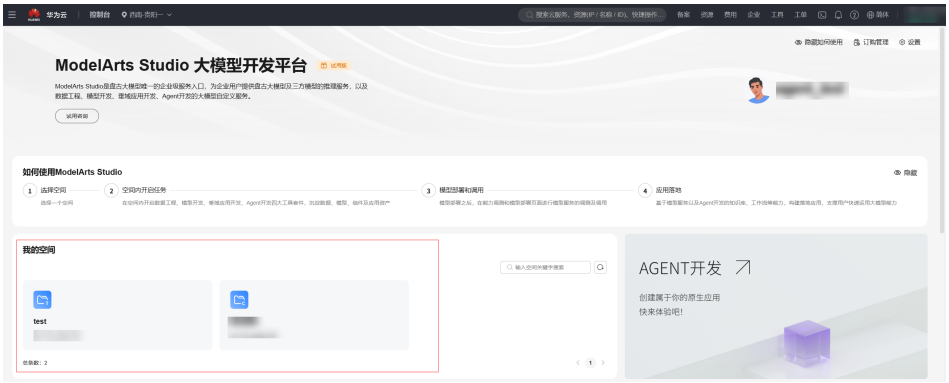
如果在创建图片类数据集标注任务时启用了标注审核功能，则在完成标注后可以审核标注结果。对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。

该操作需要具备**标注审核员**角色权限。

审核视频类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-42 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

   如果需要将该审核任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
4. 进入审核页面后，可通过单击“通过”或“不通过”逐一对数据进行审核，直至所有数据审核完成。

说明

在完成数据标注审核后，需在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。

生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

管理标注后的图片类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

- 生成：在完成数据标注审核后，需超级管理员、管理员、标注管理员角色在“标注任务”页面的操作列单击“生成”，生成加工数据集。

  生成后的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。
- 结束：管理员角色可在“标注任务”页面的操作列单击“更多 > 结束”，结束当前标注任务。

版权所有 © 华为云计算技术有限公司

- 编辑：如果该标注任务支持“AI辅助标注”和“审核”功能，管理员角色可在“标注任务”页面的操作列单击“更多 > 编辑”，选择是否开启“AI辅助标注”功能并选择“审核要求”。
- 删除：管理员角色可在“标注任务”页面的操作列单击“更多 > 删除”，删除当前标注任务。

4.5.4.3 配比图片类数据集

数据配比是将多个数据集按照特定比例关系组合并发布为“发布数据集”的过程，确保数据的多样性、平衡性和代表性。

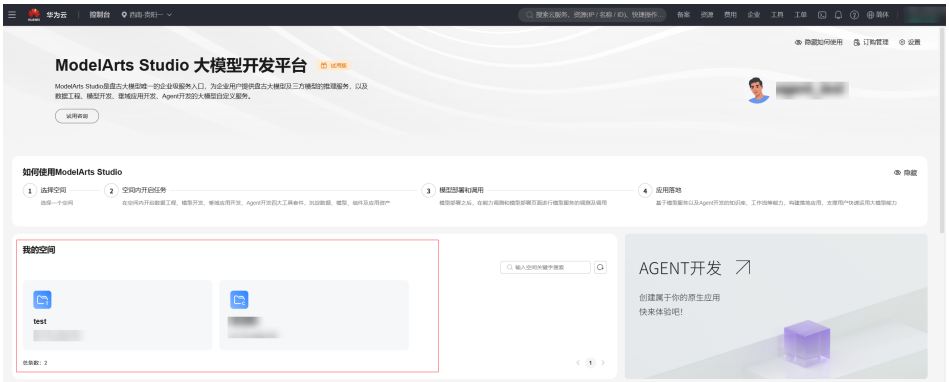
如果单个数据集已满足您的需求，**可跳过此章节至发布图片类数据集**。

创建图片类数据集配比任务

创建图片类数据集配比任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

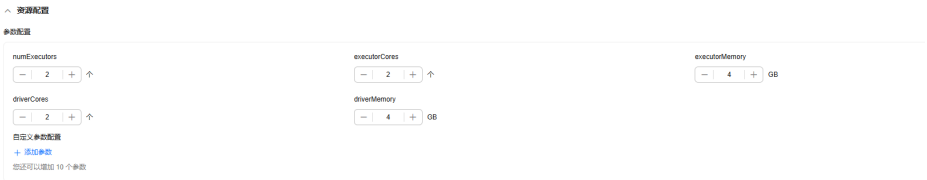
   图 4-43 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 配比任务”，单击界面右上角“创建配比任务”。
3. 在“数据集选择”页签选择需要配比的图片类数据集（至少选择两个），单击“下一步”。
4. 在“数据配比”页面，可以设置不同数据集的配比数量，单击“下一步”。
5. 数据配比完成后，单击右下角“下一步”，进行资源配置，并选择是否勾选自动生成加工数据集。
   - 资源配置

     下拉高级设置，可以设置任务资源 。也支持自定义参数配置，单击“添加参数”，输入参数名称和参数值。

     图 4-44 资源配置

版权所有 © 华为云计算技术有限公司

表 4-30 参数配置

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

- 自动生成加工数据集

  勾选，配置生成加工数据集的信息，如图4-45；单击右下角“确定”，平台将启动配比任务。配比任务运行成功后自动生成加工数据集。

  不勾选，单击右下角“确定”，平台将启动配比任务。配比任务运行成功后手动生成加工数据集。

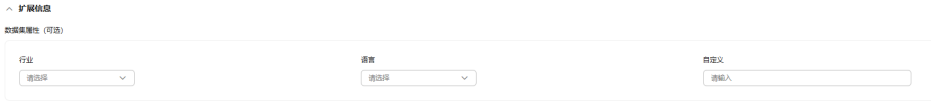
版权所有 © 华为云计算技术有限公司

图 4-45 自动生成加工数据集



- 扩展信息（可选）
  可选择行业、语言信息，或自定义数据集属性。

图 4-46 扩展信息



6. 单击“确定”，页面将返回至“数据配比”页面，配比任务运行成功后，状态将显示为“运行成功”。
7. 单击操作列“生成”，将生成“发布数据集”。
   发布数据集可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

4.5.5 加工视频类数据集

4.5.5.1 加工视频类数据集

加工视频类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建视频类数据集加工任务

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-47 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的视频类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于视频类数据集，可选择的加工算子请参见**表4-17**。

   a. 在左侧“添加算子”分页勾选所需算子。

      平台支持选择预置算子或自定义算子，如果使用自定义算子，创建步骤见**自定义数据集加工算子**。

   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

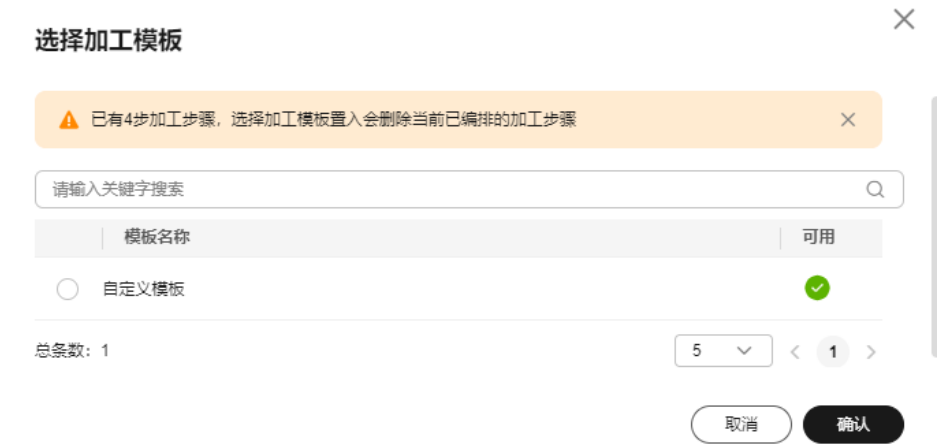
      图 4-48 算子编排

   c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

      若选择使用加工模板，将删除当前已编排的加工步骤。

      图 4-49 选择加工模板

5. 加工步骤编排完成后，单击“下一步”进入“任务配置”界面。

   – 自动生成加工数据集

      勾选，配置生成加工数据集的信息，如**图4-63**；单击右下角“确定”，平台将启动加工任务。加工任务运行成功后自动生成加工数据集。

      不勾选，单击右下角“确定”，平台将启动加工任务。加工任务运行成功后需要手动生成加工数据集。

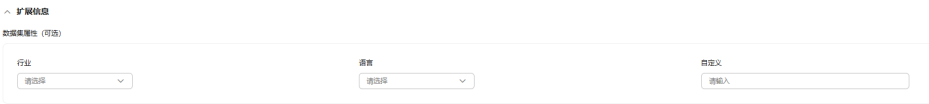
 版权所有 © 华为云计算技术有限公司

图 4-50 自动生成加工数据集



- 扩展信息（可选）
  可选择行业、语言信息，或自定义数据集属性。

图 4-51 扩展信息



6. 单击“启动加工”。当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

   说明

   - 加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.5.2 标注视频类数据集

创建视频类数据集标注任务

标注视频类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表4-31**。

表 4-31 不同角色支持的数据标注任务权限清单

角色名称	创建标注任务	标注作业任务	审核作业任务	任务管理任务
超级管理员	√	√	-	√
管理员	√	√	-	√
标注管理员	√	√	-	√
标注作业员	-	√	-	-

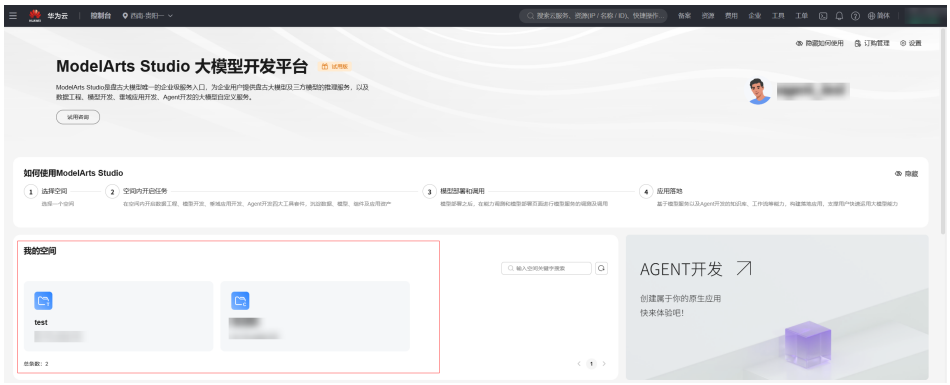
版权所有 © 华为云计算技术有限公司

角色名称	创建标注任务	标注作业任务	审核作业任务	任务管理任务
标注审核员	-	-	√	-

创建视频类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-52 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。
3. 在“创建标注任务”页面选择需要标注的视频类数据集与标注项，单击“下一步”。

   如果选择经过数据加工视频摘要生成算子处理过的视频+标注数据集，可以选择“打标结果修正”标注项，对算子打标的摘要结果进行修正检查。
4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表4-32**配置标注分配与审核。

   表 4-32 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
标注审核	是否审核	• 否，标注后不进行审核操作。 • 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。
	审核要求	• 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 • 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

版权所有 © 华为云计算技术有限公司

5. 配置完成后，单击“完成创建”。
6. 在“标注任务”页面，单击当前标注任务的“作业”，可执行标注作业任务。
   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。
   如果需要将该标注任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
7. 进入标注页面后，逐一对数据进行标注。
8. 一条数据标注完成后，单击“提交”可继续标注剩余数据。所有数据标注完成后，页面会出现标注任务成功的提示。

说明

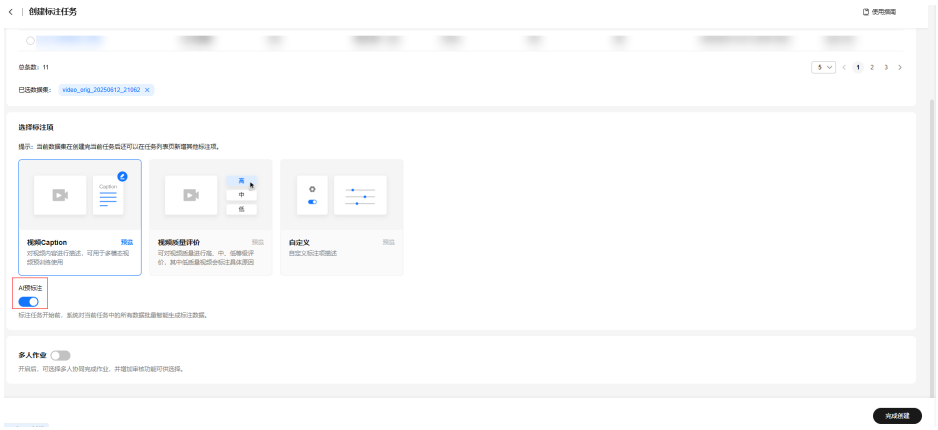
在完成数据标注后，如果无需进行标注审核，可直接在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。
生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

AI 预标注功能介绍

标注任务开始前，系统对当前任务中的所有数据批量智能生成标注数据，具体操作如下。

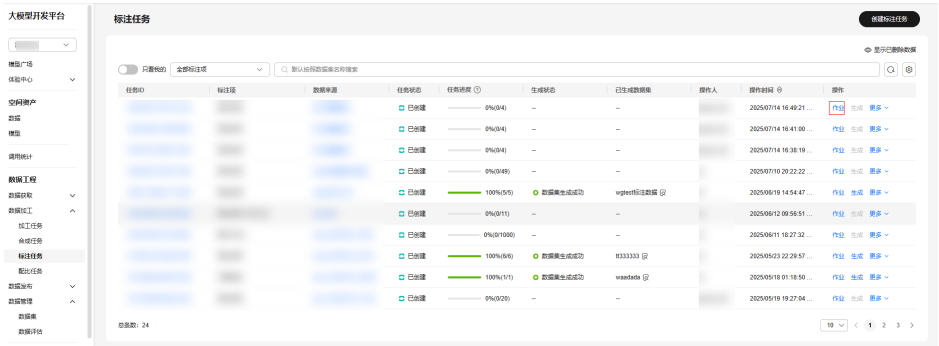
步骤1 在“创建标注任务”页面，选择数据集、开启AI预标注后单击“完成创建”。

图 4-53 AI 预标注功能



步骤2 在“标注任务”页面，单击任务名称对应操作列的“作业”，可进入标注界面。

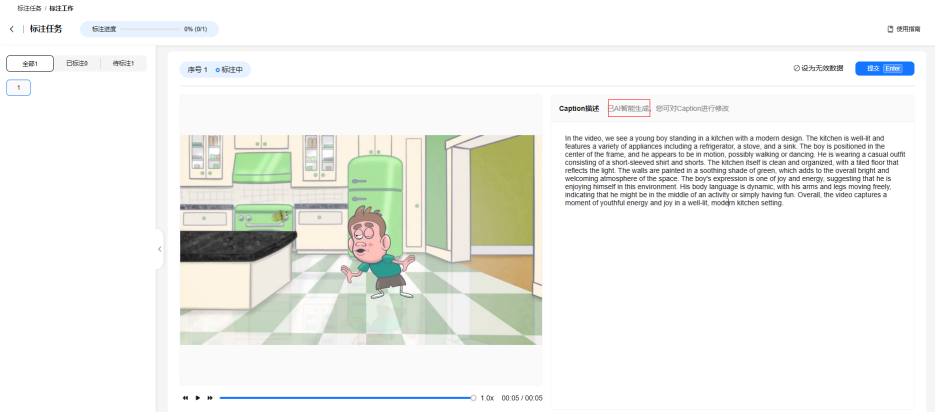
图 4-54 标注任务



版权所有 © 华为云计算技术有限公司

步骤3 “Caption描述”框中已AI智能生成描述，可对Caption进行修改后单击右上角的“提交”。

图 4-55 Caption 描述



----结束

审核标注后的视频类数据集

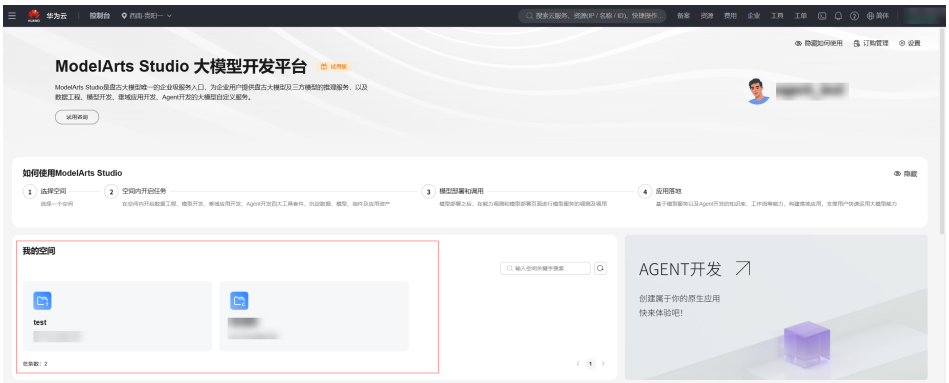
如果在**创建视频类数据集标注任务**时启用了标注审核功能，则在完成标注后可以审核标注结果。对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。

该操作需要具备**标注审核员**角色权限。

审核视频类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-56 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

   如果需要将该审核任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
4. 进入审核页面后，可通过单击“通过”或“不通过”逐一对数据进行审核，直至所有数据审核完成。

版权所有 © 华为云计算技术有限公司

说明

在完成数据标注审核后，需在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

管理标注后的视频类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

- 生成：在完成数据标注审核后，需超级管理员、管理员、标注管理员角色在“标注任务”页面的操作列单击“生成”，生成加工数据集。

  生成后的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。
- 结束：管理员角色可在“标注任务”页面的操作列单击“更多 > 结束”，结束当前标注任务。
- 编辑：如果该标注任务支持“AI辅助标注”和“审核”功能，管理员角色可在“标注任务”页面的操作列单击“更多 > 编辑”，选择是否开启“AI辅助标注”功能并选择“审核要求”。
- 删除：管理员角色可在“标注任务”页面的操作列单击“更多 > 删除”，删除当前标注任务。

4.5.5.3 配比视频类数据集

数据配比是将多个数据集按照特定比例关系组合并发布为“发布数据集”的过程，确保数据的多样性、平衡性和代表性。

如果单个数据集已满足您的需求，**可跳过此章节至发布视频类数据集**。

创建视频类数据集配比任务

创建视频类数据集配比任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-57 我的空间



2. 支持“视频+标注”配比，在左侧导航栏中选择“数据工程 > 数据加工 > 配比任务”，单击界面右上角“创建配比任务”。

版权所有 © 华为云计算技术有限公司



3. 在“数据选择”页签选择需要配比的视频类数据集（至少选择两个），单击“下一步”。
4. 在“配比数据集”页面，可以设置不同数据集的配比数量，单击“下一步”。
5. 数据配比完成后，单击右下角“下一步”，进行任务配置，并选择是否勾选自动生成加工数据集。
   - 自动生成加工数据集

     勾选，配置生成加工数据集的信息，如图4-58；单击右下角“确定”，平台将启动配比任务。配比任务运行成功后自动生成加工数据集。

     不勾选，单击右下角“确定”，平台将启动配比任务。配比任务运行成功后手动生成加工数据集。

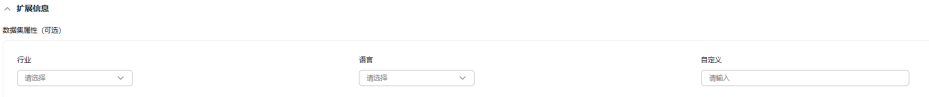
     图 4-58 自动生成加工数据集

   - 扩展信息（可选）

     可选择行业、语言信息，或自定义数据集属性。

     图 4-59 扩展信息

版权所有 © 华为云计算技术有限公司

6. 单击“确定”，页面将返回至“数据配比”页面，配比任务运行成功后，状态将显示为“运行成功”。
7. 加工数据集可在配比任务栏-“已生成数据集”中查看，亦可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.6 加工气象类数据集

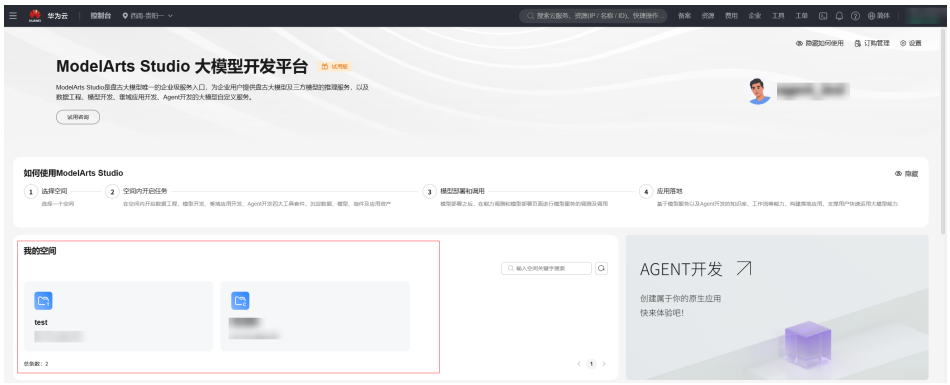
加工气象类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建气象类数据集加工任务

创建气象类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-60 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的气象类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于气象类数据集，可选择的加工算子请参见**表 4-19**。
   a. 在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

版权所有 © 华为云计算技术有限公司

图 4-61 算子编排



c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

若选择使用加工模板，将删除当前已编排的加工步骤。

图 4-62 选择加工模板



5. 加工步骤编排完成后，单击“下一步”进入“任务配置界面”。
   - 自动生成加工数据集

     勾选，配置生成加工数据集的信息，如图4-63；单击右下角“确定”，平台将启动加工任务。加工任务运行成功后自动生成加工数据集。

     不勾选，单击右下角“确定”，平台将启动加工任务。加工任务运行成功后需要手动生成加工数据集。

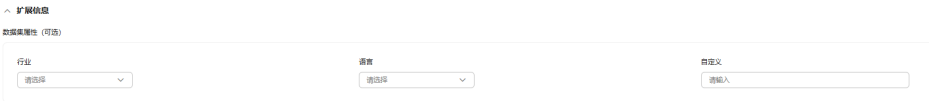
 版权所有 © 华为云计算技术有限公司

图 4-63 自动生成加工数据集



- 扩展信息（可选）
  可选择行业、语言信息，或自定义数据集属性。

图 4-64 扩展信息



6. 单击“启动加工”。当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

   说明

   - 在完成数据加工后，如果无需使用数据标注、数据合成功能，可直接在“加工任务”页面单击操作列“生成”，生成加工数据集。
   - 加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.7 加工预测类数据集

4.5.7.1 加工预测类数据集

加工预测类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建预测类数据集加工任务

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

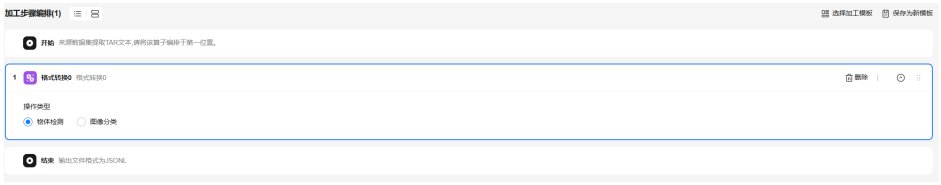
版权所有 © 华为云计算技术有限公司

图 4-65 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的预测类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于预测类数据集，仅支持自定义算子，创建步骤见**自定义数据集加工算子**。
   a. 选择预置算子，在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

   图 4-66 算子编排

   c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

   若选择使用加工模板，将删除当前已编排的加工步骤。

   图 4-67 选择加工模板

版权所有 © 华为云计算技术有限公司

5. 加工步骤编排完成后，单击“下一步”进入“任务配置”界面。
   - 自动生成加工数据集

     勾选，配置生成加工数据集的信息，如图4-63；单击右下角“确定”，平台将启动加工任务。加工任务运行成功后自动生成加工数据集。

     不勾选，单击右下角“确定”，平台将启动加工任务。加工任务运行成功后需要手动生成加工数据集。

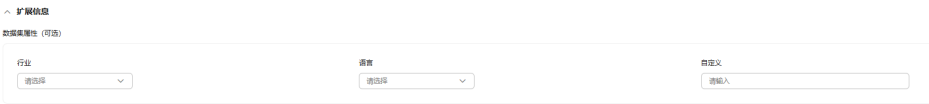
     图 4-68 自动生成加工数据集

   - 扩展信息（可选）

     可选择行业、语言信息，或自定义数据集属性。

     图 4-69 扩展信息

6. 单击“启动加工”。当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

   说明

   - 加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.7.2 配比预测类数据集

预测数据配比是将单个数据集中目标列各分类值按照特定比例关系组合并生成“加工数据集”的过程，确保数据的多样性、平衡性和代表性。

如果单个数据集已满足您的需求，**可跳过此章节至发布预测类数据集**。

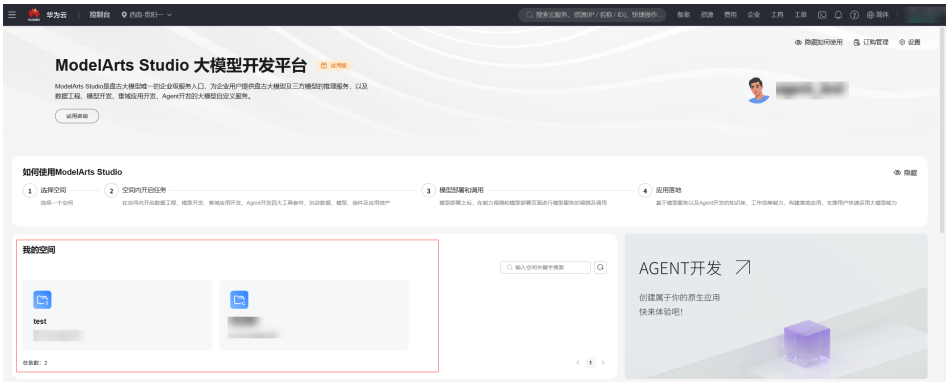
版权所有 © 华为云计算技术有限公司

创建预测类数据集配比任务

创建预测类数据集配比任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-70 我的空间



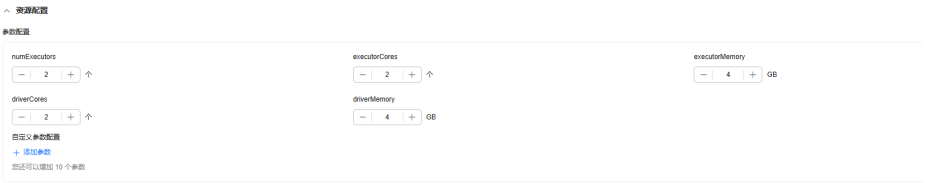
2. 仅分类数据支持配比，在左侧导航栏中选择“数据工程 > 数据加工 > 配比任务”，单击界面右上角“创建配比任务”。



3. 在“数据集选择”页签选择需要配比的预测类数据集，单击“下一步”。
4. 在“数据配比”页面，可以设置数据集的配比方式，如按目标列。根据选择的目标列，为分类值设置配比数量。填写好配比数量，单击“下一步”。
5. 数据配比完成后，单击右下角“下一步”，进行资源配置，并选择是否勾选自动生成加工数据集。
   - 资源配置

     下拉高级设置，可以设置任务资源 。也支持自定义参数配置，单击“添加参数”，输入参数名称和参数值。

图 4-71 资源配置



版权所有 © 华为云计算技术有限公司

参数说明见表4-33：

表 4-33 参数配置

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

- 自动生成加工数据集

  勾选，配置生成加工数据集的信息，如图4-45；单击右下角“确定”，平台将启动配比任务。配比任务运行成功后自动生成加工数据集。

  不勾选，单击右下角“确定”，平台将启动配比任务。配比任务运行成功后手动生成加工数据集。

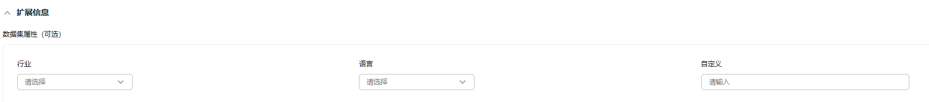
版权所有 © 华为云计算技术有限公司

图 4-72 自动生成加工数据集



- 扩展信息（可选）
  可选择行业、语言信息，或自定义数据集属性。

  图 4-73 扩展信息



6. 单击“确定”，页面将返回至“数据配比”页面，配比任务运行成功后，状态将显示为“运行成功”。
7. 加工数据集可在配比任务栏-“已生成数据集”中查看，亦可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.8 加工音频类数据集

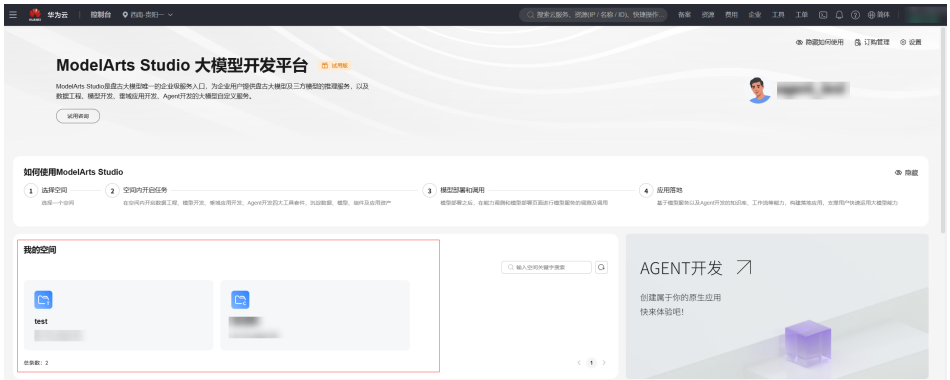
4.5.8.1 加工音频类数据集

加工音频类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

创建视频类数据集加工任务

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

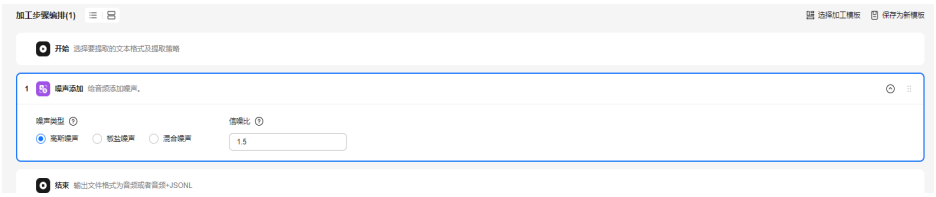
   图 4-74 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的音频类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于音频类数据集，可选择的加工算子请参见**音频类加工算子介绍**。
   a. 在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

   图 4-75 算子编排

   c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

   若选择使用加工模板，将删除当前已编排的加工步骤。

   图 4-76 选择加工模板

5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。

   当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

   说明

   加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

 版权所有 © 华为云计算技术有限公司

4.5.8.2 标注音频类数据集

创建音频类数据集标注任务

标注音频类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表1 不同角色支持的数据标注任务权限清单**。

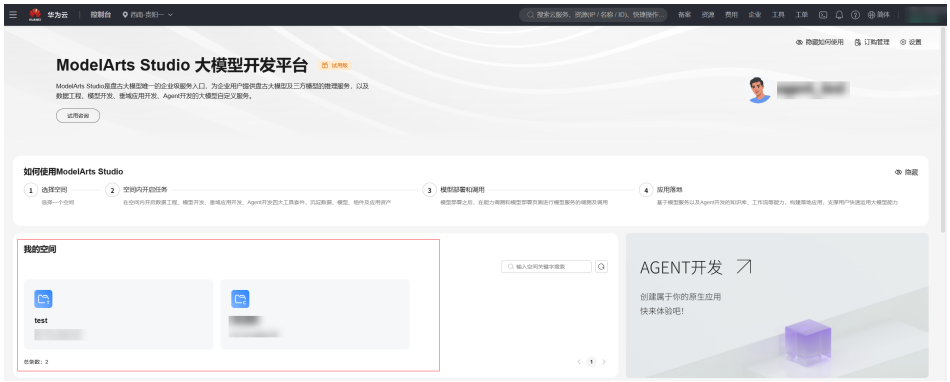
表 4-34 不同角色支持的数据标注任务权限清单

角色名称	创建标注任务	标注作业任务	审核作业任务	任务管理任务
超级管理员	√	√	-	√
管理员	√	√	-	√
标注管理员	√	√	-	√
标注作业员	-	√	-	-
标注审核员	-	-	√	-

创建音频类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 4-77 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。
3. 在“创建标注任务”页面选择需要标注的音频类数据集与标注项，单击“下一步”。
4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表4-35**配置标注分配与审核。

版权所有 © 华为云计算技术有限公司

表 4-35 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
标注审核	是否审核	• 否，标注后不进行审核操作。 • 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。
	审核要求	• 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 • 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

5. 配置完成后，单击“完成创建”。
6. 在“标注任务”页面，单击当前标注任务的“作业”，可执行标注作业任务。

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

   如果需要将该标注任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
7. 进入标注页面后，逐一对数据进行标注。
8. 一条数据标注完成后，单击“提交”可继续标注剩余数据。所有数据标注完成后，页面会出现标注任务成功的提示。

说明

在完成数据标注后，如果无需进行标注审核，可直接在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。

生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

审核标注后的音频类数据集

如果在**创建音频类数据集标注任务**时启用了标注审核功能，则在完成标注后可以审核标注结果。对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。

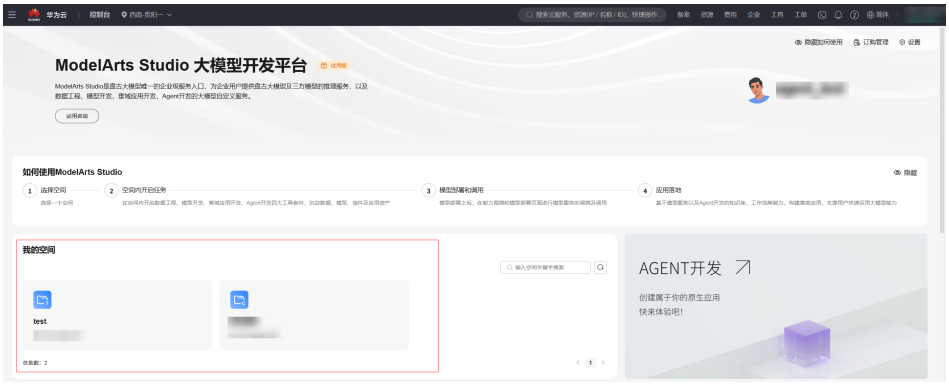
该操作需要具备**标注审核员**角色权限。

审核音频类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-78 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

   如果需要将该审核任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。
4. 进入审核页面后，可通过单击“通过”或“不通过”逐一对数据进行审核，直至所有数据审核完成。

说明

在完成数据标注审核后，需在“数据标注 > 任务管理”页面单击“生成”，生成加工数据集。生成的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

管理标注后的音频类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

- 生成：在完成数据标注审核后，需超级管理员、管理员、标注管理员角色在“标注任务”页面的操作列单击“生成”，生成加工数据集。

  生成后的加工数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。
- 结束：管理员角色可在“标注任务”页面的操作列单击“更多 > 结束”，结束当前标注任务。
- 编辑：如果该标注任务支持“AI辅助标注”和“审核”功能，管理员角色可在“标注任务”页面的操作列单击“更多 > 编辑”，选择是否开启“AI辅助标注”功能并选择“审核要求”。
- 删除：管理员角色可在“标注任务”页面的操作列单击“更多 > 删除”，删除当前标注任务。

4.5.9 加工其他类数据集

加工其他类数据集任务前，请先完成数据导入操作，具体步骤请参见**导入数据至盘古平台**。

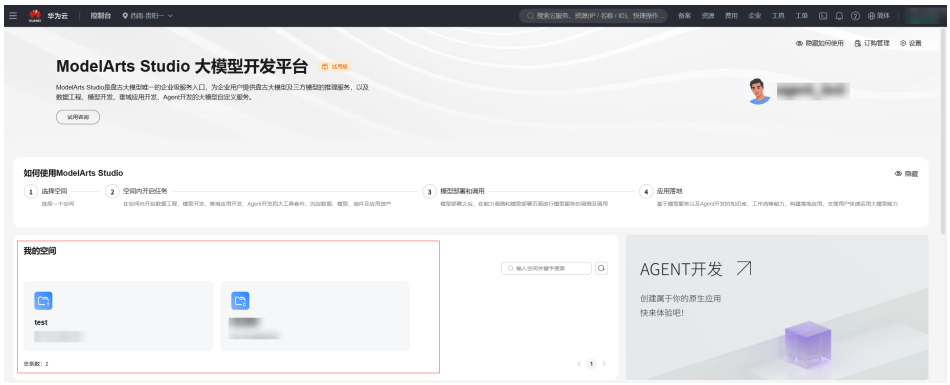
创建其他类数据集加工任务

创建其他类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-79 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的其他类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。
   a. 在左侧“添加算子”分页勾选所需算子。

      对于其他类数据集，平台仅支持选择自定义算子，其创建步骤见**自定义数据集加工算子**。

   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。
   c. 在编排过程中，可单击右上角“保存为新模板”将当前编排流程保存为模板。后续创建新的数据加工任务时，可直接单击“选择加工模板”进行使用。

      若选择使用加工模板，将删除当前已编排的加工步骤。

      图 4-80 选择加工模板



5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。

   当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。
6. 在完成数据加工后，在“加工任务”页面单击操作列“生成”，生成加工数据集。

 版权所有 © 华为云计算技术有限公司

加工后的数据集可在“数据工程 > 数据管理 > 数据集 > 加工数据集”中查看。

4.5.10 自定义数据集加工算子

除了预置的加工算子外，ModelArts Studio平台还支持用户创建自定义加工算子，以满足特定的数据处理需求和业务场景，用户可以根据实际情况灵活定义处理逻辑，进一步提升模型训练的效果和适应性。

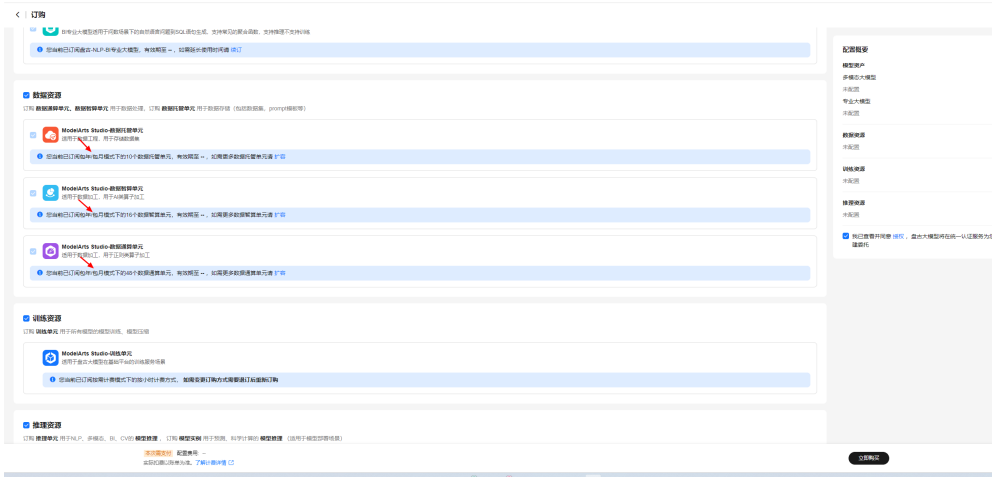
约束限制

该功能仅对包年/包月订购用户开放。

图 4-81 订购服务



图 4-82 包年包月资源



版权所有 © 华为云计算技术有限公司

4.5.10.1 算子配置文件规范

表 4-36 基本信息配置规范

参数名称	类型	是否必填	描述	约束
id	string	是	算子英文简称。	以英文字符开头，可包含英文字符、阿拉伯数字和下划线“_”，字符长度不超过128，超过部分自动截断。 算子更新配置时不可修改id。 全空间可见算子id全空间唯一，本空间可见算子id在当前空间下唯一。
name	string	是	算子展示名称。	字符长度不超过128，超过部分自动截断。
description	string	否	算子描述。	字符长度不超过2000，超过部分自动截断。
author	string	否	开发者名称。	字符长度不超过128，超过部分自动截断。
version	string	是	算子版本。	字符长度不超过32，超过部分自动截断，格式为：数字.数字.数字。
tags	**tags** object	是	算子标签，用于分类筛选。	见**表 label配置规范**
labels	Array of **label** objects	否	打标算子输出的标签。	见**表 label配置规范**。
runtime	**runtime** object	是	算子运行配置。	见**表 label配置规范**。
arguments	Array of **argument** objects	否	算子输入参数列表。	见**表4-40**。

表 4-37 tag 配置规范

参数名称	类型	是否必填	描述	约束
language	Array of strings	是	算子支持处理的数据语种列表，例如zh，en等。	单个字符串长度不超过32，超过部分自动截断。 填写国际语种简码。

版权所有 © 华为云计算技术有限公司

参数名称	类型	是否必填	描述	约束
format	Array of strings	是	算子支持处理的数据集文件格式后缀列表，例如json，csv，mp4等。	单个字符串长度不超过32，超过部分自动截断。
category	string	是	算子类别。	单选，可选取值如下： - 数据提取 - 数据抽样 - 数据转换 - 数据过滤 - 数据去重 - 数据打标 - 其他
modal	Array of strings	是	算子支持处理的数据模态列表。	多选，可选取值如下： - TEXT（文本） - IMAGE（图片） - VIDEO（视频） - AUDIO（音频） - OTHER（含气象、预测）
custom	Array of strings	否	用户自定义算子标签列表。	单个字符串长度不超过32，超过部分自动截断。

表 4-38 runtime 配置规范

参数名称	类型	是否必填	描述	约束
cpu-arch	Array of strings	是	算子运行支持的CPU架构列表。	多选，可选取值如下： - ARM - X86
xpu-devices	Array of strings	否	算子运行的支持卡设备型号列表，当算子运行涉及NPU时必填。	可选取值如下： - SNT9B 当参数runtime.resources中存在某个resource的npu个数大于0的情况时，该参数必填，可选值为SNT9B。
environment	string	是	算子包类型，仅支持纯python算子包类型。	单选，可选取值如下： - PYTHON

版权所有 © 华为云计算技术有限公司

参数名称	类型	是否必填	描述	约束
entrypoint	string	是	算子运行的启动命令。	字符长度不超过128，超过部分自动截断。 当environment=python时填写固定名称process.py，暂不支持修改。
auto-data-loading	boolean	是	是否自动加载数据	当值为true时，由框架处理输入输出；当值为false时，由用户自己处理输入输出。
resources	Array of **resource** objects	否	算子单实例运行所需的资源大小配置列表。	当environment=python时必填。

表 4-39 resource 配置规范

参数名称	类型	是否必填	描述	约束
cpu	integer	是	单实例默认CPU资源量（单位：vCPU）。	用户需结合实际订购的通算单元数量进行配置。
memory	integer	是	单实例默认memory资源量（单位：MB）。	用户需结合实际订购的通算单元数量进行配置。
npu	integer	否	单实例默认NPU资源量（单位：卡）。	用户需结合实际订购的智算单元规格和数量进行配置。

表 4-40 argument 配置规范

参数名称	类型	是否必填	描述	约束
key	string	是	参数id，参数列表内必须唯一。	字符长度不超过128，超过部分自动截断。
name	string	是	参数展示名称。	字符长度不超过128，超过部分自动截断。

版权所有 © 华为云计算技术有限公司

参数名称	类型	是否必填	描述	约束
type	string	是	参数类型.	单选，可选取值如下： - STRING - FLOAT - INT - ENUM（单选） - LIST（复选） - OBS（前端使用，选择OBS文件的前端组件参数） - BOOLEAN
tips	string	否	参数说明。	字符长度不超过2000，超过部分自动截断。
min	float	否	参数最小值。当type=INT或FLOAT时选填。	浮点数最多保留4位小数。
max	float	否	参数最小值。当type=INT或FLOAT时选填。	浮点数最多保留4位小数。
between	boolean	否	参数是否为范围型数值参数。当type=INT或FLOAT时选填，默认为false。	-
items	Array of **item** objects	否	参数枚举项列表。当type=ENUM /LIST时必填。	当type=ENUM /LIST时该参数必填且枚举项个数至少为1。
required	boolean	否	参数是否必填。	填写true或者false。
visible	boolean	否	参数是否前端可见。	填写true或者false。
default	string	否	参数默认值。	当visible=false且requires=true时必须填默认值。 多个默认值时用逗号分隔，例如“SD,HD”。 数值范围型参数默认值格式为"min;max"。

版权所有 © 华为云计算技术有限公司

表 4-41 item 配置规范

参数名称	类型	是否必填	描述	约束
name	string	是	枚举项名称。	字符长度不超过128，超过部分自动截断。
value	string	是	枚举项对应的值。	字符长度不超过128，超过部分自动截断。
tips	string	否	枚举项描述或说明。	字符长度不超过2000，超过部分自动截断。

表 4-42 label 配置规范

参数名称	类型	是否必填	描述	约束
key	string	是	标签的key值。	字符长度不超过128，超过部分自动截断，key值在不同算子间唯一。
name	string	是	标签的名称。	字符长度不超过128，超过部分自动截断。
type	string	是	标签的类型。	枚举值，单选，仅支持STRING、NUMERIC、ENUM、OBJECT
min	float	否	标签最小值	当type为NUMERIC型时必填。
max	float	否	标签最大值	当type为NUMERIC型时必填。
items	Array of **labelItem** objects	否	标签枚举列表	当type为ENUM型时必填。
dimensions	Array of **labelDimension** objects	否	标签一级维度	当type为OBJECT型必填。

表 4-43 labelItem 配置规范

参数名称	类型	是否必填	描述	约束
name	string	是	枚举名称。	字符长度不超过128，超过部分自动截断。

版权所有 © 华为云计算技术有限公司

参数名称	类型	是否必填	描述	约束
value	string	是	枚举值。	字符长度不超过128，超过部分自动截断。

表 4-44 labelDimension 配置规范

参数名称	类型	是否必填	描述	约束
key	string	是	二级标签key。	字符长度不超过128，超过部分自动截断。
name	string	是	二级标签名称。	字符长度不超过128，超过部分自动截断。
type	string	是	二级标签类型。	枚举值，单选，仅支持STRING、NUMERIC、ENUM
min	float	否	二级标签的最小值。	当type为NUMERIC型时必填。
max	float	否	二级标签的最大值。	当type为NUMERIC型时必填。
items	Array of **labelItem** objects	否	二级标签的枚举列表。	当type为ENUM型时必填。

算子配置文件是一个YAML文件，用于描述算子的基本信息、功能参数、运行环境以及资源需求等。算子配置文件示例如下：

```
id: video_clip    #【必填】算子英文简称，英文字母开头，只能包含英文字符、数字和下划线，最多128个字符，超过做截断
name: 视频Clip切分 #【必填】算子展示名称，最多128个字符，超过做截断
description: 将长视频按镜头切分为多个片段。 #算子描述，不超过2000个字符
author: "xxx Technology Co., Ltd." #可选，开发者或团队名称，例如 "Data Team"，不超过128个字符
version: 1.0.0  # 规则 "\\d+\\.\\d+\\.\\d+$"
tags:  #【必填】算子标签，用于特定场景分类筛选
  language: #【必填】算子支持处理的语言类型, 语言简码，多选，选项只能选择国际标准语言简码，不超过32个字符
    - zh
    - en
    -...
  format: #【必填】算子支持处理的数据集文件格式，多选，不超过32个字符
      - JSONL
      - TXT
      - CSV
      - HTML
      - MOBI
      - EPUB
      - DOCX
      - PDF
      - MP4
      - AVI
```

 版权所有 © 华为云计算技术有限公司

```
    -...
  category: # 【必填】算子类型，用于前台的算子分类展示，单选，选项固定如下
    - 数据提取
    - 数据抽样
    - 数据转换
    - 数据过滤
    - 数据去重
    - 数据打标
    - 其他
  modal:  # 【必填】算子支持处理的数据模态。多选，选项固定如下
    - TEXT
    - IMAGE
    - VIDEO
    - AUDIO
    - OTHER
  custom: #用户自定义算子标签，多选，不超过32个字符
    - 数据增强
    - 预标注
    -...
  task-categories:
    - BACKFLOW  # 回流

runtime:
  cpu-arch:
    - ARM
  xpu-devices:
    - SNT9B
  # 算子资源配置：
  #   npu算子配置：1卡 24core 192GB
  #   cpu算子配置：1core 2G、1core 4G、2core 4G、4core 8G、8core 16G
  resources:
    - cpu: 1
      memory: 2048
    - cpu: 24
      memory: 196608 # 单位：MB
      npu: 1
  environment: python
  entrypoint: process.py
  auto-data-loading: true # 是否需要自动加载数据
#所有的业务参数均为算子的输入参数，参数的数据类型包含：STRING、FLOAT、INT、ENUM（单选框）、LIST（复选框）、OBS
arguments:
  # STING类型参数示例
  - key: filter_keywords  #【必填】
    name: 过滤关键词
    type: STRING
    tips: 命中的关键词所在的样本将会被过滤，多个关键词以英文字符","分隔。   #界面tips
    required: true
    visible: true
    default: 大保健  #参数默认值,可选，多个默认值使用,隔开

  # INT/FLOAT数值范围类型参数示例
  - key: length_of_characters  #【必填】
    name: 过滤时长范围
    type: FLOAT  #【必填】当type=int/float时，between字段为必填。
    between: true    # 是否取数值范围，当type=int/float时，between字段为必填。
    min: 1.0  #参数最小值的范围，可选；float最多【4位小数】，超过做截断
    max: 500.0    #参数最大值的范围，可选
    tips: 过滤时长范围包含填入的边界值，float类型，单位：字符。文本长度在指定范围内的样本将被保留。 #界面tips
    required: true
    visible: true
    default: 100.0;300.0  #可选, 参数最大最小值的默认值，可选，未填写时使用"min;max"作为默认

  # INT/FLOAT数值类型参数示例
  - key: max_cropping_area_ratio
    name: 最大裁剪面积占比
    type: FLOAT
    between: false # 是否是范围型，默认为false
```

 版权所有 © 华为云计算技术有限公司

```
  min: 0.0  #参数最小值的范围
  max: 100.0   #参数最大值的范围
  tips: 值在0.0~100.0之间，float类型，单位：%。裁剪面积占比（裁剪视频面积/原始视频面积）大于该值的样本将被过滤掉。
  visible: true
  required: true
  default: 100  #参数默认值

# ENUM类型参数示例
- key: font_conversion
  name: 文本字体转换
  type: ENUM  #单选
  items:
    - name: 简体转繁体
      value: traditional
    - name: 繁体转简体
      value: simplified
  required: true
  visible: true
  default: simplified

# LIST类型参数示例
- key: resolution
  name: 分辨率
  type: LIST  #可多选
  items:  #参数选项，ENUM/LIST类型必填
    - name: 流畅  #【必填】
      value: SM  #【必填】
      tips: 480>分辨率≥360
    - name: 标清
      value: SD
      tips: 720>分辨率≥480
    - name: 高清
      value: HD
      tips: 1080>分辨率≥720
  required: true
  visible: true
  default: SD,HD # 多个默认值使用,隔开

# OBS类型参数示例
- key: sensitive_word
  name: 敏感词库文件OBS路径
  type: OBS
  tips: 敏感词库文件。
  required: true
  visible: true
  default: NLP/system_resource/sensitive_word.csv # 默认词库的OBS路径。
# BOOLEAN类型参数示例
- key: parse_all
  name: 是否解析所有文件
  type: BOOLEAN
  items:
    - name: 是
      value: true
    - name: 否
      value: false
  visible: true
  required: true
  default: false
```

4.5.10.2 算子包开发规范

Python 算子包目录规范

假设算子包名称为video_clip.tar，算子包解压后目录结构：

```
+--- video_clip # 与tar包名称保证一致
| +--- program_package # python算子目录
```


版权所有 © 华为云计算技术有限公司

```
| | +--- install.sh # 安装脚本，可选
| | +--- process.py # 算子代码，必填
```

process.py 文件开发规范

算子包必须包含一个名为process.py脚本，根据算子配置文件中**auto-data-loading**字段值分为两种开发模式。

模式一（ auto-data-loading 为 true ）

适用场景：除了以下三类场景，推荐使用该模式。

1. 输入数据集为多模态数据集场景：例如图片+文本，视频+文本等由两种及以上模态组合的数据集。
2. 输出数据集模态不在以下范围内：文本、图片、视频、音频。
3. 需要将整个数据集作为输入的场景：例如去重类算。

process.py开发规范如下所示，process.py文件包含三个类：

1.PreProcess：算子预处理相关逻辑（ 选填 ）。

算子调用GPU/NPU做模型推理之前的、于CPU中计算的部分。拆分算子预处理逻辑有利于CPU计算与GPU/NPU计算的解耦，提高GPU/NPU利用率。

2.Process：算子推理相关逻辑 **（ 必填 ）**。

建议只包含模型的加载及推理部分，将预处理和后处理编写在PreProcess及PostProcess中。

3.PostProcess: 算子后处理相关逻辑 （ 可选 ）。

如算子在推理完成后，仍然有较重的、于CPU中计算的后处理逻辑，也建议将这部分逻辑拆分，编写在PostProcess中。

算子框架的调用顺序为preproces->process->postprocess。

```
import pandas as pd
import ma_utils as utils
logger = utils.FileLogger.get_logger()

class PreProcess():
def __init__(self, args):
"""
:param args: 算子参数，包含算子的业务参数以及默认参数obs_input_path（ 输入obs路径 ）和
obs_output_path(输出obs路径)
"""
pass

def __call__(self, input: pd.DataFrame) -> pd.DataFrame:
"""
:param input:输入参数
文本模态-JSONL/CSV文件：
框架提取JSONL/CSV的样本内容传递给算子，每个JSONL/CSV文件调用一次算子,DataFrame的字段名与
JSONL/CSV文件中的定义一致
文本模态-其它文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子。DataFrame的字段名：file_path（ 输入数据在算
子本地的完整路径
（ 算子框架会将数据集下载到算子本地目录中 ））和file_name（ 数据文件的文件名称 ）
图片/视频/音频-原始文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子，DataFrame的字段名：file_path（ 输入数据在算
子本地的完整路径
（ 算子框架会将数据集下载到算子本地目录中 ））和file_name（ 数据文件的文件名称 ）
```


版权所有 © 华为云计算技术有限公司

```
图片/视频/音频-Parquet文件：
框架提取Parquet的样本内容传递给算子，每个Parquet文件调用一次算子,DataFrame的字段名与Parquet文件中的定义一致，
其中有两个系统预定义字段:file_path（输入数据在算子本地的完整路径（算子框架会将数据集下载到算子本地目录中））
和file_name（数据文件的文件名称（为文件的相对路径））
:return:输出参数，算子输出数据样本列表DataFrame，由框架做后续处理
"""
pass


class Process():
def __init__(self, args):
"""
:param args: 算子参数，包含算子的业务参数以及默认参数obs_input_path（输入obs路径）和obs_output_path(输出obs路径)
"""
pass

def __call__(self, input: pd.DataFrame) -> pd.DataFrame:
"""
:param input:输入参数
文本模态-JSONL/CSV文件：
框架提取JSONL/CSV的样本内容传递给算子，每个JSONL/CSV文件调用一次算子,DataFrame的字段名与JSONL/CSV文件中的定义一致
文本模态-其它文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子。DataFrame的字段名：file_path（输入数据在算子本地的完整路径
（算子框架会将数据集下载到算子本地目录中））和file_name（数据文件的文件名称）
图片/视频/音频-原始文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子，DataFrame的字段名：file_path（输入数据在算子本地的完整路径
（算子框架会将数据集下载到算子本地目录中））和file_name（数据文件的文件名称）
图片/视频/音频-Parquet文件：
框架提取Parquet的样本内容传递给算子，每个Parquet文件调用一次算子,DataFrame的字段名与Parquet文件中的定义一致，
其中有两个系统预定义字段:file_path（输入数据在算子本地的完整路径（算子框架会将数据集下载到算子本地目录中））
和file_name（数据文件的文件名称（为文件的相对路径））
:return:输出参数，算子输出数据样本列表DataFrame，由框架做后续处理
"""
pass


class PostProcess():
def __init__(self, args):
"""
:param args: 算子参数，包含算子的业务参数以及默认参数obs_input_path（输入obs路径）和obs_output_path(输出obs路径)
"""
pass

def __call__(self, input: pd.DataFrame) -> pd.DataFrame:
"""
:param input:输入参数
文本模态-JSONL/CSV文件：
框架提取JSONL/CSV的样本内容传递给算子，每个JSONL/CSV文件调用一次算子,DataFrame的字段名与JSONL/CSV文件中的定义一致
文本模态-其它文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子。DataFrame的字段名：file_path（输入数据在算子本地的完整路径
（算子框架会将数据集下载到算子本地目录中））和file_name（数据文件的文件名称）
图片/视频/音频-原始文件：
框架获取数据集文件列表传递给算子，每个文件调用一次算子，DataFrame的字段名：file_path（输入数据在算子本地的完整路径
（算子框架会将数据集下载到算子本地目录中））和file_name（数据文件的文件名称）
图片/视频/音频-Parquet文件：
框架提取Parquet的样本内容传递给算子，每个Parquet文件调用一次算子,DataFrame的字段名与Parquet文件中的定义一致，
```

 版权所有 © 华为云计算技术有限公司

```
其中有两个系统预定义字段:file_path（输入数据在算子本地的完整路径（算子框架会将数据集下载到算子本地目录中））
和file_name（数据文件的文件名称（为文件的相对路径））
:return:输出参数，算子输出数据样本列表DataFrame，由框架做后续处理
"""
pass
```

模式二（auto-data-loading 为 false）：

适用于以下三类场景：

1）输入数据集为多模态数据集场景：例如图片+文本，视频+文本等由两种及以上模态组合的数据集；

2）输出数据集模态不在以下范围内：文本、图片、视频、音频；

3）需要将整个数据集作为输入的场景：例如去重类算子。

process.py开发规范如下所示，process.py文件包含三个类：

1.PreProcess：算子预处理相关逻辑（选填）

算子调用GPU/NPU做模型推理之前的、于CPU中计算的部分。拆分算子预处理逻辑有利于CPU计算与GPU/NPU计算的解耦，提高GPU/NPU利用率。

2.Process：算子推理相关逻辑 **（必填）**

建议只包含模型的加载及推理部分，将预处理和后处理编写在PreProcess及PostProcess中。

3.PostProcess: 算子后处理相关逻辑 （可选）

如算子在推理完成后，仍然有较重的、于CPU中计算的后处理逻辑，也建议将这部分逻辑拆分，编写在PostProcess中。

算子框架的调用顺序为preproces->process->postprocess。

```
import pandas as pd

class Process():
def __init__(self, args):
"""
:param args: 算子参数，包含算子的业务参数以及默认参数obs_input_path（输入obs路径）和obs_output_path(输出obs路径)
"""
pass

def __call__(self, input: pd.DataFrame):
"""
:param input 输入参数，在模式二的情况下，input为空的DataFrame
:return: 无返回值
"""
pass
```

算子基础镜像预置依赖清单

算子基础镜像预置的依赖包以及版本号如下表所示：

 版权所有 © 华为云计算技术有限公司

表 4-45 自定义基础镜像预置依赖清单

依赖包名	版本号
absl-py	2.1.0
accelerate	1.0.1
aclruntime	0.0.2
addict	2.4.0
aiohappyeyeballs	2.4.3
aiohttp	3.11.1
aiosignal	1.3.1
ais-bench	0.0.2
akg	2.2
albumentations	1.3.1
antlr4-python3-runtime	4.9.3
apex	0.1.dev20241029+ascend
ascend-faultdiag	6.0.0
ascend-training-accuracy-tools	1.0
ascendebug	0.1.0
astroid	3.0.3
asttokens	2.4.1
async-timeout	5.0.1
attrs	23.2.0
audioread	3.0.1
auto-tune	0.1.0
av	12.0.0
blinker	1.9.0
blobfile	3.0.0
blosc2	2.5.1
boto3	1.23.10
botocore	1.26.10
certifi	2024.12.14
cffi	1.16.0

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
charset-normalizer	3.4.1
click	8.1.7
click-aliases	1.0.5
cloudpickle	3.1.0
coloredlogs	15.0.1
contourpy	1.3.0
coverage	7.3.0
crc32c	2.7.1
cycler	0.12.1
Cython	3.0.2
dask	2024.2.1
dataflow	0.0.1
datasets	3.0.1
debugpy	1.8.8
decorator	4.4.2
decord	0.6.0
dill	0.3.8
easydict	1.12
einops	0.8.0
entrypoints	0.4
esdk-obs-python	3.23.12
et-xmlfile	2.0.0
exceptiongroup	1.2.2
executing	2.1.0
filelock	3.16.1
flask	2.3.3
flatbuffers	24.12.23
fonttools	4.55.0
frozenlist	1.5.0
fsspec	2024.6.1

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
fuzzywuzzy	0.18.0
gnureadline	8.2.10
gpytorch	1.12
greenlet	3.1.1
grpcio	1.60.0
grpcio-tools	1.60.0
gunicorn	21.2.0
h5py	3.9.0
hccl	0.1.0
hccl-parser	0.1
huaweicloud-sdk-python-modelarts-dataset	0.1.5
huggingface-hub	0.26.2
humanfriendly	10.0
idna	3.10
ijson	3.3.0
imageio	2.36.1
imageio-ffmpeg	0.5.1
importlib-metadata	8.5.0
importlib-resources	6.4.5
iniconfig	2.0.0
iopath	0.1.10
ipykernel	6.7.0
ipython	8.18.1
isort	5.13.2
itsdangerous	2.2.0
jaxtyping	0.2.19
jedi	0.19.2
jieba	0.42.1
jinja2	3.1.4

 版权所有 © 华为云计算技术有限公司

依赖包名	版本号
jmespath	1.0.1
joblib	1.4.2
jsonlines	4.0.0
jupyter-client	7.4.9
jupyter-core	5.7.2
keras	3.2.1
kiwisolver	1.4.5
lazy-import	0.2.2
lazy-loader	0.4
levenshtein	0.26.0
libclang	18.1.1
libcst	1.2.0
librosa	0.10.2.post1
linear-operator	0.5.3
llm-datadist	0.0.1
llvmlite	0.41.1
locket	1.0.0
lxml	5.1.0
Markdown	3.7
markdown-it-py	3.0.0
MarkupSafe	3.0.2
matplotlib	3.7.3
matplotlib-inline	0.1.7
mccabe	0.7.0
mdurl	0.1.2
mindspore-lite	2.4.0
mindstudio-probe	1.0.2
mindx-elastic	0.0.1
ml-dtypes	0.5.0
mmcv	2.0.1

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
mmengine	0.10.5
modelarts-pytorch-model-server	1.0.6
moviepy	2.1.2
moxing-framework	2.2.10
mpi4py	3.1.6
mpmath	1.3.0
msgpack	1.1.0
msit	7.0.0rc630
msit-benchmark	7.0.0rc2
msit-compare	7.0.0rc2
msit-llm	7.0.0rc2
msit-surgeon	7.0.0rc2
msprof-analyze	1.2.0
multidict	6.1.0
multiprocess	0.70.16
mypy-extensions	1.0.0
msobjdump	0.1.0
namex	0.0.8
ndindex	1.9.2
nest-asyncio	1.6.0
netifaces	0.11.0.post20240306102544
networkx	3.2.1
numba	0.58.1
numexpr	2.8.6
numpy	1.26.4
omegaconf	2.3.0
onnx	1.17.0
onnxconverter-common	1.14.0
onnxruntime	1.18.0
op-compile-tool	0.1.0

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
op-gen	0.1
op-test-frame	0.1
opc-tool	0.1.0
opencv-python	4.9.0.80
opencv-python-headless	4.8.1.78
openpyxl	3.1.5
optree	0.13.1
orjson	3.10.7
packaging	24.2
pandas	1.3.5
parso	0.8.4
partd	1.4.2
pathlib2	2.3.7.post1
peft	0.7.1
pexpect	4.9.0
pillow	10.4.0
pip	21.0.1
platformdirs	4.3.6
pluggy	1.5.0
ply	3.11
pooch	1.8.2
portalocker	2.10.1
prettytable	3.12.0
proglog	0.1.10
prometheus-client	0.14.1
prompt-toolkit	3.0.48
propcache	0.2.0
protobuf	3.20.2
psutil	6.0.0
ptyprocess	0.7.0

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
pure-eval	0.2.3
py-cpuinfo	9.0.0
pyarrow	18.0.0
pybind11	2.13.6
pyclipper	1.3.0.post6
pycocotools	2.0.7
pycparser	2.22
pycryptodome	3.21.0
pycryptodomex	3.21.0
pygments	2.18.0
pylint	3.0.2
pynndescent	0.5.13
pyparsing	3.2.0
pypng	0.20220715.0
pytest	7.4.3
python-dateutil	2.9.0.post0
python-dotenv	1.0.1
pytz	2024.2
PyWavelets	1.4.1
PyYAML	6.0.1
pyzmq	26.2.0
pydub	0.25.1
qudida	0.0.4
rapidfuzz	3.10.1
redis	5.1.1
regex	2024.9.11
requests	2.32.2
rich	13.9.4
rouge-metric	1.0.1
s3transfer	0.5.2

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
safetensors	0.4.5
schedule-search	0.0.1
scikit-image	0.22.0
scikit-learn	1.5.1
scipy	1.10.1
sentencepiece	0.2.0
setuptools	75.8.0
shapely	2.0.3
show-kernel-debug-data	0.1.0
shyaml	0.6.2
six	1.16.0
skl2onnx	1.17.0
soundfile	0.12.1
soxr	0.5.0.post1
SQLAlchemy	2.0.36
stack-data	0.6.3
sympy	1.13.0
setuptools	57.5.0
soundfile	0.13.1
tables	3.9.2
tabulate	0.9.0
tailor	0.3.4
te	0.4.0
tensorboard	2.18.0
tensorboard-data-server	0.7.2
termcolor	2.5.0
terminaltables	3.1.10
tfrecord	1.14.5
threadpoolctl	3.5.0
tifffile	2024.8.30

版权所有 © 华为云计算技术有限公司

依赖包名	版本号
tiktoken	0.7.0
timm	1.0.9
tokenizers	0.20.3
tomli	2.1.0
tomlkit	0.13.2
toolz	1.0.0
torch	2.1.0
torch-npu	2.1.0
torchvision	0.16.0
tornado	6.4.1
tqdm	4.66.3
traitlets	5.14.3
transformers	4.45.0
transformers-stream-generator	0.0.5
typeguard	4.4.1
typing-extensions	4.8.0
typing-inspect	0.9.0
tzdata	2024.2
tenacity	9.1.2
umap-learn	0.5.6
urllib3	1.26.7
wcwidth	0.2.13
werkzeug	3.0.3
wheel	0.45.1
XlsxWriter	3.2.0
xmltodict	0.13.0
xxhash	3.5.0
yacs	0.1.8
yapf	0.43.0
yarl	1.17.1

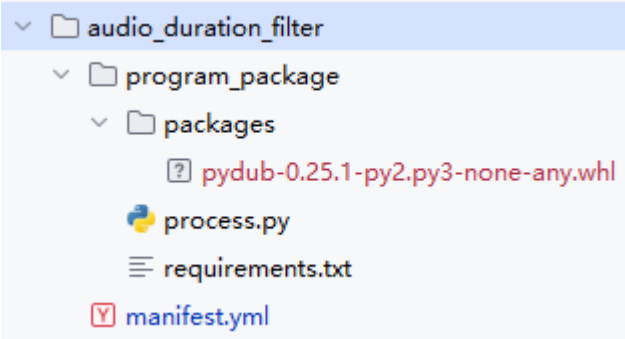
版权所有 © 华为云计算技术有限公司

依赖包名	版本号
ydata-profiling	4.16.1
zhdate	0.1
zipp	3.21.0
zstandard	0.22.0

4.5.10.3 算子开发典型样例

样例 1：音频时长过滤算子

算子包目录结构：



manifest.yml文件内容：

```
id: audio_duration_filter
name: 音频时长过滤
description: 保留音频时长在指定访问内的样本。
author: "Huawei Cloud/EI"

tags:
  language:
    - zh
  format:
    - wav
    - mp3
  category: 数据过滤
  modal:
    - AUDIO
runtime:
  cpu-arch:
    - ARM
  resources:
    - cpu: 4
      memory: 8192
  environment: PYTHON
  entrypoint: process.py
  auto-data-loading: true
arguments:
  - key: duration_range
    name: 保留时长（秒）
    type: int
    between: true
    min: 0
    tips: 保留时长在指定访问的的音频文件。
```


版权所有 © 华为云计算技术有限公司

```
  required: true
  visible: true
  default: 5;3600
```

process.py文件内容：

```
import os
import numpy as np
import pandas as pd
from pydub import AudioSegment
import argparse
import logging
logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(name)s - %(levelname)s - %(message)s')
logger = logging.getLogger(__name__)
def get_audio_duration(file_path):
    try:
        audio = AudioSegment.from_file(file_path)
        duration = len(audio) / 1000  # 转换为秒
    except:
        duration = None
    return duration


class Process:
    def __init__(self, args: argparse.Namespace):
        logger.info(f'self.args is {args},type is {type(args)}')
        min_duration, max_duration = args.duration_range.split(';')
        self.min_duration = int(min_duration) if min_duration else 0
        self.max_duration = int(max_duration) if max_duration else np.inf
    def __call__(self, df: pd.DataFrame) -> pd.DataFrame:
        duration_list = []
        filter_list = []
        for row in df.itertuples():
            duration = get_audio_duration(row.file_path)
            duration_list.append(duration)
            filter_list.append(0)
        df['duration'] = duration_list

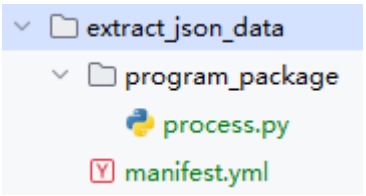
        # 根据duration筛选文件，0表示保留，1表示滤除。滤除动作由上层算子框架执行。
        df['filter'] = np.where((df['duration'] >= self.min_duration)
                                & (df['duration'] <= self.max_duration), 0, 1)
        return df
```

requirements.txt文件内容：

```
pydub
```

样例 2：提取文本中的 JSON 对象

算子包目录结构：



manifest.yml文件内容：

```
id: extract_json_data
name: 提取文本中的JSON对象
description: 提取文本中的JSON对象
author: "Huawei Cloud/EI"
version: 1.0.0
```


版权所有 © 华为云计算技术有限公司

```
tags:
  language:
    - zh
  format:
    - jsonl
    - txt
    - json
  category: 数据转换
  modal:
    - TEXT
    - OTHER
runtime:
  cpu-arch:
    - ARM
  resources:
    - cpu: 4
      memory: 8192
  environment: PYTHON
  entrypoint: process.py
  auto-data-loading: false
```

process.py文件内容：

```
import os
import pandas as pd
import argparse
import json
import logging
logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(name)s - %(levelname)s - %(message)s')
logger = logging.getLogger(__name__)
from moxing.file import list_directory, copy
def extract_json_data(input_file, output_file):
    """
    从输入文件中提取正文JSON数据，并写入输出文件（JSONL格式）
    :param input_file: 输入文件名
    :param output_file: 输出文件名
    """
    with open(input_file, 'r', encoding='utf-8') as fin, \
            open(output_file, 'w', encoding='utf-8') as fout:
        for line in fin:
            line = line.strip()
            if not line:
                continue

            # 使用分隔符" : "分割每行数据（最多分割一次）
            parts = line.split(' : ', 1)
            if len(parts) < 2:
                print(f"跳过无效行: {line}")
                continue

            json_str = parts[1].strip()
            try:
                # 尝试直接解析JSON
                json_data = json.loads(json_str)
            except json.JSONDecodeError:
                try:
                    # 替换常见中文引号后重试解析
                    fixed_str = json_str.replace('“', '"').replace('”', '"')
                    json_data = json.loads(fixed_str)
                except:
                    print(f"解析失败，跳过: {json_str}")
                    continue

            # 将有效的JSON对象写入输出文件（JSONL格式）
            fout.write(json.dumps(json_data, ensure_ascii=False) + '\n')
class Process:
    def __init__(self, args):
        # 建本地目录，用于存放从obs上下载的文件
        self.input_dir = r'input_dir'
```


版权所有 © 华为云计算技术有限公司

```
    os.makedirs(self.input_dir, exist_ok=True)
    # 建本地目录，用于存放算子生成的jsonl文件
    self.output_dir = r'output_dir'
    os.makedirs(self.output_dir, exist_ok=True)
    self.args = args
    logger.info(f'self.args is {self.args},type is {type(self.args)}')
def __call__(self, input):
    input_obs_path = self.args.obs_input_path
    output_obs_path = self.args.obs_output_path
    file_name_list = list_directory(input_obs_path, recursive=False)
    file_path_dict = {file_name: os.path.join(input_obs_path, file_name) for file_name in file_name_list}
    # 数据处理
    for file_name, obs_file_path in file_path_dict.items():
        input_file = os.path.join(self.input_dir, file_name)
        output_file = os.path.join(self.output_dir, file_name)
        copy(obs_file_path, input_file)
        extract_json_data(input_file, output_file)
        copy(output_file, os.path.join(output_obs_path, file_name))
```

4.5.10.4 管理自定义算子

创建自定义加工算子

查看自定义加工算子列表

查看自定义加工算子详情

更新自定义加工算子

删除自定义加工算子

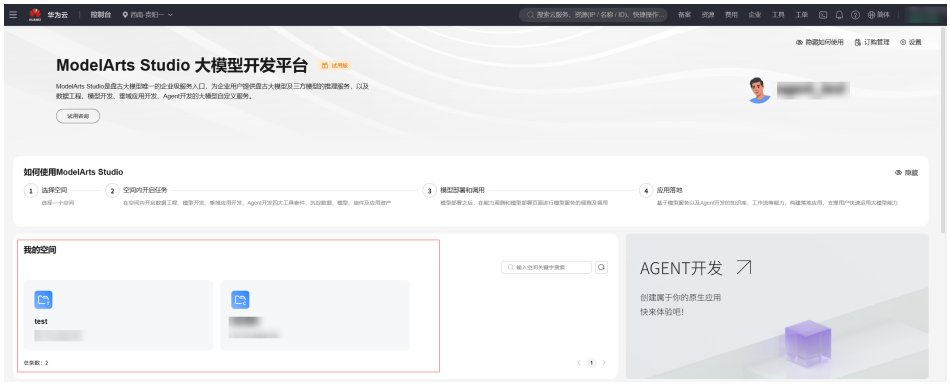
恢复自定义加工算子

彻底删除自定义加工算子

创建自定义加工算子

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-83 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”。
3. 在“加工算子管理 > 自定义”页面，单击右上角“创建自定义算子”。
4. 在“创建自定义算子”页面，单击“下载样例”可查看算子配置文件与算子包的规格说明，使用OBS服务上传算子配置文件与算子包，设置空间可见性后，单击右下角“确定”，完成自定义算子的创建。

版权所有 © 华为云计算技术有限公司

图 4-84 创建自定义算子

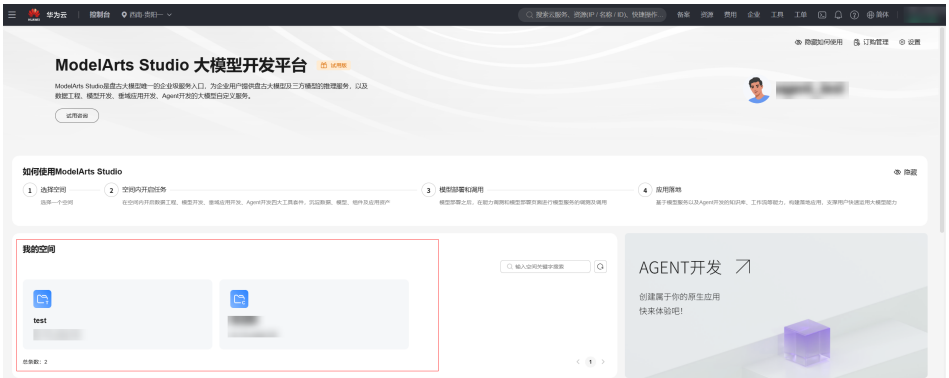


5. 创建成功的自定义算子可在**加工文本类数据集**、**加工图片类数据集**、**加工视频类数据集**、**加工音频类数据集**、**加工气象类数据集**、**加工预测类数据集**、**加工其他类数据集**时选择使用。

查看自定义加工算子列表

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

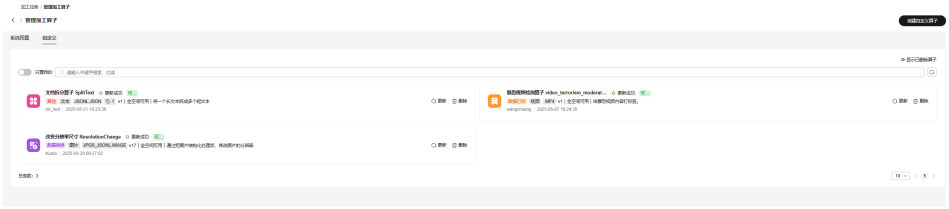
图 4-85 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面。

版权所有 © 华为云计算技术有限公司

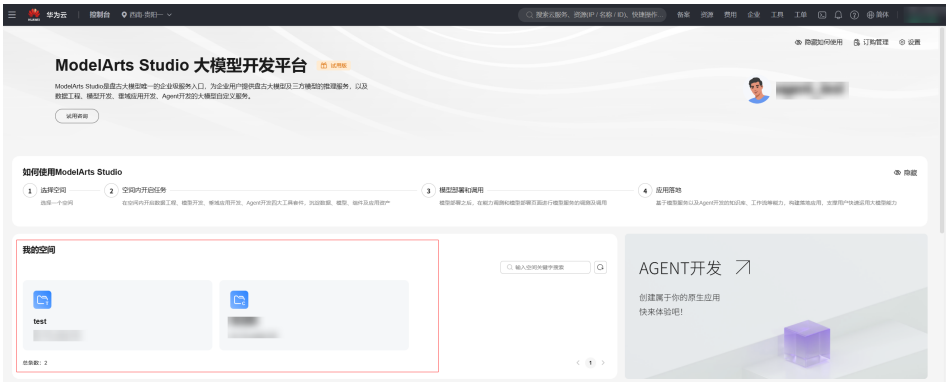
图 4-86 自定义算子列表



查看自定义加工算子详情

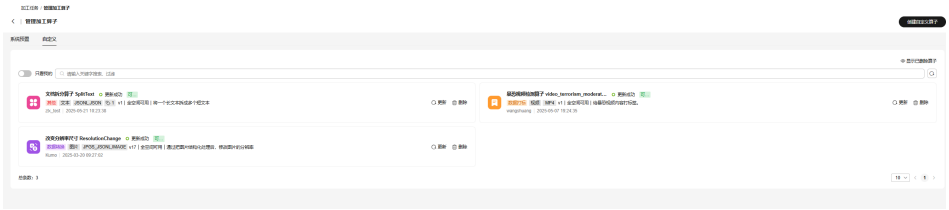
1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-87 我的空间



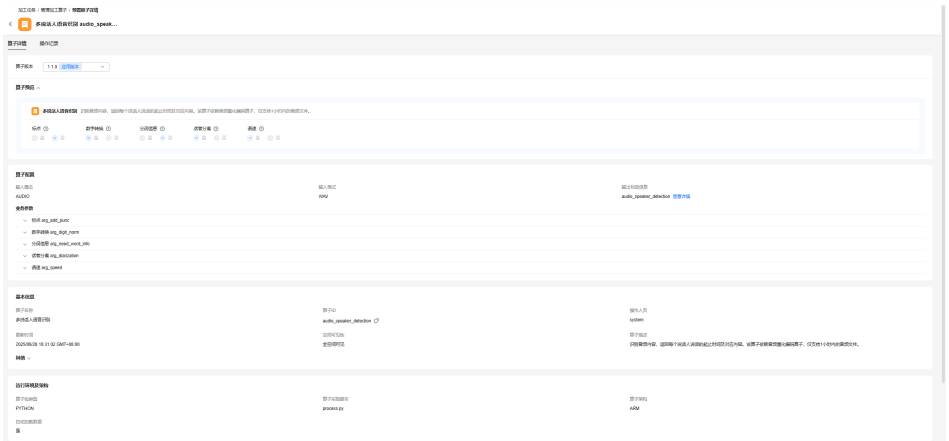
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面。

图 4-88 自定义算子列表



3. 单击算子“audio_speaker_detection”，查看算子详情。

图 4-89 算子详情

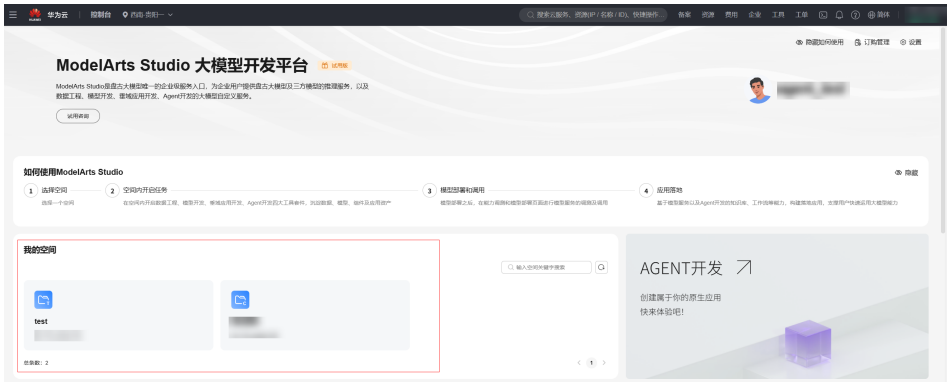


版权所有 © 华为云计算技术有限公司

更新自定义加工算子

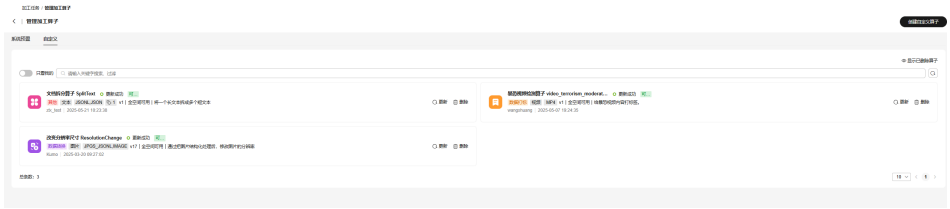
1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-90 我的空间



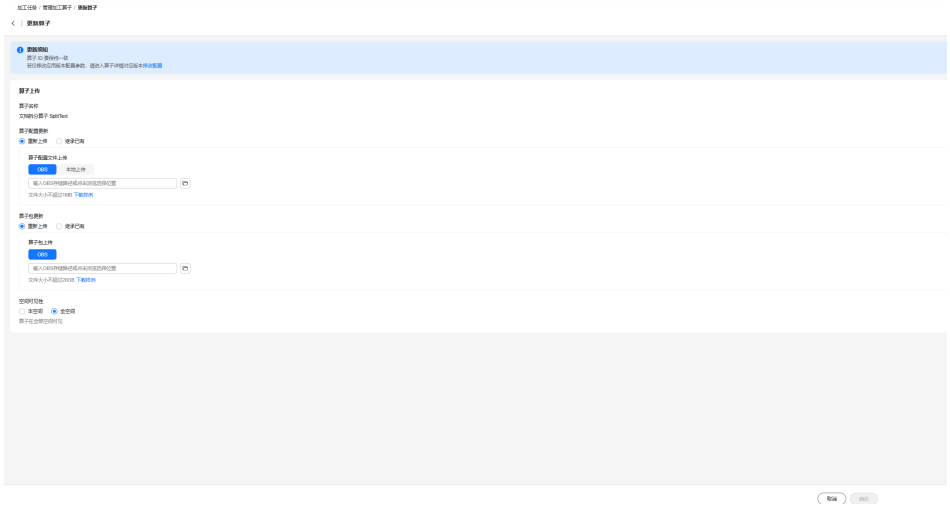
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面，单击自定义算子页签。

图 4-91 自定义算子列表



3. 单击算子”audio_speaker_detection”“更新”按钮。更新算子操作支持仅更新算子配置，仅更新算子包，同时更新算子和算子包。

图 4-92 更新自定义算子

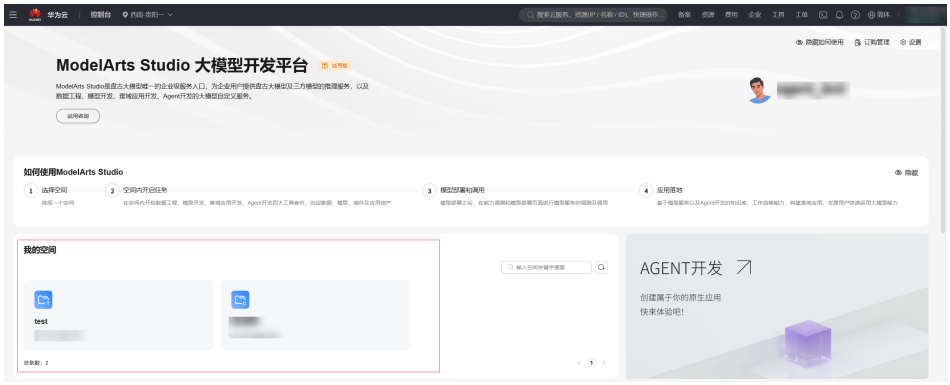


删除自定义加工算子

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

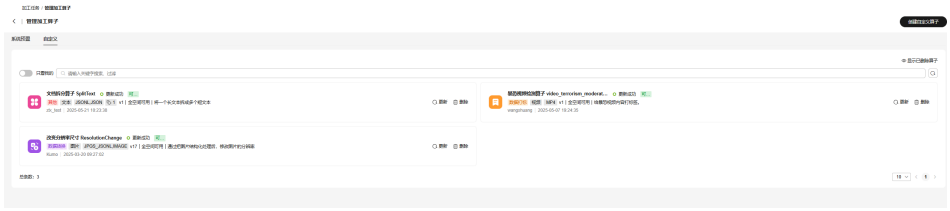
版权所有 © 华为云计算技术有限公司

图 4-93 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面，单击“自定义”算子页签。

图 4-94 自定义算子列表



3. 选择任意算子，单击“删除”，单击“确定”。

图 4-95 删除自定义算子

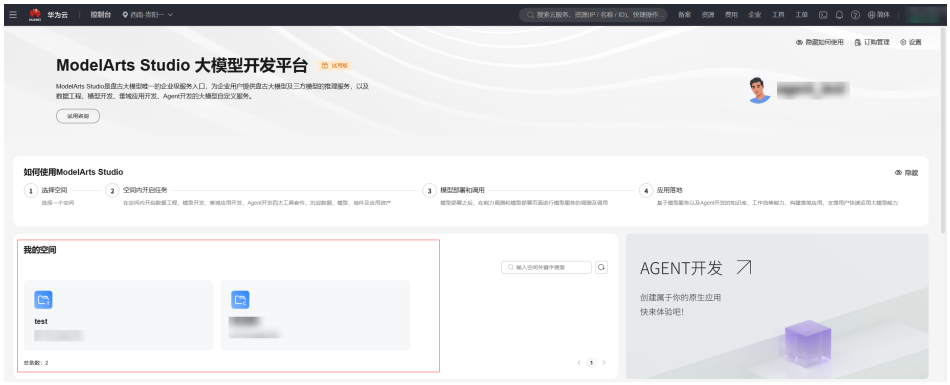


恢复自定义加工算子

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

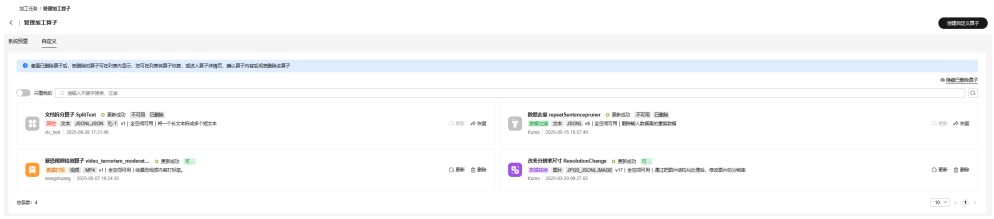
版权所有 © 华为云计算技术有限公司

图 4-96 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面，单击“自定义”算子页签，单击显示已删除算子。

图 4-97 自定义算子列表



3. 选择任意算子，单击“更新”。

图 4-98 恢复自定义算子



彻底删除自定义加工算子

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

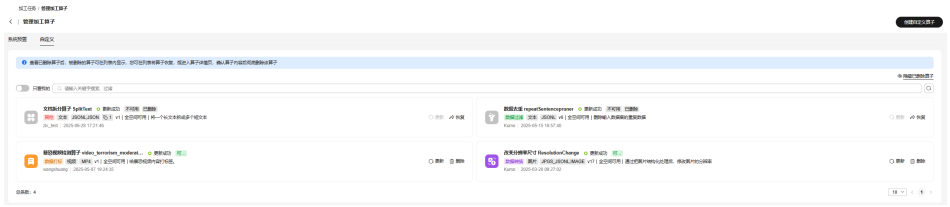
版权所有 © 华为云计算技术有限公司

图 4-99 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“管理加工算子”进入加工列表页面，单击“自定义”算子页签，单击显示已删除算子。

图 4-100 自定义算子列表



3. 选择任意已删除算子，进入详情页面，单击上方“彻底删除”，并做二次确认。

图 4-101 彻底删除自定义算子

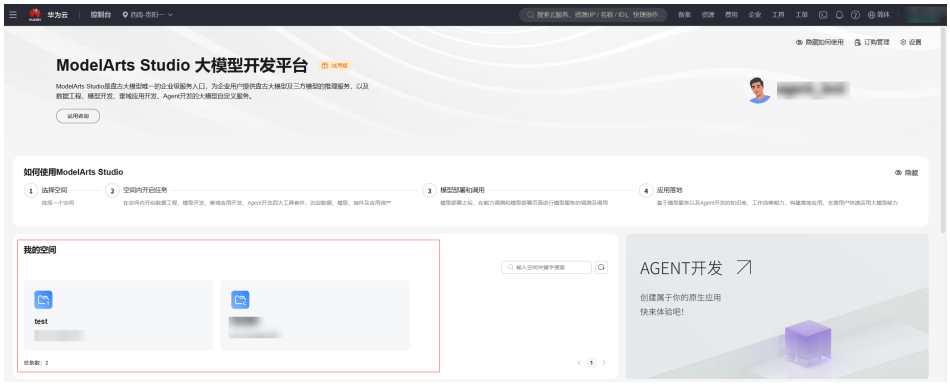


4.5.11 使用日志查看数据集加工进度

平台通过记录任务运行日志提供对加工数据集的查看加工进度操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-102 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”。
3. 单击任务ID查看加工任务详情页的日志管理页签下的运行日志页签。
   - 在“运行日志”页签可实时查看数据集加工的进度。
   - 可以找到每个任务开始结束的时间戳，可以算出任务时长。
   - 可以查看每个任务对应作业的运行状态和结果。

图 4-103 日志管理

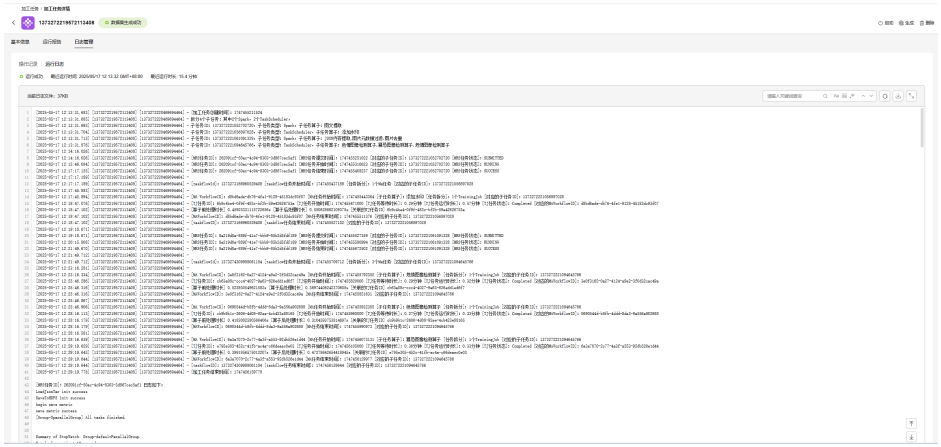
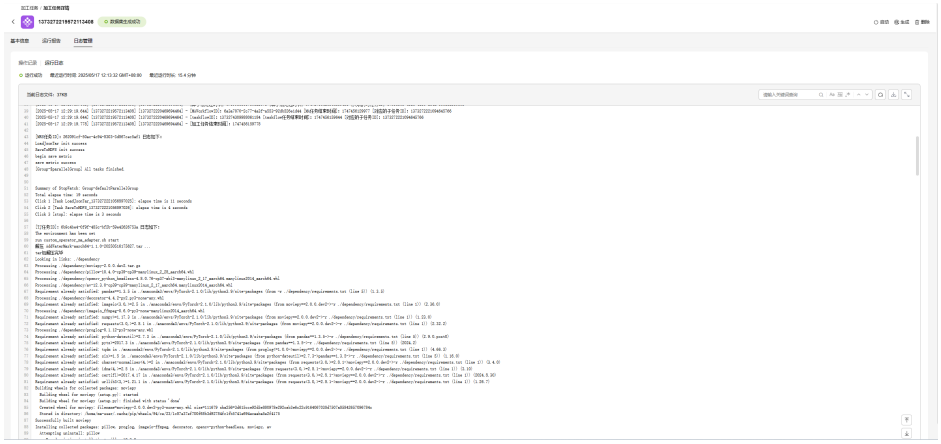


图 4-104 日志管理



运行日志内容介绍

数据集加工任务运行日志包括3个部分。具体如下。

- **数据集加工任务总览**
  加工任务创建后打印。显示该加工任务包含的算子。

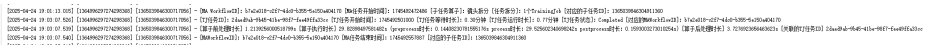
  图 4-105 数据集加工任务总览

  ```
  [2025-04-24 19:01:00.854][1364996297274298368][1365039846300717056] - [加工任务创建时间]: 1745492460572
  [2025-04-24 19:01:00.856][1364996297274298368][1365039846300717056] - 拆分3个子任务：其中2个MA，1个TaskScheduler，
  [2025-04-24 19:01:00.874][1364996297274298368][1365039846300717056] - 子任务ID：1365039846304911360，子任务类型：MA，子任务算子：镜头拆分
  [2025-04-24 19:01:00.904][1364996297274298368][1365039846300717056] - 子任务ID：1365039846715953152，子任务类型：TaskScheduler，子任务算子：运动幅度评分,添加水印
  [2025-04-24 19:01:00.926][1364996297274298368][1365039846300717056] - 子任务ID：1365039846724341760，子任务类型：MA，子任务算子：视频裁剪
  ```

- **算子运行详情日志**
  记录算子在运行时关键节点的信息。

版权所有 © 华为云计算技术有限公司

图 4-106 算子运行详情日志



- **MRS/MA任务运行日志**
  记录算子内部MRS/MA任务运行的详细信息。

图 4-107 MRS/MA 任务运行日志

```
25
26  [TJ任务ID]: 2daed9ab-9b45-41be-98f7-fee49ffa33cc 日志如下:
27  [GCC 10.3.1]
28  INFO:root:Using MoXing-v2.1.6.879ab2f4-879ab2f4
29  INFO:root:Using OBS-Python-SDK-3.20.9.1
30  INFO:__main__:The start time of the framework:1745492505.1990206
31  INFO:__main__:Downloading input data from obs...
32  INFO:__main__:The start time of downloading input data from obs is 1745492505.362652
33  INFO:root:List OBS time cost: 0.04 seconds.
34  ['-movflags', ' faststart']
35  INFO:root:Copy parallel total time cost: 0.25 seconds.
36  INFO:__main__:The end time of downloading input data from obs is 1745492505.7128587
37  INFO:__main__:The start time of converting to DataFrame is 1745492505.803619
38
39  0it [00:00, ?it/s]
40  3it [00:00, 142.17it/s]
41  INFO:__main__:Preparing input annotations..
42  INFO:__main__:The end time of converting to DataFrame is 1745492506.1408036
43  INFO:__main__:Initializing the operator...
44  INFO:__main__:The start time of initializing the operator is 1745492506.3248634
45  INFO:__main__:The end time of initializing the operator is 1745492506.4129462
46  INFO:OperatorLogger:Pre-operator finished,total time: 1.2139256000518799s
47  INFO:__main__:Running the operator...
48  INFO:__main__:The start time of running the operator is 1745492506.6356113
49
50    0%|          | 0/3 [00:00<?, ?it/s]INFO:OperatorLogger:The start time of preprocessing0 is 1745492506.7272844
51  INFO:OperatorLogger:The end time of preprocessing0 is 1745492506.781274
52  INFO:OperatorLogger:The start time of processing0 is 1745492506.8247027
53
54
```

4.5.12 管理加工后的数据集

完成数据加工、数据合成、数据标注或数据配比任务的数据集，在对应任务列表执行“生成”操作，将生成“加工数据集”被平台统一管理，并用于后续的发布任务。

平台支持对加工数据集查看基本信息、数据血缘等管理操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-108 我的空间



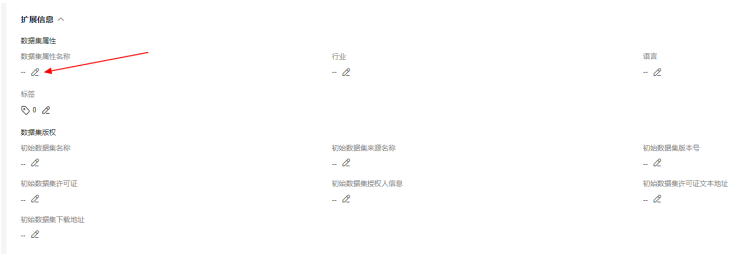
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据集 > 加工数据集”。
3. 数据集支持“数据集名称”、“数据集模态”、“数据集属性名称”等属性类型进行搜索、过滤。

版权所有 © 华为云计算技术有限公司



4. 单击数据集名称查看加工数据集的基本信息、数据预览、数据血缘以及操作记录。
   - 在“基本信息”页签，可以查看数据详情、数据来源以及扩展信息。在扩展信息内，可以根据用户需求设置数据集属性，包括数据集属性名称、行业、语言、自定义标签，如图4-109。

     图 4-109 编辑扩展信息

   - 在“数据预览”页签可查看加工后的数据内容。
   - 在“数据血缘”页签查看该数据集所经历的操作，如导入、合成等操作。数据血缘上钻至OBS源文件，下钻至模型训练。
   - 在“操作记录”页签可以查看数据集所经历的操作及状态等信息。
5. 单击操作列的“删除”，可删除不需要的数据集。
   - 如果需要恢复删除的数据集，可单击右上角“显示已删除数据”，被删除的数据集将在列表显示，可将数据集恢复。
   - 如果需要彻底删除数据集，可单击数据集名称进入详情页，确认数据集内容后彻底删除该数据集。

   说明

   删除“加工数据集”属于高危操作，删除前，请确保该数据集不再使用。

版权所有 © 华为云计算技术有限公司

4.5.13 管理加工任务资源

与MRS相关的加工任务、发布任务在创建时，可以进行任务资源的参数配置，具体的参数可见**表1 任务资源参数清单**

表 4-46 任务资源参数清单

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

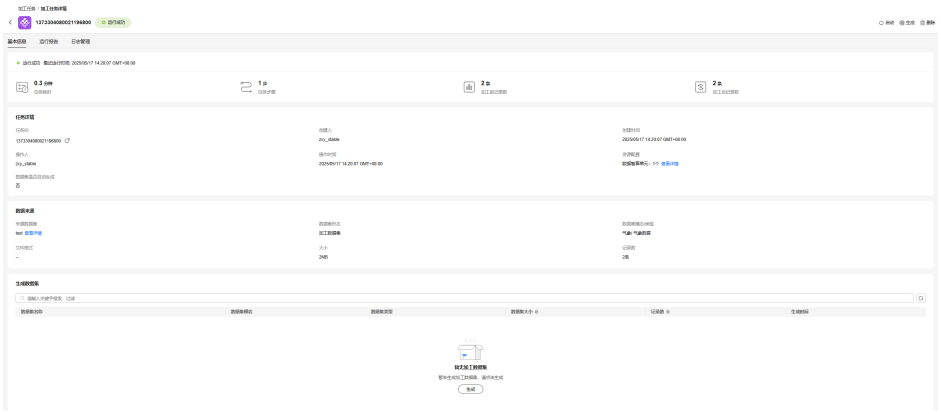
4.5.14 加工任务生成数据集

加工任务详情页面增加生成数据集相关提示。

1. 当加工任务执行结束且没有生成数据集时，页面提供生成数据集按钮。

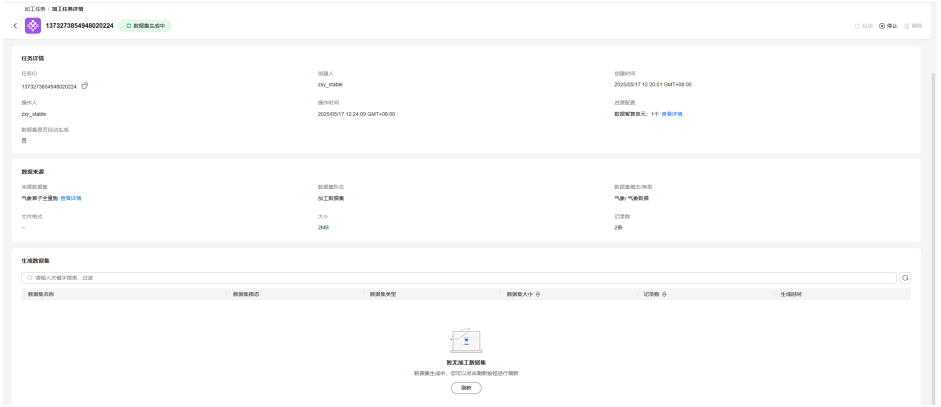
版权所有 © 华为云计算技术有限公司

图 4-110 生成数据集按钮



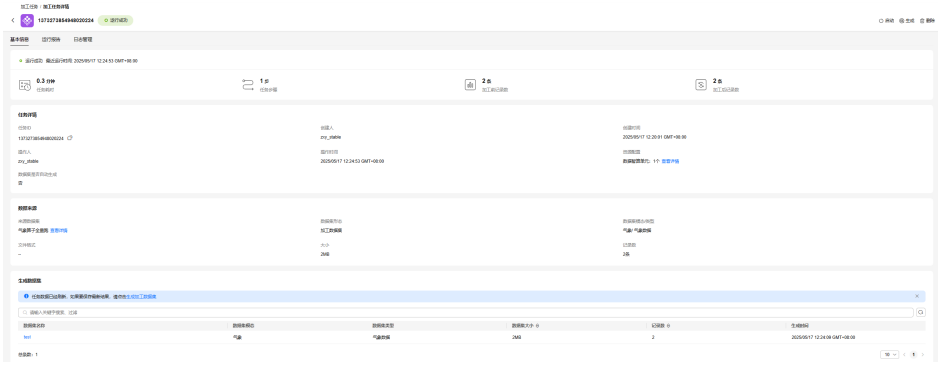
2. 当任务无生成数据集且处理生成数据集中状态时，提供刷新按钮。

图 4-111 刷新按钮



3. 当加工任务最新一次执行任务成功，且没有生成最新数据集时，页面提示：任务数据已经刷新，如果要保存最新结果，请单击生成加工数据集。

图 4-112 任务数据已经刷新，如果要保存最新结果，请单击生成加工数据集



4.6 发布数据集

版权所有 © 华为云计算技术有限公司

4.6.1 数据集发布场景介绍

数据发布介绍

ModelArts Studio大模型开发平台提供的数据发布功能涵盖数据评估和数据发布操作，旨在通过数据质量评估，确保数据满足大模型训练的多样性、平衡性和代表性需求，促进数据的高效流通和应用。

数据发布不仅包括将数据发布为适合使用的格式，还要求根据任务需求评估数据集效果，确保数据集在规模、质量和内容上符合模型训练的标准。

- **数据评估**

  平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。

- **数据发布**

  数据发布是将数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。支持的发布格式为标准格式、盘古格式（适用于训练盘古大模型时）。目前，仅文本类和图片类数据集支持发布为“盘古格式”。

通过这些功能，平台能够帮助用户科学管理和发布数据集，确保数据集质量符合大模型训练的需求，从而提高后续模型训练的效果。

数据发布意义

数据发布不仅仅是将数据转换为不同格式，还包括根据任务需求评估数据集效果，确保数据在规模、质量和内容上满足训练标准。具体而言，数据发布具备以下几个重要意义：

- **多格式支持**

  对于文本类、图片类数据集，平台支持多种数据发布格式，包括“标准格式”、“盘古格式”，以满足不同训练任务的需求。通过这些格式的转换，用户可以确保数据与特定模型（如盘古大模型）兼容，并优化训练效果。

- **提高训练效率**

  发布符合标准的数据集可以大幅提升数据处理效率，减少后续调整工作，帮助用户快速进入模型训练阶段。

数据集发布是数据工程中的关键环节，确保数据集符合模型训练要求。通过平台提供的数据发布功能，用户能够根据具体任务需求，灵活选择数据发布格式，保证数据的兼容性与一致性，从而为后续模型训练和应用部署打下坚实基础。

支持数据发布的数据集类型

支持数据发布的数据集类型见表4-47。

表 4-47 支持数据发布的数据集类型

数据类型	数据评估	数据发布
文本类	√	√
图片类	√	√

版权所有 © 华为云计算技术有限公司

数据类型	数据评估	数据发布
视频类	√	√
音频类	-	√
气象类	-	√
预测类	-	√
其他类	-	√

ModelArts Studio大模型开发平台支持将**文本类、图片类数据集**发布为两种格式：

- **标准格式**：适用于广泛的数据使用场景，满足大多数模型训练的标准需求。该格式的数据集将发布到资产中，但下游模型开发不可见。
- **盘古格式**：专为盘古大模型训练设计的格式，确保数据集在盘古模型训练中的兼容性和一致性。该格式的数据集将被用于ModelArts Studio大模型开发平台的模型开发功能使用。

说明

除**文本类、图片类**数据集外，其余类型的数据集当前仅支持发布为**标准格式**。

4.6.2 发布文本类数据集

4.6.2.1 评估文本类数据集

发布文本类数据集前，ModelArts Studio大模型开发平台支持对数据集进行评估操作，帮助用户优化数据质量，确保数据满足高标准，提升模型性能。

如果无需使用数据评估操作，**可跳过此章节至发布文本类数据集。**

创建文本类数据集评估标准

ModelArts Studio大模型开发平台针对文本类数据集预设了一套基础评估标准，涵盖了数据准确性、完整性、一致性、格式规范等多个维度，用户可以直接使用该标准或在该标准的基础上创建评估标准。

若您希望使用平台预置的评估标准，**可跳过此章节至创建文本类数据集评估任务**。

创建文本类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-113 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，在“人工评估标准”页签，平台预置了文本类数据集评估标准“NLP数据质量标准V1.0”，单击评估标准名称，可以查看具体的评估项。

图 4-114 预置文本类数据集评估标准



3. 在“人工评估标准”页面，单击“创建自定义标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 编辑评估项。

   用户可以基于实际需求删减评估项，或创建自定义评估项。创建自定义评估项时，需要将评估类别、评估项、评估项说明填写清晰，填写时确保描述无歧义。
5. 单击“完成创建”以创建评估标准。

   评估标准创建完成后可以在“人工评估标准”页面查看创建的评估标准，并支持编辑与删除操作。

创建文本类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

创建文本类数据集评估任务前，请参考**加工文本类数据集**，生成一个“加工数据集”。

创建文本类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-115 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，单击界面右上角“创建评估任务”。
3. 选择需要评估的加工数据集，并设置抽样样本的数量。
4. 单击“下一步”，选择评估标准。单击“下一步”设置评估人员，单击“下一步”填写任务名称。
5. 单击“完成创建”，将返回至“数据评估”页面，评估任务创建成功后状态将显示为“已创建”。
6. 单击操作列的“评估”，进入评估页面。
7. 在评估页面，可参考评估项对当前数据的问题进行标注，且满足则单击“通过”，不满足则单击“不通过”。

   如图4-116，对于文本类数据集而言，可选中问题内容后，右键标记数据问题。

图 4-116 标记数据集问题



8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。

   单击操作列“报告”，可查看数据集质量评估报告。

4.6.2.2 发布文本类数据集

数据发布是将数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

文本类数据集支持发布的格式为：

- **标准格式：**数据工程功能支持的原始格式。

  标准格式的示例如下，其中，**context**和**target**是键值对。

  ```
  {"context": "你好，请介绍自己", "target": "我是盘古大模型"}
  ```

- **盘古格式：**训练盘古大模型时，需要将数据集格式发布为“盘古格式”。为适配不同盘古模型的数据规范，文本类型-单轮问答、单轮问答（带人设）、多轮问

版权所有 © 华为云计算技术有限公司

答、多轮问答（带人设）、DPO、DPO（带人设）发布时需选择盘古格式-非思维链或盘古格式-思维链进行发布。

盘古格式-非思维链的示例如下，其中，**context**和**target**是键值对。与标准格式不同，**context**是一个数组。

```
{"context":["你好，请介绍自己"],"target":"我是盘古大模型"}
```

盘古格式-思维链模式的示例如下，其中，**context,target**和**thinking**是键值对。与盘古-非思维链格式相比，多了一个**thinking**键值对，**thinking**的值是一个数组。

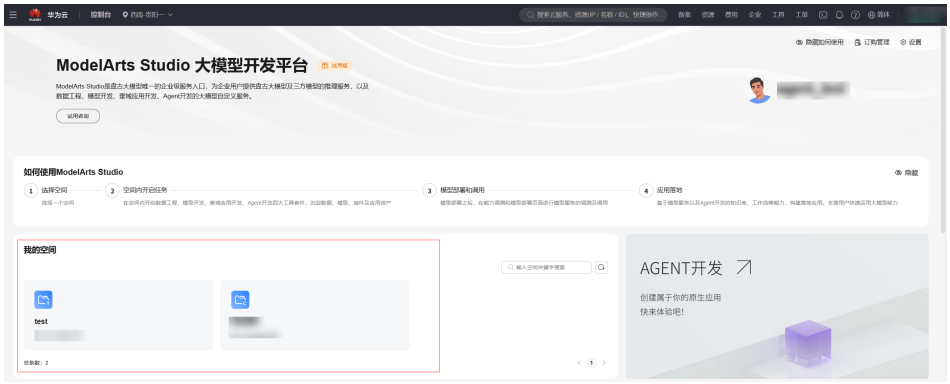
```
{"context":["你好，请介绍自己"],"target":"我是盘古大模型","thinking": ["好的，用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景"]}
```

创建文本类数据集发布任务

创建文本类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-117 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“文本 > 单轮问答”类型的数据集。

图 4-118 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、使用场景和格式配置，如图4-119。

由于数据工程需要支持对接盘古大模型或三方大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

当前支持标准格式、盘古格式：

- **标准格式**：数据工程功能支持的原始格式。该格式的数据集可发布到资产中，但下游模型开发不可见。“使用场景”选择“其他”时使用。

版权所有 © 华为云计算技术有限公司

- **盘古格式**：专门针对训练模型为盘古大模型而定义的数据集格式。该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。为适配不同盘古模型的数据规范，在数据集格式配置时，需要选择“选择盘古格式-非思维链”或“盘古格式-思维链进行发布”。“使用场景”为“STUDIO模型训练”时使用。

图 4-119 格式配置



说明

- 当使用发布的数据集训练**盘古**大模型时，请选择“格式配置”为**盘古格式。**
- 当发布文本数据集为如下类型时，请选择“格式配置”为“盘古格式-非思维链”或“盘古格式-思维链”。
  - 文本类型-单轮问答
  - 单轮问答（带人设）
  - 多轮问答
  - 多轮问答（带人设）
  - DPO
  - DPO（带人设）
- 当发布文本数据集为“预训练”类型时，请选择“格式配置”为“盘古格式”。
- 当发布成“盘古格式-非思维链”时会把数据集中符合“盘古格式-思维链”的数据剔除，反之发布“盘古格式-思维链”则会把“盘古格式-非思维链”的数据剔除。

6. 填写数据集名称、描述，设置扩展信息后，单击“确认”，发布数据集。

   扩展信息会自动保留所选数据集的属性，并且在选择发布“盘古格式-思维链”数据集之后，后端会自动给扩展信息的标签（simple_tag）加上“reasoner”标签，模型会根据该标签对数据集进行过滤。

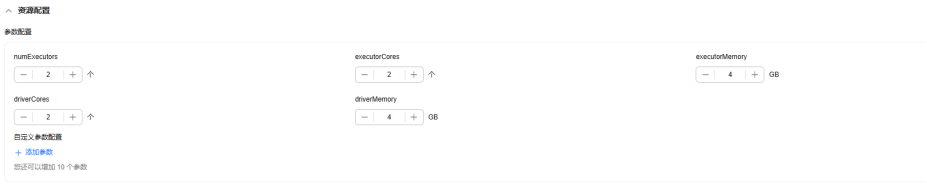
版权所有 © 华为云计算技术有限公司

图 4-120 生成发布数据集



7. （可选）如果发布数据集为“预训练”类型时，在完成扩展信息设置之后需要单击“下一步”进行资源设置后才能单击“确认”进行发布。

图 4-121 资源配置



参数说明见表4-48：

表 4-48 参数配置

参数名称	参数说明
numExecutors	Executor的数量，默认值2。Executor是运行在工作节点上的进程，它负责执行具体的task任务，并将计算结果返回给Driver。每个Executor每个核同时能跑一个task任务，所以增加了Executor的个数相当于增大了任务的并发度。在资源充足的情况下，可以相应增加Executor的个数，以提高运行效率。 numExecutors * executorMemory最小值为4，最大值为16。
executorCores	每个Executor进程使用的CPU内核数量，默认值2。每个Executor多个核同时能跑多个task任务，相当于增大了任务的并发度。但是由于所有核共用Executor的内存，所以要在内存和核数之间做好平衡。 numExecutors * executorMemory最小值为4，最大值为16。executorCores和executorMemory的比例需要在1:2~1:4之间。
executorMemory	每个Executor进程使用的内存数量，默认值4。Executor的内存主要用于任务执行、通信等。当一个任务很大的时候，可能需要较多资源，因而内存也可以做相应的增加；当一个任务较小运行较快时，就可以增大并发度减少内存。 executorCores和executorMemory的比例需要在1:2~1:4之间。

 版权所有 © 华为云计算技术有限公司

参数名称	参数说明
driverCores	驱动程序进程使用的CPU内核数量，默认值2。Driver负责任务的调度，和Executor之间的消息通信。 driverCores和driverMemory的比例需要在1:2~1:4之间。
driverMemory	驱动程序进程使用的内存数量，默认值4。Driver负责任务的调度，和Executor之间的消息通信。当任务数变多，任务平行度增大时，Driver内存都需要相应增大。 driverCores和driverMemory的比例需要在1:2~1:4之间。

8. 当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

4.6.3 发布图片类数据集

4.6.3.1 评估图片类数据集

发布图片类数据集前，ModelArts Studio大模型开发平台支持对数据集进行评估操作，帮助用户优化数据质量，确保数据满足高标准，提升模型性能。

如果无需使用数据评估操作，**可跳过此章节至发布图片类数据集。**

创建图片类数据集评估标准

ModelArts Studio大模型开发平台针对图片类数据集预设了一套基础评估标准，涵盖了图像清晰度、分辨率、标签准确性、图像一致性等多个质量维度，用户可以直接使用该标准或在该标准的基础上创建评估标准。

若您希望使用平台预置的评估标准，**可跳过此章节至创建图片类数据集评估任务。**

创建图片类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-122 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，在“人工评估标准”页签，平台预置了图片类数据集评估标准“图片数据质量标准V1.0”，单击评估标准名称，可以查看具体的评估项。

图 4-123 预置图片类数据集评估标准



3. 在“人工评估标准”页面，单击“创建自定义标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 编辑评估项。

   用户可以基于实际需求删减评估项，或创建自定义评估项。创建自定义评估项时，需要将评估类别、评估项、评估项说明填写清晰，填写时确保描述无歧义。
5. 单击“完成创建”以创建评估标准。

   评估标准创建完成后可以在“人工评估标准”页面查看创建的评估标准，并支持编辑与删除操作。

创建图片类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

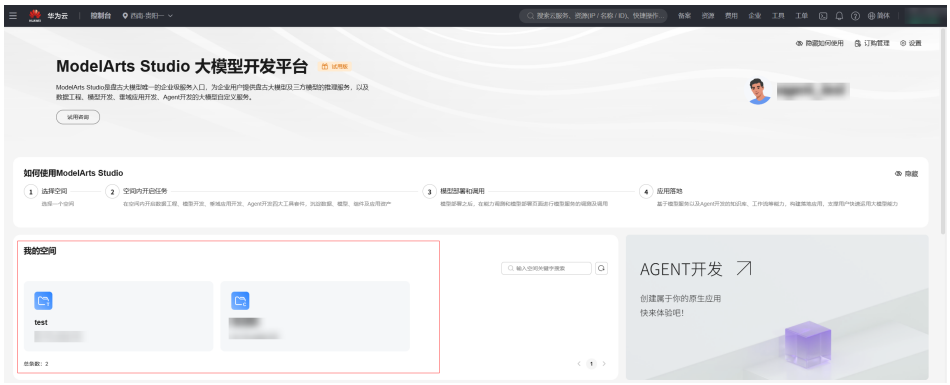
创建图片类数据集评估任务前，请参考**加工图片类数据集**，生成一个“加工数据集”。

创建图片类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-124 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，单击界面右上角“创建评估任务”。
3. 选择需要评估的加工数据集，并设置抽样样本的数量。
4. 单击“下一步”，选择评估标准。单击“下一步”设置评估人员，单击“下一步”填写任务名称。
5. 单击“完成创建”，将返回至“数据评估”页面，评估任务创建成功后状态将显示为“已创建”。
6. 单击操作列的“评估”，进入评估页面。
7. 在评估页面，可参考评估项对当前数据的问题进行标注，且满足则单击“通过”，不满足则单击“不通过”。
8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。

   单击操作列“报告”，可查看数据集质量评估报告。

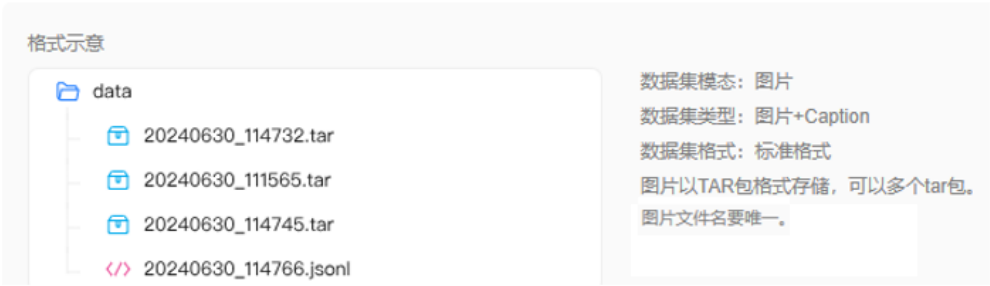
4.6.3.2 发布图片类数据集

数据发布是将数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

图片类数据集支持发布的格式为：

- **标准格式：**如图4-125，平台默认的格式。该格式的数据集可发布到资产中，但下游模型开发不可见。

图 4-125 图片类数据集标准格式示例



- **盘古格式：**如图4-126，训练盘古大模型时，需要将数据集格式发布为“盘古格式”，该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。

版权所有 © 华为云计算技术有限公司

图 4-126 图片类数据集盘古格式示例

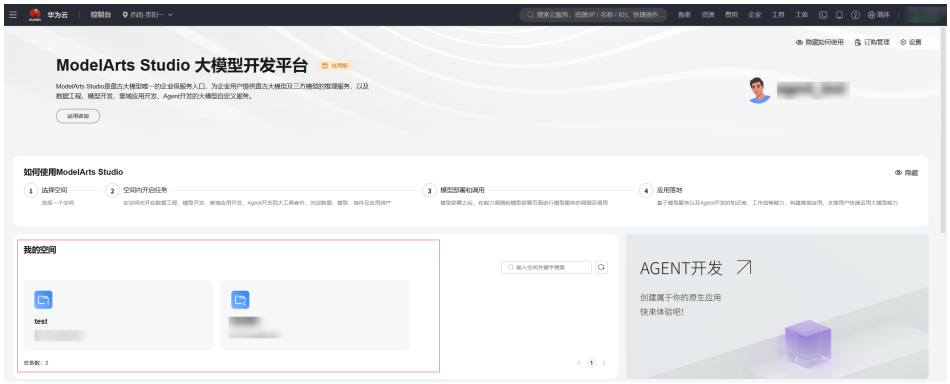


创建图片类数据集发布任务

创建图片类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-127 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“图片 > 图片+Caption”类型的数据集。

图 4-128 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。

   由于数据工程需要支持对接盘古大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

   当前支持标准格式、盘古格式：

   - **标准格式**：数据工程功能支持的原始格式。该格式的数据集可发布到资产中，但下游模型开发不可见。

版权所有 © 华为云计算技术有限公司

- **盘古格式**：使用盘古大模型训练时所需要使用的数据格式，该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。

说明

如果使用该数据集训练盘古大模型，请将发布格式配置为**盘古格式**。

6. 根据数据集是否进行拆分，勾选“数据集拆分”。如勾选，设置好训练集和验证集的比例，如图4-129所示：

图 4-129 数据集拆分示例



7. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

8. 单击任务id，可以查看发布任务详情。发布任务详情界面包含“基本信息”和“日志管理”2个tab页，其中基本信息又包含“任务详情”、“配置信息”、“数据来源”、“生成数据集”几个部分。在“生成数据集”表格的记录数列中，展示训练集和验证集各自的数量，如图4-131所示。

图 4-130 查看任务详情



图 4-131 生成数据集表格示例



4.6.4 发布视频类数据集

4.6.4.1 评估视频类数据集

发布视频类数据集前，ModelArts Studio大模型开发平台支持对数据集进行评估操作，帮助用户优化数据质量，确保数据满足高标准，提升模型性能。

如果无需使用数据评估操作，**可跳过此章节至发布视频类数据集。**

创建视频类数据集评估标准

ModelArts Studio大模型开发平台针对视频类数据集预设了一套基础评估标准，涵盖了视频的清晰度、帧率、完整性、标签准确性等多个质量维度，用户可以直接使用该标准或在该标准的基础上创建评估标准。

版权所有 © 华为云计算技术有限公司

若您希望使用平台预置的评估标准，**可跳过此章节至创建视频类数据集评估任务**。

创建视频类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-132 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，在“人工评估标准”页签，平台预置了视频类数据集评估标准“视频数据质量标准V1.0”，单击评估标准名称，可以查看具体的评估项。

图 4-133 预置视频类数据集评估标准



3. 在“人工评估标准”页面，单击“创建自定义标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 编辑评估项。

   用户可以基于实际需求删减评估项，或创建自定义评估项。创建自定义评估项时，需要将评估类别、评估项、评估项说明填写清晰，填写时确保描述无歧义。
5. 单击“完成创建”以创建评估标准。

   评估标准创建完成后可以在“人工评估标准”页面查看创建的评估标准，并支持编辑与删除操作。

创建视频类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

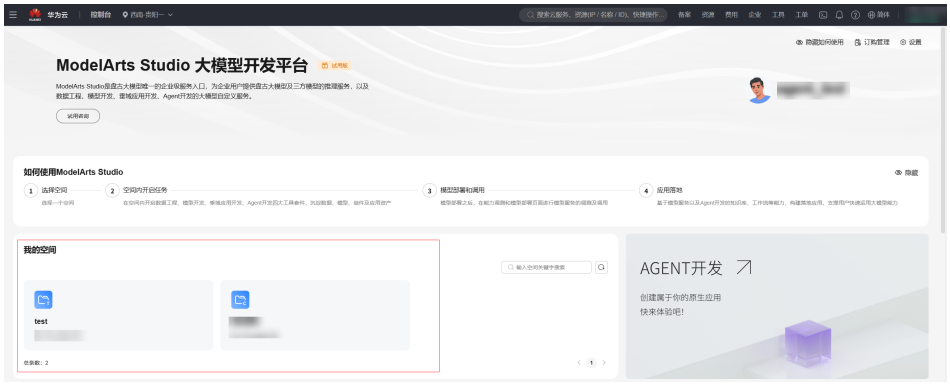
版权所有 © 华为云计算技术有限公司

创建视频类数据集评估任务前，请参考**加工视频类数据集**，生成一个“加工数据集”。

创建视频类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 4-134 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，单击界面右上角“创建评估任务”。
3. 选择需要评估的加工数据集，并设置抽样样本的数量。
4. 单击“下一步”，选择评估标准。单击“下一步”设置评估人员，单击“下一步”填写任务名称。
5. 单击“完成创建”，将返回至“数据评估”页面，评估任务创建成功后状态将显示为“已创建”。
6. 单击操作列的“评估”，进入评估页面。
7. 在评估页面，可参考评估项对当前数据的问题进行标注，且满足则单击“通过”，不满足则单击“不通过”。
8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。

   单击操作列“报告”，可查看数据集质量评估报告。

4.6.4.2 发布视频类数据集

数据发布是将数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

视频类数据集当前仅支持发布为“标准格式”。

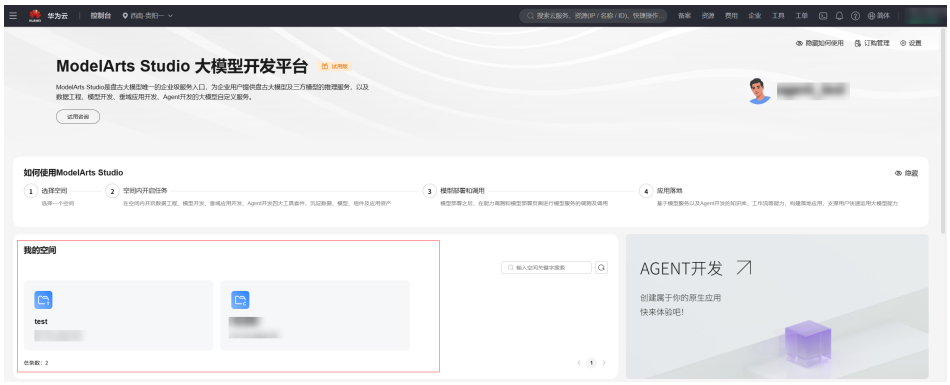
创建视频类数据集发布任务

创建视频类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-135 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“视频 > 仅视频”类型的数据集。

图 4-136 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前视频类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

4.6.5 发布音频类数据集

数据发布是将数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

音频类数据集当前仅支持发布为“标准格式”。

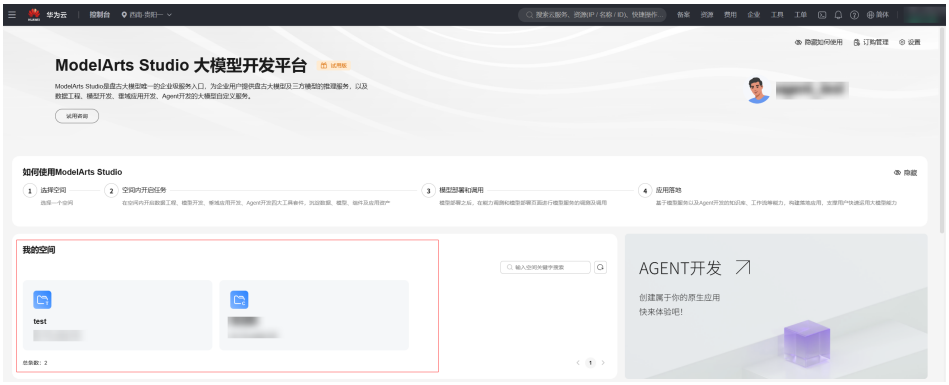
创建音频类数据集发布任务

创建视频类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-137 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“音频 > 音频”类型的数据集。

图 4-138 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前音频类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

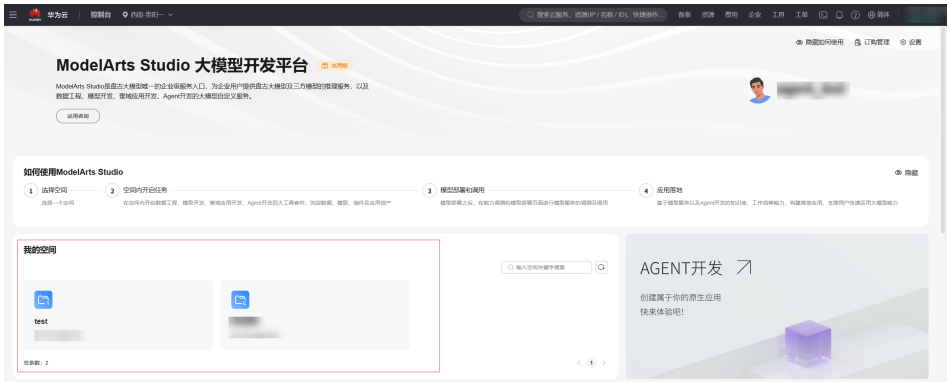
4.6.6 发布气象类数据集

操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-139 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“气象 > 气象数据”类型的数据集。

图 4-140 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

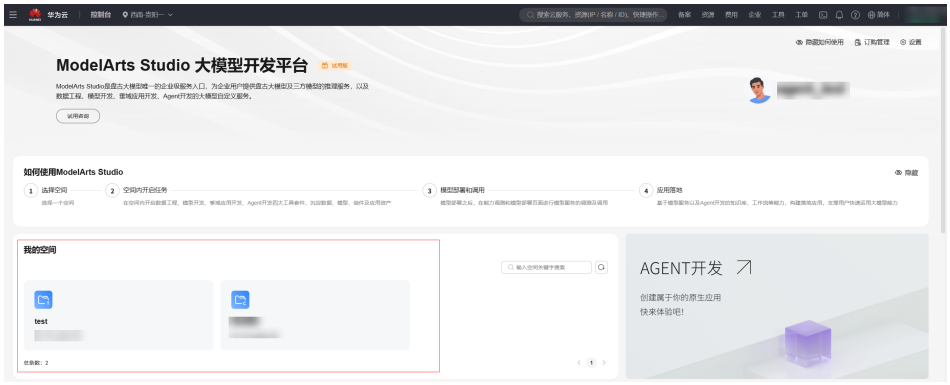
4.6.7 发布预测类数据集

操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-141 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“预测 > 时序（分类）”类型的数据集。

图 4-142 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

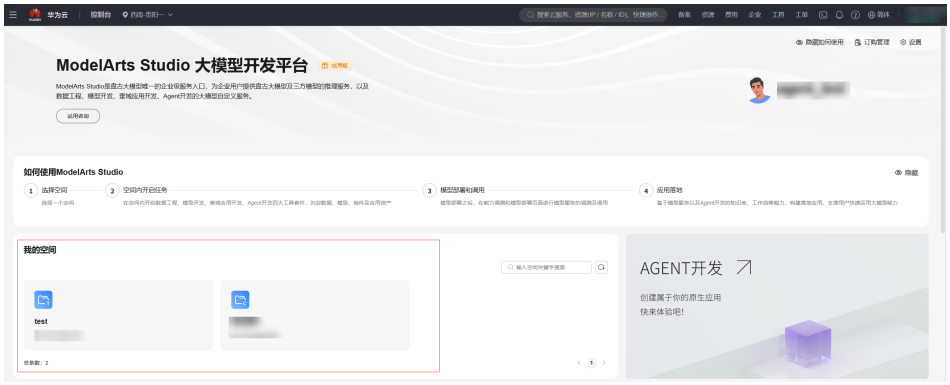
4.6.8 发布其他类数据集

操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-143 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“其他 > 自定义”类型的数据集。

图 4-144 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

4.6.9 管理发布后的数据集

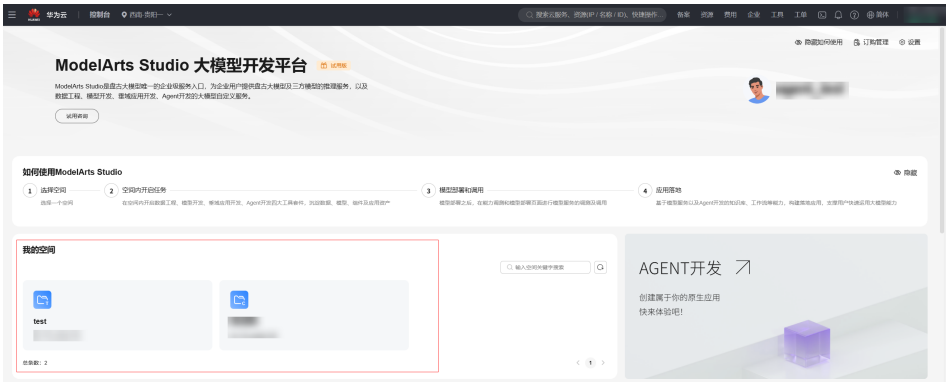
完成数据配比、或数据流通任务的数据集，在对应任务列表执行“生成”操作，将生成“发布数据集”被平台统一管理，并用于后续的发布任务。

平台支持对发布数据集查看基本信息、数据血缘等管理操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 4-145 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据集 > 发布数据集”。
3. 数据集支持“数据集名称”、“数据集模态”、“数据集属性名称”等属性类型进行搜索、过滤。



4. 单击数据集名称查看发布数据集的基本信息、数据预览、数据血缘以及操作记录。
   - 在“基本信息”页签，可以查看数据详情、数据来源以及扩展信息。在扩展信息内，可以根据用户需求设置数据集属性，包括数据集属性名称、行业、语言、自定义标签，如图4-146。

 版权所有 © 华为云计算技术有限公司

图 4-146 编辑扩展信息



- 在“数据预览”页签可查看发布后的数据内容。
- 在“数据血缘”页签查看该数据集所经历的操作，如导入、合成、训练等操作。数据血缘上钻至OBS源文件，下钻至模型训练。
- 在“操作记录”页签可以查看数据集所经历的操作及状态等信息。

5. 单击操作列的“删除”，可删除不需要的数据集。
   - 如果需要恢复删除的数据集，可单击右上角“显示已删除数据”，被删除的数据集将在列表显示，可将数据集恢复。
   - 如果需要彻底删除数据集，可单击数据集名称进入详情页，确认数据集内容后彻底删除该数据集。

说明

删除“发布数据集”属于高危操作，删除前，请确保该数据集不再使用。

4.7 数据集格式转换

4.7.1 使用限制

支持转换的数据集类型

自定义数据集结构：支持图片后缀（jpg、jpeg、png、bmp、tif、tiff）

假设自定义数据集文件名称为：test_自定义数据集

+---test_自定义数据集

| +--- 图片001 # python算子目录

| +--- 图片002 # python算子目录

| +--- 图片003 # python算子目录

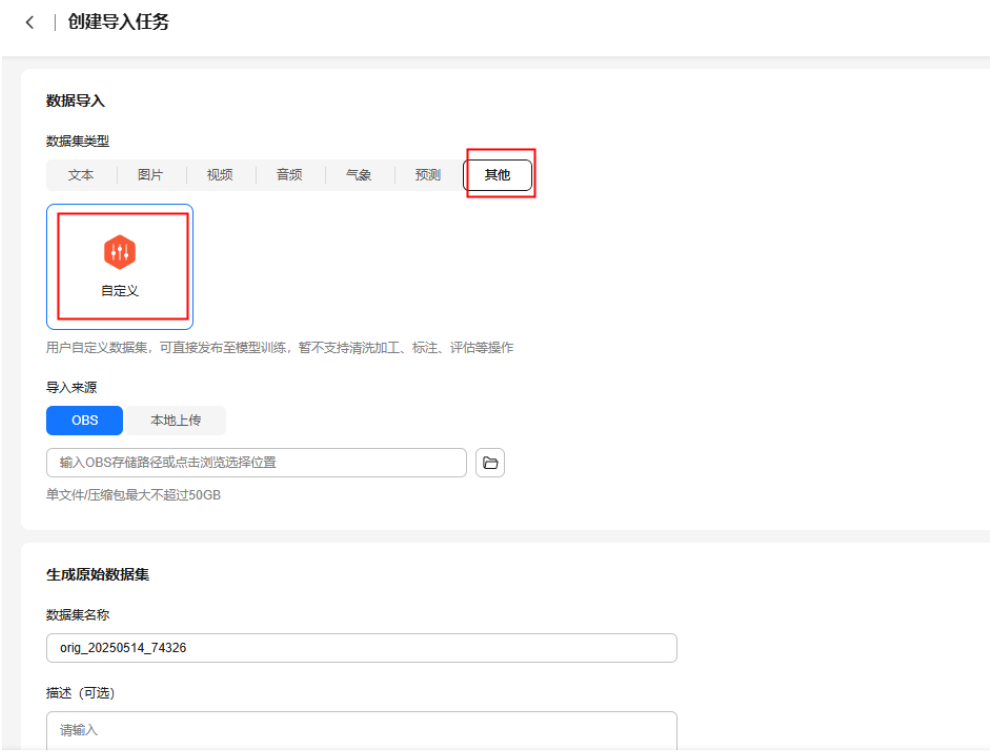
| +--- CSV文件

支持转换的 CV 数据集类型

仅支持自定义数据集转换成图片（物体检测、图像分类、语义分割、实例分割）数据集。

 版权所有 © 华为云计算技术有限公司

图 4-147 导入自定义数据集



4.7.2 格式转换操作流程

构建测试数据 (图1)

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 4-148 我的空间



2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击右上角“创建导入任务”。
3. 在“创建导入任务”页面选择其他--自定义 ，导入需要转换数据集，测试数据如图4-149所示。

版权所有 © 华为云计算技术有限公司

图 4-149 测试数据



4. 返回导航栏，选择“数据工程 > 数据加工 > 加工任务 > 管理加工算子 > 创建自定义算子”，如图4-150所示。

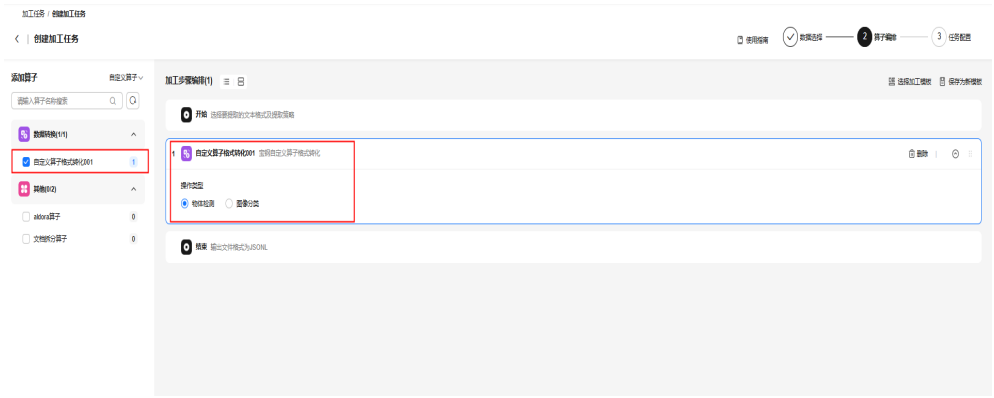
图 4-150 创建自定义算子



5. 传入对应的YML文件和TAR包，构建格式转换算子。
6. 返回导航栏，选择“数据工程 > 数据加工>加工任务> 创建加工任务”。

 版权所有 © 华为云计算技术有限公司

图 4-151 创建加工任务



7. 选择数据集和转换算子，开始格式转换加工任务。

说明

1. 物体检测和图像分类支持jpg、jpeg、bmp、png、tif、tiff格式。
2. csv格式为UTF-8 且读取内容在第一个sheet页。

4.7.3 转换算子说明

转换算子构建包括：①YML配置文件 ②实现功能的python脚本TAR包

①YML文件示例：

```
### python算子示例

id: FormatConvert2
name: 自定义算子格式转化
description: 自定义算子格式转化
author: Huawei Cloud Technology Co., Ltd.
tags:
  language:
    - zh
    - en
  format:
    - CSV
    - TIF
  category: 数据转换
  modal:
    - IMAGE
runtime:
  cpu-arch:
    - X86
  resources:
    - cpu: 1
      memory: 4096
  environment: python
  entrypoint: process.py
  auto-data-loading: false # 是否需要自动加载数据
arguments:
  - key: operateType
    name: 操作类型
    type: ENUM  #单选
    items:
      - name: 物体检测
        value: object_detection
      - name: 图像分类
        value: image_classification
    required: true
    visible: true
    default: object_detection
```

 版权所有 © 华为云计算技术有限公司

```
    """
    return "parquet"
```

②TAR包：

假设算子包名称为operator-package.tar，算子包解压后目录结构：

+---operator-package

| +--- program_package # python算子目录

| | +---dependency

| | | +---requirements.txt #需求说明，可选

| | +--- install.sh # 安装脚本，可选

| | +--- process.py # 算子代码，必填

实现格式转换功能的process.py文件

```
import PIL.Image
import PIL.ImageOps
import numpy as np
import ast
import os
import pandas as pd
import xml.etree.ElementTree as ET
from xml.dom import minidom
import PIL
import moxing as mox
import ma_utils as utils

logger = utils.FileLogger.get_logger()

import argparse


arg_parser = argparse.ArgumentParser()
arg_parser.add_argument("--operateType", type=str, required=True, help="")
arg_parser.add_argument("--input_obs_path", type=str, required=True, help="")
arg_parser.add_argument("--output_obs_path", type=str, required=True, help="")


def prettify_xml(elem):
    rough_string = ET.tostring(elem, 'utf-8')
    reparsed = minidom.parseString(rough_string)
    return reparsed.toprettyxml(indent="  ")


def create_pascal_voc_xml(folder, filename, path, width, height, depth, objects, deu, db, dp, ls, lv, lse, lc, srct,
                          srrt, src, srr, rct, rrt, rc, rr, dest__engine_uuid=None):
    annotation = ET.Element('annotation')
    # 目的端引擎uuid
    dest__engine_uuid = ET.SubElement(annotation, 'DEST_ENGINE_UUID')
    dest__engine_uuid.text = deu
    # 目的端桶(逻辑分区)
    dest__engine_uuid = ET.SubElement(annotation, 'DEST_BUCKET')
    dest__engine_uuid.text = db
    # 目的端路径
    dest_path = ET.SubElement(annotation, 'DEST_PATH')
    dest_path.text = folder
    # 文件名称
    filename_element = ET.SubElement(annotation, 'FILE_NAME')
    filename_element.text = filename
    # 标签组全路径
    dest_path = ET.SubElement(annotation, 'DEST_PATH')
    dest_path.text = dp
    # 标签组全路径
```

 版权所有 © 华为云计算技术有限公司

```
    label_value = ET.SubElement(annotation, 'LABEL_VALUE')
    label_value.text = lv
    # 标签形状 点 / 线 / 矩形(两点) / 矩形(bbox) / 矩形(中心点) / 多边形 / 轮廓
    label_shape = ET.SubElement(annotation, 'LABEL_SHAPE')
    label_shape.text = ls
    # 标签形状 点 / 线 / 矩形(两点) / 矩形(bbox) / 矩形(中心点) / 多边形 / 轮廓
    label_coordinate = ET.SubElement(annotation, 'LABEL_COORDINATE')
    label_coordinate.text = lc
    # 源端记录创建时间
    rec_time = ET.SubElement(annotation, 'SOURCE_REC_CREATE_TIME')
    rec_time.text = srct
    # 源端记录修改时间
    rec_rev_time = ET.SubElement(annotation, 'SOURCE_REC_REVISE_TIME')
    rec_rev_time.text = srrt
    # 源端记录创建人
    rec_ctor = ET.SubElement(annotation, 'SOURCE_REC_CREATOR')
    rec_ctor.text = src
    # 源端记录修改人
    rec_r_ctor = ET.SubElement(annotation, 'SOURCE_REC_REVISOR')
    rec_r_ctor.text = srr
    # 源端记录创建时间
    d_cd_time = ET.SubElement(annotation, 'REC_CREATE_TIME')
    d_cd_time.text = rct
    # 记录修改时间
    r_time = ET.SubElement(annotation, 'REC_REVISE_TIME')
    r_time.text = rrt
    # 记录创建人
    rec_c = ET.SubElement(annotation, 'REC_CREATOR')
    rec_c.text = rc
    # 记录修改人
    recd_time = ET.SubElement(annotation, 'REC_REVISOR')
    recd_time.text = rr
    tree = ET.ElementTree(annotation)
    return tree
def create_pascal_voc_xml1(folder, filename, path, width, height, depth, objects):
    annotation = ET.Element('annotation')

    folder_element = ET.SubElement(annotation, 'folder')
    folder_element.text = folder

    filename_element = ET.SubElement(annotation, 'filename')
    filename_element.text = filename

    path_element = ET.SubElement(annotation, 'path')
    path_element.text = path

    source = ET.SubElement(annotation, 'source')
    database = ET.SubElement(source, 'database')
    database.text = 'Unknown'

    size = ET.SubElement(annotation, 'size')
    width_element = ET.SubElement(size, 'width')
    width_element.text = str(width)
    height_element = ET.SubElement(size, 'height')
    height_element.text = str(height)
    depth_element = ET.SubElement(size, 'depth')
    depth_element.text = str(depth)

    segmented = ET.SubElement(annotation, 'segmented')
    segmented.text = '0'

    for obj in objects:
        obj_element = ET.SubElement(annotation, 'object')
        name_element = ET.SubElement(obj_element, 'name')
        name_element.text = obj['name']

        pose = ET.SubElement(obj_element, 'pose')
        pose.text = 'Unspecified'
```


版权所有 © 华为云计算技术有限公司

```
        truncated = ET.SubElement(obj_element, 'truncated')
        truncated.text = '0'

        difficult = ET.SubElement(obj_element, 'difficult')
        difficult.text = '0'

        bndbox = ET.SubElement(obj_element, 'bndbox')
        xmin = ET.SubElement(bndbox, 'xmin')
        xmin.text = str(obj['xmin'])
        ymin = ET.SubElement(bndbox, 'ymin')
        ymin.text = str(obj['ymin'])
        xmax = ET.SubElement(bndbox, 'xmax')
        xmax.text = str(obj['xmax'])
        ymax = ET.SubElement(bndbox, 'ymax')
        ymax.text = str(obj['ymax'])

    tree = ET.ElementTree(annotation)
    return tree

def get_value(row, name):
    value = '' if pd.isna(row[name]) else str(row[name])
    return value


def copy_local_to_obs(local_file_path, obs_file_path):
    mox.file.copy(local_file_path, obs_file_path)


def csv_to_pascal_voc(args, csv_file, local_tif_file_path_dict, local_output_file_dir):
    operate_type = getattr(args, 'operateType', 'object_detection')
    logger.info(operate_type)
    df = pd.read_csv(csv_file)
    grouped = df.groupby('FILE_NAME')

    for filename, group in grouped:
        if local_tif_file_path_dict.get(filename, None):
            if operate_type == 'object_detection':

                objects = []
                for _, row in group.iterrows():
                    bndbox = ast.literal_eval(row['LABEL_COORDINATE'])
                    label = row['LABEL_VALUE']
                    obj = {
                        'name': label,
                        'xmin': bndbox[0],
                        'ymin': bndbox[1],
                        'xmax': bndbox[2],
                        'ymax': bndbox[3]
                    }
                    objects.append(obj)

                width, height, depth = 0, 0, 0

                folder = os.path.join(row['DEST_BUCKET'], row['DEST_PATH'])
                path = os.path.join(folder, filename)
#                 real_path = os.path.join(test_root, path)
                with PIL.Image.open(local_tif_file_path_dict.get(filename)) as img:
                    width, height = img.size
                    depth = len(img.getbands())
                tree = create_pascal_voc_xml1(folder, filename, path, width, height, depth, objects)
                pretty_xml = prettify_xml(tree.getroot())

                local_output_xml_file_path = os.path.join(local_output_file_dir, os.path.splitext(filename)[0] +
'.xml')
                with open(local_output_xml_file_path, 'w', encoding='utf-8') as f:
                    f.write(pretty_xml)
                fname = replacetype(filename,'1')
                copy_local_to_obs(local_output_xml_file_path,os.path.join(args.output_obs_path, fname))
                copy_local_to_obs(local_tif_file_path_dict.get(filename), os.path.join(args.output_obs_path,
```


版权所有 © 华为云计算技术有限公司

```
filename))
            elif operate_type == 'image_classification':
                for _, row in group.iterrows():
                    logger.info(f'-------------循环row--')

                lv = get_value(row, 'LABEL_VALUE')
                txtstr = create_pascal_voc_txt(lv)
                fname = replacetype(filename,'2')
                local_output_txt_file_path = os.path.join(local_output_file_dir, fname)
                with open(local_output_txt_file_path, 'w', encoding='utf-8') as f:
                    f.write(txtstr)
                copy_local_to_obs(local_output_txt_file_path,
                                  os.path.join(args.output_obs_path, fname))
                copy_local_to_obs(local_tif_file_path_dict.get(filename), os.path.join(args.output_obs_path,
filename))




def replacetype(filename,type):
    if type == '1':
        filename = filename.replace('.jpg', '.xml')
        filename = filename.replace('.bmp', '.xml')
        filename = filename.replace('.jpeg', '.xml')
        filename = filename.replace('.png', '.xml')
        filename = filename.replace('.JPG', '.xml')
        filename = filename.replace('.BMP', '.xml')
        filename = filename.replace('.JPEG', '.xml')
        filename = filename.replace('.PNG', '.xml')
    elif type == '2':
        filename = filename.replace('.jpg', '.txt')
        filename = filename.replace('.bmp', '.txt')
        filename = filename.replace('.jpeg', '.txt')
        filename = filename.replace('.png', '.txt')
        filename = filename.replace('.JPG', '.txt')
        filename = filename.replace('.BMP', '.txt')
        filename = filename.replace('.JPEG', '.txt')
        filename = filename.replace('.PNG', '.txt')
    return filename


def create_pascal_voc_txt(lv):
    # tree='DEST_ENGINE_UUID:'+deu+'\n'+'DEST_BUCKET:'+db
#     txttree = 'DEST_ENGINE_UUID:' + deu + '\n' + 'DEST_BUCKET:' + db + '\n' + 'DEST_PATH:' + dp + '\n' +
'LABEL_SET:' + ls + '\n' + 'LABEL_VALUE:' + lv + '\n' + 'LABEL_SHAPE:' + lse + '\n' + 'LABEL_COORDINATE:' +
lc + '\n' + 'S#OURCE_REC_CREATE_TIME:' + srct + '\n' + 'SOURCE_REC_REVISE_TIME:' + srrt + '\n' +
'SOURCE_REC_CREATOR:' + src + '\n' + 'SOURCE_REC_REVISOR:' + srr + '\n' + 'REC_CREATE_TIME:' + rct +
'\n' + 'REC_REVISE_TIME:' + rrt + '\n' + 'RE#C_CREATOR:' + rc + '\n' + 'REC_REVISOR:' + rr
    txttree = lv

    return txttree


class Process:
    def __init__(self, args):
        # 创建本地目录用以存储从obs传入的文件
        self.local_input_file_dir = r'/tmp/obs_input/'
        os.makedirs(self.local_input_file_dir, exist_ok=True)
        # 创建本地目录用以存储处理后的文件，从而传输到obs
        self.local_output_file_dir = r'/tmp/obs_output/'
        os.makedirs(self.local_output_file_dir, exist_ok=True)
        # 设置参数，针对算子控制参数，设置默认值
        self.args = args
        self.input_obs_path=args.obs_input_path
        self.output_obs_path=args.obs_output_path
        self.args.input_obs_path=args.obs_input_path
        self.args.output_obs_path=args.obs_output_path
        logger.info(f'-------算子执行--------')
```


版权所有 © 华为云计算技术有限公司

```
    def __call__(self, input):
        # 获取obs路径下的文件名
        file_name_list = mox.file.list_directory(self.args.input_obs_path, recursive=False)
        file_path_dict = {file_name: os.path.join(self.args.input_obs_path, file_name) for file_name in
file_name_list}
        # 对每个文件进行处理
        local_tif_file_path_dict, local_csv_file_path = {}, ''
        for file_name, obs_file_path in file_path_dict.items():
            # 将文件从obs复制到本地

#             if file_name.endswith('.jpg'):
#                 local_tif_file_path = os.path.join(self.local_input_file_dir, file_name)
#                 logger.info(f'---------------4'+local_tif_file_path)
#                 local_tif_file_path_dict[file_name] = local_tif_file_path
#                 mox.file.copy(obs_file_path, local_tif_file_path)
#             else:
#                 local_csv_file_path = os.path.join(self.local_input_file_dir, file_name)
#                 logger.info(f'---------------5'+local_csv_file_path)
#                 mox.file.copy(obs_file_path, local_csv_file_path)

            if file_name.endswith('.csv'):

                local_csv_file_path = os.path.join(self.local_input_file_dir, file_name)
                mox.file.copy(obs_file_path, local_csv_file_path)
            else:
                local_tif_file_path = os.path.join(self.local_input_file_dir, file_name)
                local_tif_file_path_dict[file_name] = local_tif_file_path
                mox.file.copy(obs_file_path, local_tif_file_path)

        csv_to_pascal_voc(self.args, local_csv_file_path, local_tif_file_path_dict, self.local_output_file_dir)
```

4.8 数据工程常见报错与解决方案

数据工程常见报错及解决方案请详见**表4-49**。

版权所有 © 华为云计算技术有限公司

表 4-49 数据工程常见报错与解决方案

功能模块	常见报错	解决方案
数据获取	数据解析错误，详情请查看日志。	进入导入任务详情，切换tab页到日志管理，查看运行日志： 1.File format mismatch, require [XXX] 请检查创建数据集时使用的数据，与平台要求的文件内容格式是否一致。 2. Verification failed. Please check the content format is consistent with the template requirements. Line number: [XXX]. 请检查创建数据集时使用的数据，与平台要求的文件内容格式是否一致。 3.There is no jsonl file in the directory. 缺少JSONL文件文件。 4.invalid target column: [XXX]. 预测数据集存在无效的目标行。 5.The file [XXX] size exceeds the maximum limit of [XXX] MB. 文件大小超过上限。
	服务内部异常，请联系技术人员。	进入导入任务详情，切换tab页到日志管理，查看运行日志： 1.Internal service error. The error message is: [Corrupted TAR archive.]。 tar包格式错误，请检查tar包格式是否正确。 2.Parse token failed. 解析用户token失败，检查用户信息。
	参数配置错误。	进入导入任务详情，切换tab页到日志管理，查看运行日志： Illegal argument [%s].检查参数是否正确。
	依赖服务(Gallery)错误。	进入导入任务详情，切换tab页到日志管理，查看运行日志： 1.Request Gallery failed. Gallery response: [XXX]. 请确认Gallery服务是否正常。

版权所有 © 华为云计算技术有限公司

功能模块	常见报错	解决方案
	依赖服务(MRS)错误。	进入导入任务详情，切换tab页到日志管理，查看运行日志： 1.Call mrs api failed: [XXX].请确认MRS服务是否正常。 2.Mrs job is failed, job id: [XXX].请确认MRS服务是否正常。 3.The user does not order mrs service periodic.请确认是否订购MRS服务。
	依赖服务(OBS)错误。	进入导入任务详情，切换tab页到日志管理，查看运行日志： The user has not subscribed to OBS-related services.请确认是否订购OBS服务。
数据加工	dataset is not online.	数据加工使用的数据集未上线，请先执行上线操作。
	invalid obs path.	请检查数据集对应的OBS路径是否有效，是否可正常访问。
数据合成	模型未配置。	重新部署模型。
	依赖服务【 MA 】错误，请确认服务【 MA 】是否正常。	请确认MA服务是否正常。
	执行MA任务失败，详情请查看运行日志。	请查看运行日志。
	任务不存在。	重新创建合成任务。
	服务内部异常，请联系技术人员。	请联系技术人员。
	数据集生成失败，请检查数据管理状态并查看OBS资源是否充足。	请检查数据管理状态并查看OBS资源是否充足。
数据标注	数据解析错误，详情请查看日志。	请检查标注数据集是否符合规范。
	依赖资源{数据集/任务}错误，请确认{数据集/任务}正常。	请确认{数据集/任务}正常。
	服务内部异常，请联系技术人员。	请联系技术人员。
	依赖服务{MRS/MA/OBS/Gallery/TICS}错误，请确认服务{MRS/MA/OBS/Gallery/TICS}是否正常。	请确认服务{MRS/MA/OBS/Gallery/TICS}是否正常。

版权所有 © 华为云计算技术有限公司

功能模块	常见报错	解决方案
	数据库访问异常。	请检查数据库是否正常。
数据配比	参数配置错误，请调整配置参数。	请调整配置参数。
	权限(本地目录/服务访问)错误，请配置相应权限。	请配置相应权限。
	依赖资源{数据集/任务}错误，请确认{数据集/任务}正常。	请确认{数据集/任务}正常。
	服务内部异常，请联系技术人员。	请联系技术人员。
	依赖服务{MRS/MA/OBS/GaLLery/TICS}错误，请确认服务{MRS/MA/OBS/GaLLery/TICS}是否正常。	请确认服务{MRS/MA/OBS/Gallery/TICS}是否正常。
	依赖组件（itep/空间/biLLingmanager）错误。	请确认组件（itep/空间/biLLingmanager）是否正常。
数据评估	annotate type is invalid.	请检查上传的数据中，使用的数据标注类型、数据标注要求与平台要求的是否一致。
	annotate data not exist.	待评测数据不存在，请检查数据是否导入成功，OBS桶是否为空。
	obs url invalid.	请检查数据集对应的OBS路径是否有效，是否可正常访问。
	standard item not exist.	请检查评估标准是否存在，是否被删除。
	the corresponding data has been marked as deleted.	请检查评估数据是否被删除。
	the current data not exist.	请检查评估数据是否存在，是否被删除。
	dataset file type does not match standard file type.	请检查上传的数据集文件类型与平台要求的标准文件类型是否一致。
	data management query dataset data invalid.	请检查数据集中是否有异常格式的数据。
	dataset obs file empty.	检查数据集文件是否还存在于原先的OBS桶中。
数据发布	参数配置错误，请调整配置参数。	请调整配置参数。

版权所有 © 华为云计算技术有限公司

功能模块	常见报错	解决方案
	权限(本地目录/服务访问)错误，请配置相应权限。	请配置相应权限。
	依赖资源{数据集/任务}错误，请确认{数据集/任务}正常。	请确认{数据集/任务}正常。
	服务内部异常，请联系技术人员。	请联系技术人员。
	依赖服务{MRS/MA/OBS/GaLLery/TICS}错误，请确认服务{MRS/MA/OBS/GaLLery/TICS}是否正常。	请确认服务{MRS/MA/OBS/Gallery/TICS}是否正常。
	依赖组件（itep/空间/biLLingmanager）错误。	请确认组件（itep/空间/biLLingmanager）是否正常。

版权所有 © 华为云计算技术有限公司

5 开发盘古 NLP 大模型

使用数据工程构建NLP大模型数据集
训练NLP大模型
压缩NLP大模型
部署NLP大模型
评测NLP大模型
调用NLP大模型

5.1 使用数据工程构建 NLP 大模型数据集

NLP 大模型支持接入的数据集类型

盘古NLP大模型仅支持接入文本类数据集，数据集文件内容包括：预训练文本、单轮问答、多轮问答、带人设单轮问答、带人设多轮问答、偏好优化DPO等，同时新增了Reasoner模型，Reasoner模型是一种基于逻辑推理或知识推理的AI模型。不同训练方式所需要使用的数据见**表5-1**，该数据集格式要求请参见**文本类数据集格式要求**。

表 5-1 训练 NLP 大模型数据集类型要求

基模型	训练场景	数据集类型	数据集内容	格式	文件格式
NLP	预训练	文本	预训练文本	NA	jsonl
	微调（快思考模型）	文本	单轮问答	盘古格式-非思维链	jsonl、csv
		文本	多轮问答	盘古格式-非思维链	jsonl

版权所有 © 华为云计算技术有限公司

基模型	训练场景	数据集类型	数据集内容	格式	文件格式
		文本	单轮问答（人设）	盘古格式-非思维链	jsonl、csv
		文本	多轮问答（人设）	盘古格式-非思维链	jsonl
	微调（慢思考模型）	文本	单轮问答	盘古格式-思维链	jsonl、csv
		文本	多轮问答	盘古格式-思维链	jsonl
		文本	单轮问答（人设）	盘古格式-思维链	jsonl、csv
		文本	多轮问答（人设）	盘古格式-思维链	jsonl
	强化学习-RFT（快思考模型）	文本	单轮问答	盘古格式-非思维链	jsonl
	强化学习-DPO（快思考模型）	文本	偏好优化DPO	盘古格式-非思维链	jsonl

训练 NLP 大模型所需数据量

使用数据工程构建盘古NLP大模型数据集进行模型训练时，所需数据量见**表5-2**。

表 5-2 构建 NLP 大模型所需数据量

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N1	微调	-	1000条/每场景	≥ 1万条/每场景	32K
N2	微调	-	1000条/每场景	≥ 1万条/每场景	32K

版权所有 © 华为云计算技术有限公司

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N4	微调	-	1000条/每场景	≥ 1万条/每场景	4K版本：4096 32K版本：32768
	预训练	4GB(等价10亿Tokens)	-	-	4K版本：4096 32K版本：32768
	RFT强化学习	100 ~ 1万条	100条	-	
	DPO强化学习	5万 ~ 15万条	1000条	-	

评测 NLP 大模型所需数据量

要求所有文本大小最大不超过100MB，目录下文件数量最多不超过100个。数据条数范围为：3-1000条。

构建 NLP 大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程构建盘古NLP大模型数据集流程见表5-3。

表 5-3 盘古 NLP 大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。慢思考模型的数据相比快思考模型的数据新增了思考部分的内容。例如对于单轮对话微调数据：快思考一条数据样例的格式：{"context":["XX"], "target":"答案"}，慢思考模型的数据样例：{"context":["XX"], "target":"<think>xx</think> 答案"}	导入数据至盘古平台
加工文本类数据集	加工文本类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	加工文本类数据集

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	合成文本类数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。	合成文本类数据集
	标注文本类数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。	标注文本类数据集
	配比文本类数据集	数据配比是将多个数据集按特定比例组合的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。	配比文本类数据集
发布文本类数据集	评估文本类数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。	评估文本类数据集
	发布文本类数据集	发布流程是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。为适配不同盘古模型的数据规范，文本类型-单轮问答、单轮问答（带人设）、多轮问答、多轮问答（带人设）、DPO、DPO（带人设）发布时需选择盘古格式-非思维链或盘古格式-思维链进行发布。 • 标准格式：平台默认的格式。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。	发布文本类数据集

5.2 训练 NLP 大模型

版权所有 © 华为云计算技术有限公司

5.2.1 NLP 大模型训练流程与选择建议

NLP 大模型训练流程介绍

NLP大模型的训练分为两个关键阶段：预训练和微调。

- **预训练阶段：**在这一阶段，模型通过学习大规模通用数据集来掌握语言的基本模式和语义。这一过程为模型提供了处理各种语言任务的基础，如阅读理解、文本生成和情感分析，但它还未能针对特定任务进行优化。

  针对**预训练阶段**，还可以继续进行训练，这一过程称为**增量预训练**。增量预训练是在已经完成的预训练的基础上继续训练模型。增量预训练旨在使模型能够适应新的领域或数据需求，保持其长期的有效性和准确性。

- **微调阶段：**基于预训练的成果，微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。这一阶段使模型能够精确执行如文案生成、代码生成和专业问答等特定场景中的任务。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。

  针对**微调阶段，**微调后的模型可以直接部署，也可以继续进行**强化学习（DPO、RFT、GRPO）**训练，以进一步对齐任务场景下的用户偏好，例如模型回答的风格、模型的价值观等。

NLP 大模型选择建议

选择合适的NLP大模型类型有助于提升训练任务的准确程度。您可以根据模型**可处理最大Token长度**，选择合适的模型，从而提高模型的整体效果，详见**表5-4**。

此外，不同类型的NLP大模型在训练过程中，读取中文、英文内容时，字符长度转换为Token长度的转换比有所不同，详见**表5-5**。

表 5-4 不同系列 NLP 大模型对处理文本的长度差异

模型名称	可处理最大序列长度（输入+输出）
Pangu-NLP-N1-32K-3.1.34	32K
Pangu-NLP-N1-32K-3.2.36	32K
Pangu-NLP-N1-128K-3.1.34	128K
Pangu-NLP-N1-128K-3.2.36	128K
Pangu-NLP-N1-Reasoner-128K-3.0.1.1	128K
Pangu-NLP-N2-4K-3.2.35	4K
Pangu-NLP-N2-32K-3.1.35	32K
Pangu-NLP-N2-128K-3.1.35	128K
Pangu-NLP-N2-256K-3.1.35	256K
Pangu-NLP-N2-Reasoner-32K-3.0.0.1	32K
Pangu-NLP-N2-Reasoner-128K-3.0.1.1	128K

版权所有 © 华为云计算技术有限公司

模型名称	可处理最大序列长度（输入+输出）
Pangu-NLP-N2-Reasoner-128K-5.0.0.1	128K
Pangu-NLP-N4-4K-3.2.36	4K
Pangu-NLP-N4-4K-2.5.32	4K
Pangu-NLP-N4-4K-2.5.35	4K
Pangu-NLP-N4-32K-2.5.32	32K
Pangu-NLP-N4-32K-2.5.35	32K
Pangu-NLP-N4-Reasoner-32K	32K
Pangu-NLP-N4-Reasoner-128K-3.0.1.2	128K
Pangu-NLP-N4-MoE-Reasoner-32K-5.0.0.1	32K
Pangu-RAG-N1-32K-4.3.2	32K
Pangu-AgentExpert-N1-0.0.2	32K
Pangu-AgentExpert-N2-0.0.2	32K

说明

模型名称中带Reasoner的为Reasoner模型，例如：Pangu-NLP-N4-Reasoner-32K。否则为快思考模型，例如：Pangu-NLP-N4-4K。

表 5-5 Token 转换比

模型规格	Token比（Token/英文单词）	Token比（Token/汉字）
N1	0.75	1.5
N2	0.88	1.24
N4	0.75	1.5

说明

针对Token转换比，平台提供了**Token计算器**功能，可以根据您输入的文本计算Token数量，您可以通过以下方式使用该功能：

- 在左侧导航栏选择“体验中心”，单击“选择其他服务”，选择预置大模型或自己的大模型后，单击右上角“Token计算器”使用该功能。
- 使用API调用Token计算器，详见《API参考》“API > Token计算器”。

 版权所有 © 华为云计算技术有限公司

NLP 大模型训练类型选择建议

平台针对NLP大模型提供了三种训练类型，包括预训练、微调、强化学习（DPO/RFT/GRPO），三者区别详见表5-6。

表 5-6 预训练、微调训练类型区别

训练方式	训练目的	训练数据	模型效果	应用场景举例
预训练	**关注通用性**：预训练旨在让模型学习广泛的通用知识，建立词汇、句法和语义的基础理解。通过大规模的通用数据训练，模型可以掌握丰富的语言模式，如语言结构、词义关系和常见的句型。	**使用大规模通用数据**：通常使用海量的无监督数据（如文本语料库、百科文章），这些数据覆盖广泛的领域和语言表达方式，帮助模型掌握广泛的知识。	**适合广泛应用**：经过预训练后，模型可以理解自然语言并具备通用任务的基础能力，但还没有针对特定的业务场景进行优化。预训练后的模型主要用于多个任务的底层支持。	通过使用海量的互联网文本语料对模型进行预训练，使模型理解人类语言的基本结构。
微调	**关注专业性**：微调是对预训练模型的参数进行调整，使其在特定任务中达到更高的精度和效果。微调的核心在于利用少量的特定任务数据，使模型的表现从通用性向具体任务需求过渡。	**使用小规模的特定任务数据**：微调通常需要小规模但高质量的标注数据，直接与目标任务相关。通过这些数据，模型可以学习到任务特定的特征和模式。	**在特定任务上具有更高的准确性**：微调后的模型在具体任务中表现更优。相较于预训练阶段的通用能力，微调能使模型更好地解决细分任务的需求。	在一个客户服务问答系统中，可以用特定领域（如电商、保险）的对话数据对预训练模型进行微调，使其更好地理解和回答与该领域相关的问题。

版权所有 © 华为云计算技术有限公司

训练方式	训练目的	训练数据	模型效果	应用场景举例
强化学习（DPO）	**关注偏好性：**DPO（Direct Preference Optimization，直接偏好优化）通过利用偏好数据来提升模型的对话质量，增强其安全性、可控性，并且优化其输出的风格和语气。	**使用模型生成的多个回答的排序数据：**这些偏好数据通常来自于已有的高质量对话数据集。关键在于数据的质量，需要保证生成的偏好数据能够有效反映人类的实际偏好和期望。	• **对话质量**：通过优化生成的回复，使其更自然、更符合语境，并减少不当或低质量的内容。 • **语气和风格调整**：可以根据需要调整模型的语气（如正式或亲和）和输出风格（如简洁或详细）。 • **安全性和可控性**：减少有害或偏见内容的生成，增强模型的可控性，能够避免生成不符合伦理或社会规范的回答。	在内容创作和写作辅助中，DPO能够帮助创作类工具生成更加符合用户风格和偏好的文本内容，支持个性化写作。

 版权所有 © 华为云计算技术有限公司

训练方式	训练目的	训练数据	模型效果	应用场景举例
强化学习（RFT）	**关注正确性：**RFT（Reinforcement Fine-Tuning，强化微调）是一种结合规则化奖励信号的强化学习方法。其训练目的是引导模型在具有明确“正确答案”的复杂任务中不断优化输出结果，使其逐步逼近最优解。RFT通过少量高质量反馈数据，使模型在特定领域实现高精度推理与决策能力。	**使用规则化奖励函数对模型生成的回答打分：**RFT的训练数据通常包括输入问题与标答。由模型自动生成的多个候选回答，每个候选回答会通过一个预定义的规则化奖励函数进行评分。这些得分作为强化学习中的奖励信号，用于指导策略更新，使模型更倾向于生成“正确”或“有效”的输出。	• **高精度推理能力**：RFT能显著提升模型在数学推理、代码生成、逻辑推导等任务上的准确率，使其接近甚至达到专家水平。 • **稳定性和一致性增强**：通过规则化反馈机制，模型在重复任务中表现更加稳定，减少随机性导致的错误。 • **可解释性提升**：由于依赖明确的规则奖励机制，RFT训练出的模型在输出路径上更具可追踪性和可解释性，便于调试和优化。 • **泛化能力有限但可控**：虽然RFT更聚焦于特定领域的性能优化，泛化能力相对较弱，但其输出更加可控，适合封闭式任务场景。	在数学问题求解中，RFT应用于解题系统，通过对答案是否正确进行反馈，让模型学会一步步推导出正确的解法。

版权所有 © 华为云计算技术有限公司

训练方式	训练目的	训练数据	模型效果	应用场景举例
强化学习（GRPO）	**关注相对性：**GRPO（Group Relative Policy Optimization，群组相对策略优化）是一种基于组间多个回复的相对优势水平进行训练的强化学习方法。其核心思想是根据模型多个采样回答的平均奖励计算奖励基线，并得到每个回答的相对优势值，进而用于迭代训练、不断优化提升模型的推理决策能力。	**使用组内多个回答的平均奖励作为基准：**GRPO的训练数据包括输入问题与标答。同一个输入问题的多次采样回答作为一个群组。GRPO会通过预定义的规则化奖励函数对每个采样回答进行评分。评分会用于计算群组内的基准奖励值，进而可以计算出每个回答相对于基准的优势水平，优势值会用于指导模型训练迭代。	• **复杂推理能力提升**：GRPO可以借助规则式奖励函数进行评分，显著提升模型在数学推理、逻辑问答等复杂任务中的表现。 • **计算资源降低**：GRPO的基准值计算机制替代了传统PPO算法中的价值模型，能有效减少内存和计算资源的使用。 • **稳定性增强**：GRPO通过KL约束机制防止新模型过度偏离参考模型，增强了模型的训练稳定性。	GRPO应用于数学计算、逻辑推理等复杂问题的求解场景中，激发模型的高级推理行为，包括自我反思以及长链思维，能有效提升模型的解题能力。

此外，针对微调训练任务，平台提供了两种微调方式：

- **全量微调**：适合有充足数据并关注特定任务性能的场景。在全量微调中，模型的所有参数都会调整，以适应特定任务的需求。这种方式适合样本量较大、对推理效果要求较高的任务。例如，在特定领域（如金融、医疗）中，若拥有大量标注数据，且需要更高的特定任务推理精度，则全量微调是优先选择。
- **LoRA微调**：适用于数据量较小、侧重通用任务的情境。LoRA（Low-Rank Adaptation）微调方法通过调整模型的少量参数，以低资源实现较优结果，适合聚焦于领域通用任务或小样本数据情景。例如，在针对通用客服问答的场景中，样本量少且任务场景广泛，选择LoRA微调既能节省资源，又能获得较好的效果。

微调方式选择建议：

- 若项目中数据量有限或任务场景较为广泛，可以选择**LoRA微调**以快速部署并保持较高适用性。
- 若拥有充足数据且关注特定任务效果，选择**全量微调**有助于大幅提升在特定任务上的模型精度。

创建训练任务如何配置资源：

1. 创建训练任务页面默认资源配置为该模型支持的最小资规格源配置；
2. 在训练任务创建页面，如果单实例训练单元数可以配置为8，且实例数可以配置为大于1，则说明当前模型支持多机多卡训练；

 版权所有 © 华为云计算技术有限公司



5.2.2 NLP 大模型默认 prompt

NLP大模型针对不同的模型、不同模型的版本预置了默认的prompt，具体见**表5-7**。

版权所有 © 华为云计算技术有限公司

表 5-7 模型预置默认 prompt

模型名称	prompt	样例
Pangu-NLP-N1-32K-3.1.34	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"]，"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。"，"1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。"，"1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N1-32K-3.2.36	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"], "context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"], "context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N1-Reasoner-128K-3.0.1.1	[unused9]系统：[unused10][unused9]用户：	输入示例1-**单轮对话**： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "<think>好，我现在要将输入翻译成中文</think>我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：<think>好，我现在要将输入翻译成中文</think>\n\n我每天喜欢学习新事物。[unused10] 输入示例2-**单轮对话 + system prompt**： {"system": "你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。你的解答中包含推理思考与最终响应两部分，请先在<think>...</think>标签中进行推理思考，再给出最终响应。", "context": "在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。", "target": "<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**"} 数据处理后： [unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。你的解答中包含推理思考与最终响应两部分，请先在<think>...</think>标签中进行推理思考，再给出最终响应。[unused10][unused9]用户：在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。 [unused10][unused9]助手：<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>\n\n老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**[unused10] 输入示例3-**多轮对话**： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。", "1 + 1。"], "target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"}

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
		数据处理后： [unused9]系统：[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：<think>嗯，我需要计算2453 + 4567的值</think>\n\n两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：<think>嗯，我需要思考1+1等于多少，应该为2</think>\n\n 2[unused10] 输入示例4-**多轮对话 + system prompt**： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。"，"1 + 1。"]，"target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：<think>嗯，我需要计算2453 + 4567的值</think>\n\n两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：<think>嗯，我需要思考1+1等于多少，应该为2</think>\n\n 2[unused10]

 版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N2-Reasoner-32K-3.0.0.1	[unused9]系统：[unused10][unused9]用户：	输入示例1-**单轮对话**： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "<think>好，我现在要将输入翻译成中文</think>我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：<think>好，我现在要将输入翻译成中文</think>\n\n我每天喜欢学习新事物。[unused10] 输入示例2-**单轮对话 + system prompt**： {"system": "你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。你的解答中包含推理思考与最终响应两部分，请先在<think>...</think>标签中进行推理思考，再给出最终响应。", "context": "在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。", "target": "<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**"} 数据处理后： [unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。你的解答中包含推理思考与最终响应两部分，请先在<think>...</think>标签中进行推理思考，再给出最终响应。[unused10][unused9]用户：在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。 [unused10][unused9]助手：<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>\n\n老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**[unused10] 输入示例3-**多轮对话**： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。", "1 + 1。"], "target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"}

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
		数据处理后： [unused9]系统：[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：<think>嗯，我需要计算2453 + 4567的值</think>\n\n两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：<think>嗯，我需要思考1+1等于多少，应该为2</think>\n\n 2[unused10] 输入示例4-**多轮对话 + system prompt**： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。"，"1 + 1。"]，"target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：<think>嗯，我需要计算2453 + 4567的值</think>\n\n两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：<think>嗯，我需要思考1+1等于多少，应该为2</think>\n\n 2[unused10]

 版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N2-Reasoner-128K-5.0.0.1	[unused9]系统：[unused10][unused9]用户：	输入示例1-**单轮对话**： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "<think>好，我现在要将输入翻译成中文</think>我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：[unused16]好，我现在要将输入翻译成中文[unused17]我每天喜欢学习新事物。[unused10] 输入示例2-**单轮对话 + system prompt**： {"system": "你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。", "context": "在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。", "target": "<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**"} 数据处理后： [unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。 [unused10][unused9]助手：[unused16]好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n[unused17]老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**[unused10] 输入示例3-**多轮对话**： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。", "1 + 1。"], "target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：[unused16]嗯，我需要

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
		计算2453 + 4567的值[unused17]两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：[unused16]嗯，我需要思考1+1等于多少，应该为2[unused17] 2[unused10] 输入示例4-**多轮对话 + system prompt**： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。"，"1 + 1。"]，"target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：[unused16]嗯，我需要计算2453 + 4567的值[unused17]两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：[unused16]嗯，我需要思考1+1等于多少，应该为2[unused17] 2[unused10]
Pangu-NLP-N2-4K-3.2.35	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."]，"target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"]，"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."]，"target": "我每天喜欢学习新事物。"} 数据处理后： [[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。"，"1 + 1。"]，"target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。"，"1 + 1。"]，"target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N2-32K-3.1.35	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"], "context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"], "context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N4-4K-2.5.35	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"], "context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"], "context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N4-32K-2.5.35	[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：	输入示例1-单轮对话： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例2-单轮对话 + system prompt： {"system": ["你是一个翻译高手。"], "context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个翻译高手。[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：我每天喜欢学习新事物。[unused10] 输入示例3-多轮对话： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10] 输入示例4-多轮对话 + system prompt： {"system": ["你是一个算术高手。"], "context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "两数相加的结果为:6920。", "1 + 1。"], "target": "2"} 数据处理后： [unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：2[unused10]

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
Pangu-NLP-N4-Reasoner-32K	[unused9]系统：[unused10][unused9]用户：	输入示例1-**单轮对话**： {"context": ["给定一个英文句子，翻译成中文。\nI love to learn new things every day."], "target": "<think>好，我现在要将输入翻译成中文</think>我每天喜欢学习新事物。"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给定一个英文句子，翻译成中文。\nI love to learn new things every day. [unused10][unused9]助手：[unused16]好，我现在要将输入翻译成中文[unused17]我每天喜欢学习新事物。[unused10] 输入示例2-**单轮对话 + system prompt**： {"system": "你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。", "context": "在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。", "target": "<think>好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n</think>老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**"} 数据处理后： [unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：在一次定向越野活动中，老郎的移动轨迹如下：他首先向西走85米，随后改变方向向东走102米，接着再次转向西走73米，最后向东走169米。根据这一系列的移动，请分析老郎最终与起点的相对位置关系。 [unused10][unused9]助手：[unused16]好，我现在要解决这个问题。老郎在定向越野活动中移动了四次，每次都是向西或向东走不同的距离。我需要计算他最终相对于起点的位置。\n\n首先，我应该明确方向。通常，东和西是相反的方向，所以可以将向东走定为正方向，向西走则为负方向。这样，每次移动的距离就可以转换成正数或负数，然后相加得到总位移。\n\n老郎第一次向西走了85米，所以这是-85米。第二次他向东走了102米，也就是+102米。第三次又向西走了73米，记作-73米。第四次向东走了169米，即+169米。\n\n接下来，我把这些数值加起来：-85 + 102 -73 +169。先算前两个数，-85加102等于17米。然后加上-73，17减73等于-56米。最后加上169，-56加169等于113米。\n\n所以，总位移是113米，正数表示向东。因此，老郎最终位于起点以东113米的位置。\n[unused17]老郎最终的位置在起点以东113米处。以下是详细步骤：\n\n1. 将方向转换为数值：向东为正，向西为负。\n2. 计算每次移动的数值：\n - 向西85米 → -85米\n - 向东102米 → +102米\n - 向西73米 → -73米\n - 向东169米 → +169米\n3. 累加所有数值：-85 + 102 = 17；17 - 73 = -56；-56 + 169 = 113米。\n4. 结果为正，说明老郎在起点东侧113米处。\n\n**答案：老郎最终位于起点以东113米的位置。**[unused10] 输入示例3-**多轮对话**： {"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。", "1 + 1。"], "target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：[unused16]嗯，我需要

版权所有 © 华为云计算技术有限公司

模型名称	prompt	样例
		计算2453 + 4567的值[unused17]两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：[unused16]嗯，我需要思考1+1等于多少，应该为2[unused17] 2[unused10] 输入示例4-**多轮对话 + system prompt**： {"system": ["你是一个算术高手。"]，"context": ["给出一个数学运算，计算它的结果。\n2453 + 4567。", "<think>嗯，我需要计算2453 + 4567的值</think>两数相加的结果为:6920。"，"1 + 1。"]，"target":<think>嗯，我需要思考1+1等于多少，应该为2</think> "2"} 数据处理后： [unused9]系统：你是一个算术高手。[unused10][unused9]用户：给出一个数学运算，计算它的结果。\n2453 + 4567。[unused10][unused9]助手：[unused16]嗯，我需要计算2453 + 4567的值[unused17]两数相加的结果为:6920。[unused10][unused9]用户：1 + 1。[unused10][unused9]助手：[unused16]嗯，我需要思考1+1等于多少，应该为2[unused17] 2[unused10]

5.2.3 创建 NLP 大模型训练任务

创建 NLP 大模型预训练任务

创建NLP大模型预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 5-1 我的空间

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表5-8**完成训练参数设置，参数默认值在创建训练任务的时候会带出。

版权所有 © 华为云计算技术有限公司

表 5-8 NLP 大模型预训练参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制，详细配置参见**断点续训配置**。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。 开启断点续训后对于未训练完成的模型，可以通过单击训练任务名称，进入“训练结果”页签，通过“编辑训练”入口选中需要进行断点续训的checkpoints进行续训。
训练参数	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	数据批量大小	数据集进行分批读取训练，设定每个批次数据的大小。 通常情况下，较大的数据批量可以使梯度更加稳定，从而有利于模型的收敛。然而，较大的数据批量也会占用更多的显存资源，这可能导致显存不足，并且会延长每次训练的时长。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	模型保存策略	save_checkpoint_steps/save_checkpoint_epoch，训练过程中是按迭代步数，还是训练轮数保存Checkpoint文件。
	Checkpoint保存间隔	save_checkpoint_steps，训练过程中每隔多少个训练步长保存一次模型Checkpoint文件。
	Checkpoint保存轮数	save_checkpoint_epoch，训练过程中每个多少训练轮数保存一次模型Checkpoint文件。
	数据预处理并发个数	定义了在预处理数据时，能够同时处理文件的并行进程数量。设定这个参数的主要目的是通过并发处理来加速数据预处理，从而提升训练效率。
	序列长度	sequence_length，训练单条数据的最大长度，超过该长度的数据在训练时将被截断。
训练数据配置	训练集	选择训练模型所需的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
故障自动重启	自动重启	该功能开启后，当训练作业发生故障终止本次作业时，自动从本次中断的作业继续训练。参见**故障自动重启开关配置**。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置预训练相关参数。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建 NLP 大模型增量预训练任务

在模型完成**创建NLP大模型预训练任务**预训练后，可以对训练后的模型继续训练，该过程称为“增量预训练”。

创建NLP大模型增量预训练任务前，请确保有已完成预训练的NLP大模型。

创建NLP大模型增量预训练任务的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-2 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“我的资产 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 其余参数配置等步骤同**创建NLP大模型预训练任务**。

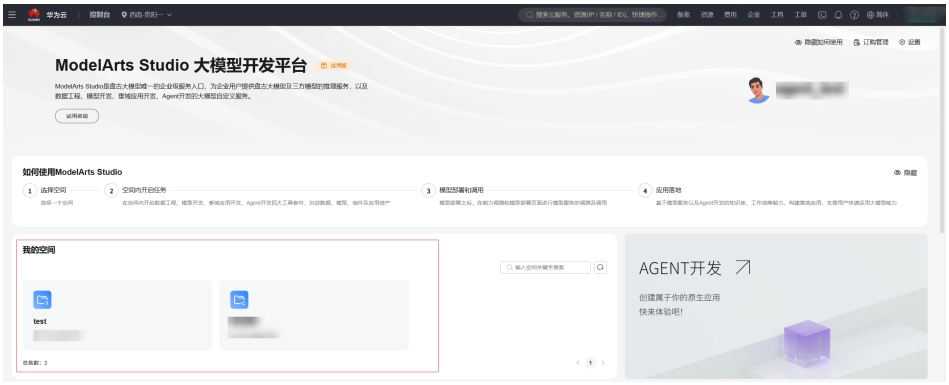
创建 NLP 大模型全量微调任务

创建NLP大模型全量微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-3 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表5-9**完成训练参数设置，参数默认值在创建训练任务的时候会带出。

表 5-9 NLP 大模型全量微调参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制，详细配置参见**断点续训配置**。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	序列长度	sequence_length，训练单条数据的最大长度，超过该长度的数据在训练时将被截断。
	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	模型保存策略	save_checkpoint_steps/save_checkpoint_epoch，训练过程中是按迭代步数，还是训练轮数保存Checkpoint文件。
	Checkpoint保存间隔	save_checkpoint_steps，训练过程中每隔多少个训练步长保存一次模型Checkpoint文件。
	Checkpoint保存轮数	save_checkpoint_epoch，训练过程中每个多少训练轮数保存一次模型Checkpoint文件。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。
训练数据配置	训练集	选择训练模型所需的数据集。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	验证集	• 若选择“分割训练集”，则需进一步配置数据拆分比例。 • 若选择“选择数据集”，则需选择导入的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练单元数。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
故障自动重启	自动重启	该功能开启后，当训练作业发生故障终止本次作业时，自动从本次中断的作业继续训练。参见**故障自动重启开关配置**。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

1. 不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置全量微调相关参数。

5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

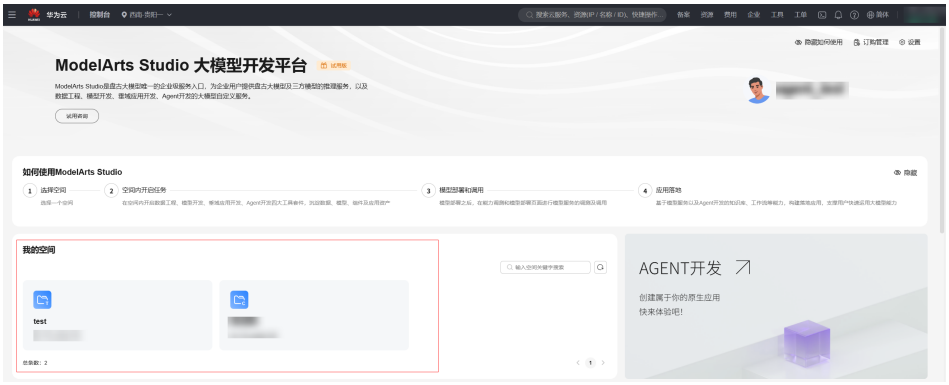
创建 NLP 大模型 LoRA 微调任务

创建NLP大模型LoRA微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表5-10**完成训练参数设置，参数默认值在创建训练任务的时候会带出。

表 5-10 NLP 大模型 LoRA 微调参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“LoRA微调”。 ● LoRA微调：在模型微调过程中，只对特定的层或模块的参数进行更新，而其余参数保持冻结状态。这种方法可以显著减少计算资源和时间消耗，同时在很多情况下，依然能够保持较好的模型性能。
训练参数	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	序列长度	sequence_length，训练单条数据的最大长度，超过该长度的数据在训练时将被截断。
	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	LoRA矩阵中的秩	lora_rank，在Lora矩阵中，Rank的值用于衡量矩阵的复杂度和信息量。数值较大，增强模型的表示能力，但会增加训练时长；数值越小可以减少参数数量，降低过拟合风险。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
训练数据配置	训练集	选择训练模型所需的数据集。
	验证集	● 若选择“分割训练集”，则需进一步配置数据拆分比例。 ● 若选择“选择数据集”，则需选择导入的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练单元数。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置LORA微调相关参数。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

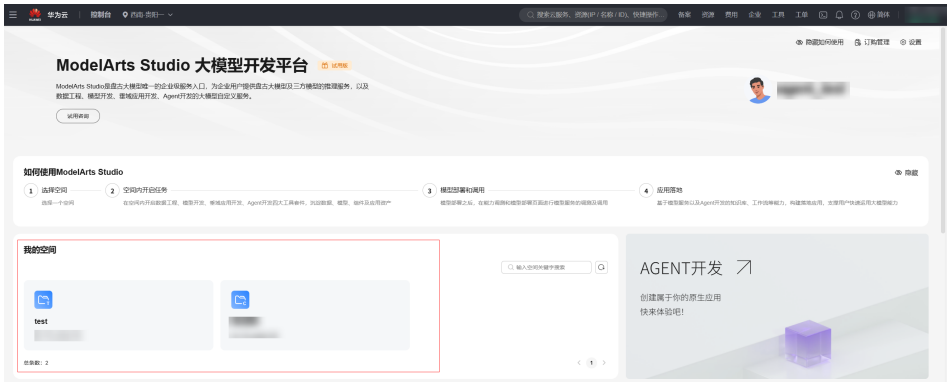
创建 NLP 大模型 DPO 强化学习任务

DPO强化学习基于偏好标注数据，针对微调后的模型继续进行强化训练，使模型输出更符合人类偏好。

创建NLP大模型DPO强化学习任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-5 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“我的资产 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表5-11完成训练参数设置，参数默认值在创建训练任务的时候会带出。

版权所有 © 华为云计算技术有限公司

表 5-11 NLP 大模型 DPO 强化学习参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“我的资产”。 ● 类型：选择“大语言模型”，并选择训练所用的模型。
	训练类型	选择“强化学习”。
	训练目标	选择“DPO”。
	高级设置checkpoints	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。详细配置参见**断点续训配置**。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	数据批量大小	数据集进行分批读取训练，设定每个批次数据的大小。 通常情况下，较大的数据批量可以使梯度更加稳定，从而有利于模型的收敛。然而，较大的数据批量也会占用更多的显存资源，这可能导致显存不足，并且会延长每次训练的时长。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	序列长度	sequence_length，训练单条数据的最大长度，超过该长度的数据在训练时将被截断。
	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	权重衰减系数	weight_decay，是一种对模型参数值大小进行衰减的正则化方法，防止模型过拟合，提高模型泛化能力。
	模型保存策略	save_checkpoint_steps，训练过程中是按迭代步数，还是训练轮数保存Checkpoint文件。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	Checkpoint保存间隔	save_checkpoint_steps，训练过程中每隔多少个训练步长保存一次模型Checkpoint文件。
	Checkpoint保存轮数	save_checkpoint_epoch，训练过程中每隔多少训练轮数保存一次模型Checkpoint文件，当值为0以save_checkpoint_steps为准，当值大于0以save_checkpoint_epoch为准。
	验证步数	eval_steps，模型每隔多少步跑一次验证集。
	旋转位置编码	rotary_base，位置编码的基底值，增强模型对序列中位置信息的捕捉能力，数值越大，模型能够处理的序列长度更长，泛化能力更好，建议使用默认值。
	DPO loss温度超参	Beta，用于控制模型输出分布的集中程度。较高的beta值会使输出更具确定性，而较低的beta值则使输出更具多样性。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	模型保存步数	每训练一定数量的步骤（或批次），模型的状态将会被保存。可以通过以下公式预估已训练的数据量： token_num = step * batch_size * sequence ● token_num：已训练的数据量（以Token为单位）。 ● step：已完成的训练步数。 ● batch_size：每个训练步骤中使用的样本数量。 ● sequence：每个数据样本中的Token数量。
训练数据配置	训练集	选择训练模型所需的数据集。
	验证集	● 若选择“选择数据集”，则需进一步配置数据拆分比例。 ● 若选择“分割训练集”，则需选择导入的数据集。
资源配置	计费模式	选择训练RFT强化任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置DPO强化学习关参数。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

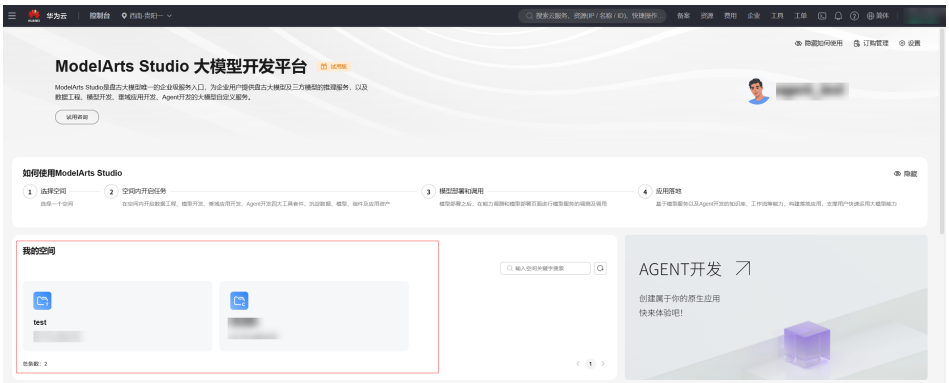
创建 NLP 大模型 RFT 强化学习任务

RFT强化学习通过强化学习的方式优化模型，使其能够在数据量较少的情况下，完成专业领域内复杂任务。

创建NLP大模型RFT强化学习任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-6 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表5-12完成训练参数设置。

 版权所有 © 华为云计算技术有限公司

表 5-12 NLP 大模型 RFT 强化学习参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“强化学习”。
	训练目标	选择“RFT”。
训练参数	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	数据批量大小	指定每个数据并行下处理的数据批量大小。在数据并行和流水线并行开启情况下，全局batch_size等于per_batch_size乘micro_size乘data_parallelism。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	完成全部训练数据集训练的次数。
	优化器	训练中对模型参数进行更新的算法，默认为adamw。
	权重衰减系数	权重衰减系数，通过在损失函数中增加一个与模型权重大小相关的惩罚项，来鼓励模型保持权重较小，从而防止模型过于复杂或过拟合训练数据。
	DPO loss温度超参	用于控制模型输出分布的集中程度。较高的beta值会使输出更具确定性，而较低的beta值则使输出更具多样性。
	打分器类型	rft训练中的样本打分器。
训练数据配置	训练集	选择训练模型所需的数据集。
资源配置	计费模式	选择训练RFT强化任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。

 版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置RFT强化学习关参数。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

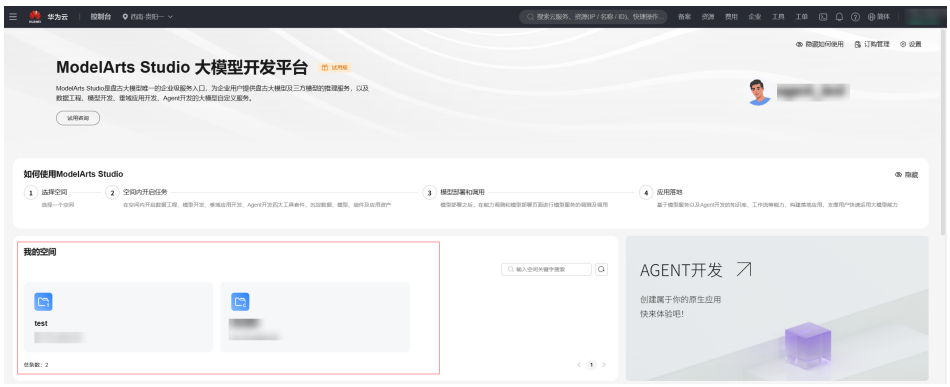
创建 NLP 大模型 GRPO 强化学习任务

GRPO强化学习通过强化学习的方式优化模型，使其能够在数据量较少的情况下，完成专业领域内复杂任务。

创建NLP大模型GRPO强化学习任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-7 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。

版权所有 © 华为云计算技术有限公司

3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表 NLP大模型GRPO强化学习参数说明**完成训练参数设置。

表 5-13 NLP 大模型 GRPO 强化学习参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“强化学习”。
	训练目标	选择“GRPO”。
训练参数	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	数据批量大小	指定每个数据并行下处理的数据批量大小。在数据并行和流水线并行开启情况下，全局batch_size等于per_batch_size乘micro_size乘data_parallelism。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	完成全部训练数据集训练的次数。
	序列长度	sequence_length，训练单条数据的最大长度，超过该长度的数据在训练时将被截断
	权重衰减系数	权重衰减系数，通过在损失函数中增加一个与模型权重大小相关的惩罚项，来鼓励模型保持权重较小，从而防止模型过于复杂或过拟合训练数据。
	单个问题的推理回答数量	在GRPO算法中，模型会对同一个问题进行多次推理，该参数控制单个问题的推理回答数量。
	单次推理的问题数量	模型在推理阶段，单次推理的prompt总数，需能被推理的data_parallel数整除。最大可设置为batch_size值。

 版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	温度	控制模型推理N条数据的随机性，当temperature趋近于0时，选择最大概率的词；temperature越大，分布越平缓，选择更随机。
	核采样	控制模型推理的多样性，取值越大，生成文本的多样性越强。
	排序K采样	控制保留概率最高的前k个候选词，并在这k个词中按概率分布进行采样。
	最大输入长度	单条问题输入（prompt）的最大长度，单位为token。最大为sequence_length-1k。
	初始KL惩罚系数	在强化学习中显著策略更新的幅度，防止过度优化局部最优。若策略更新幅度过大（如回报突增但KL散度显著上升），可逐步增大该值以约束更新。
训练数据配置	训练集	选择训练模型所需的数据集。
资源配置	计费模式	选择训练GRPO强化任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同NLP模型对应“训练参数”可能不同，请根据实际模型对应的“训练参数”配置GRPO强化学习关参数。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态

 版权所有 © 华为云计算技术有限公司

断点续训配置

断点续训练是指因为某些原因（例如容错重启、资源抢占、作业卡死等）导致训练作业还未完成就被中断，下一次训练可以在上一次的训练基础上继续进行。这种方式对于需要长时间训练的模型而言比较友好。

断点续训练是通过checkpoint机制实现。

checkpoint的机制：在模型训练的过程中，不断地保存训练结果（包括但不限于EPOCH、模型权重、优化器状态、调度器状态）。即便模型训练中断，也可以基于checkpoint接续训练。

当需要从训练中断的位置接续训练，只需要加载checkpoint，并用checkpoint信息初始化训练状态即可。

图 5-8 checkpoint 续训配置



模型训练过程中点开训练详情页面，会展示所有保存的checkpoint信息，选择需要恢复的checkpoint，可以看到三种续训方式：

1. 从所选checkpoint创建新训练；
2. 从所选checkpoint继续训练；
3. 从所选checkpoints跳过step继续训练。

三者的区别如下：

版权所有 © 华为云计算技术有限公司

表 5-14 checkpoint 续训方式清单

序号	续训方式	说明
1	从所选checkpoint创建新训练	创建新的训练任务，可以重新选择数据集，只会使用checkpoint中权重的信息，其余的如优化器状态、调度器状态不会读取。
2	从所选checkpoint继续训练	继续原有的任务，会读取模型权重、优化器状态、调度器状态并用于初始化，主要解决由于硬件、网络相关的故障导致训练的中断。
3	从所选checkpoint跳过step继续训练	继续原有的任务，会读取模型权重、优化器状态、调度器状态并用于初始化，但会跳过该checkpoint之后的部分数据（用户可配置），用于解决由于数据质量差导致的loss不收敛问题。

说明

训练详情页的checkpoint输出发布成资产之后，与正常训练结束的模型产物一致，可以支持增量微调，LoRA微调，部署等功能。

故障自动重启开关配置

当训练任务失败且无法及时重启训练时，会导致训练环境闲置，降低训练环境使用效率和训练成功率。为解决此类问题，增加了故障自动重启开关配置。

故障自动重启是指当训练作业失败时，不管什么原因系统都会自动重启训练作业，提高训练成功率和提升作业的稳定性。

开启故障自动重启见图5-9，为避免程序反复无效重启浪费算力资源，建议重启次数设置为3次。

图 5-9 开启故障自动重启



说明

开启故障自动重启开关后，只要系统检测到训练异常，就会无条件重启训练作业。为避免反复无效重启导致的算力浪费，请根据实际使用场景确认是否开启。

5.2.4 查看 NLP 大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、评估指标、训练任务详情和训练日志。

版权所有 © 华为云计算技术有限公司

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表5-15。

表 5-15 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见表5-16。

图 5-10 查看训练指标

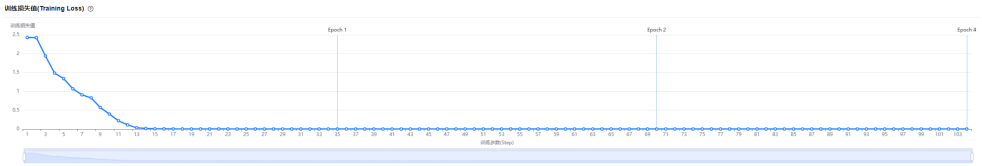


表 5-16 训练指标说明

模型	训练指标	指标说明
NLP大模型	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
NLP大模型	训练精度	强化学习（DPO）训练过程中的指标，表示优选回答（chosen response）与拒绝（rejected response）回答的奖励差值，衡量模型优选回答优于被拒回答的概率。 计算方式：一个训练步的数据中优选回答的奖励分数大于拒绝回答的奖励分数的比例。

版权所有 © 华为云计算技术有限公司

模型	训练指标	指标说明
NLP大模型	训练优选回复的奖励	强化学习（DPO）训练过程中的指标，表示模型对人类标注的优选回答（chosen response）的偏好得分。 计算方式：在一个训练步的数据中，通过模型对两个候选回复的得分差（优选回复得分减去拒绝回复得分）并经Sigmoid函数映射到[0, 1]范围来确定其奖励分数，得分越高表示越符合人类偏好。
NLP大模型	训练拒绝回复的奖励	强化学习（DPO）训练过程中的指标，表示模型对被人类拒绝的回答（rejected response）的偏好得分。 计算方式：在一个训练步的数据中，通过模型对两个候选回复的得分差（拒绝回复得分减去优选回复得分）并经Sigmoid函数映射到[0, 1]范围来确定其奖励分数，得分越低表示越不符合人类偏好。
NLP大模型	验证损失值	强化学习（DPO）训练中运行验证集的指标，表示模型在验证集上的损失值。值越小，意味着模型对验证集数据的泛化能力越好。
NLP大模型	验证精度	强化学习（DPO）训练中运行验证集的指标，表示优选回答（chosen response）与拒绝（rejected response）回答的奖励差值，衡量模型优选回答优于被拒回答的概率。 计算方式：在验证集中优选回答的奖励分数大于拒绝回答的奖励分数的比例。
NLP大模型	验证优选回复的奖励	强化学习（DPO）训练中运行验证集的指标，表示模型对人类标注的优选回答（chosen response）的偏好得分。 计算方式：在验证集中，通过模型对两个候选回复的得分差（优选回复得分减去拒绝回复得分）并经Sigmoid函数映射到[0, 1]范围来确定其奖励分数，得分越高表示越符合人类偏好。
NLP大模型	验证拒绝回复的奖励	强化学习（DPO）训练中运行验证集的指标，表示模型对被人类拒绝的回答（rejected response）的偏好得分。 计算方式：在验证集中，通过模型对两个候选回复的得分差（拒绝回复得分减去优选回复得分）并经Sigmoid函数映射到[0, 1]范围来确定其奖励分数，得分越低表示越不符合人类偏好。

获取训练日志

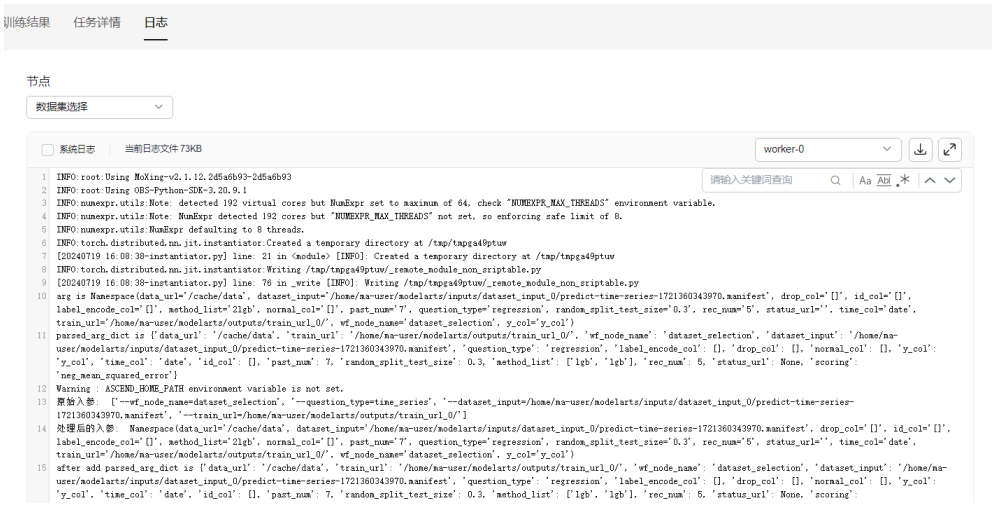
单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**NLP大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

版权所有 © 华为云计算技术有限公司

图 5-11 获取训练日志



训练异常定位方法

模型训练失败时，可以按照如下定位思路对训练任务做初步定位。

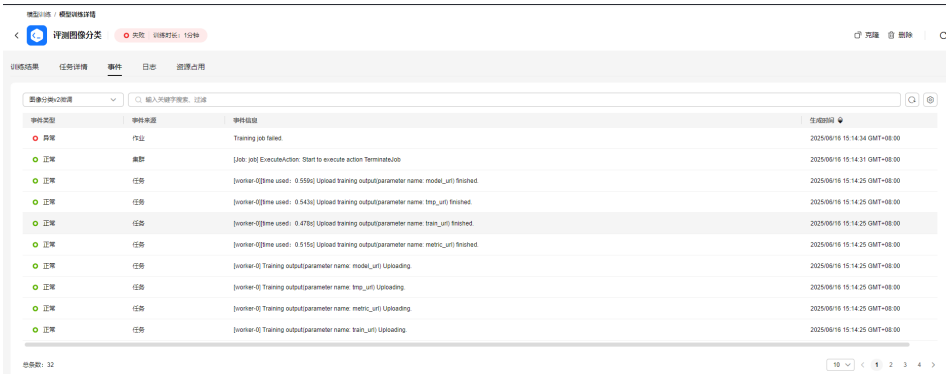
1. 在训练任务列表查看任务失败原因。

   模型启动训练后，可以在模型训练列表中查看训练任务的状态，当任务状态为“失败”时，可将鼠标放置在“失败”上，可以查看到任务失败的概要信息。



2. 单击任务名，进入任务详情页查看训练任务运行关键事件信息。

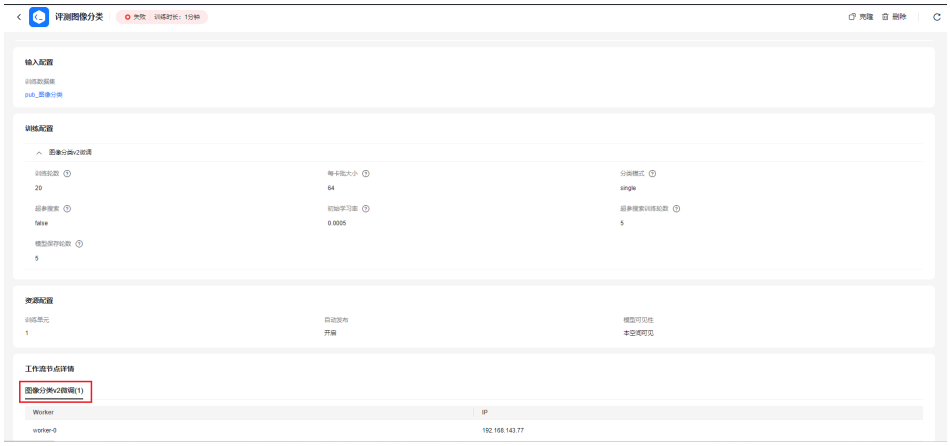
   单击“事件”页签，可查看训练任务子任务运行事件，通过事件可以观察到当前任务的运行情况。事件可以按照“正常、异常、告警”三类事件类型过滤。



3. 获取当前任务运行的worker、workerIP以及任务异常退出时的节点和退出码信息。

   在ModelArts Studio训练任务页面单击需要查看的任务名，进入“任务详情”页，在“工作流节点详情”下可以查看到当前任务所运行的节点及IP信息。

版权所有 © 华为云计算技术有限公司



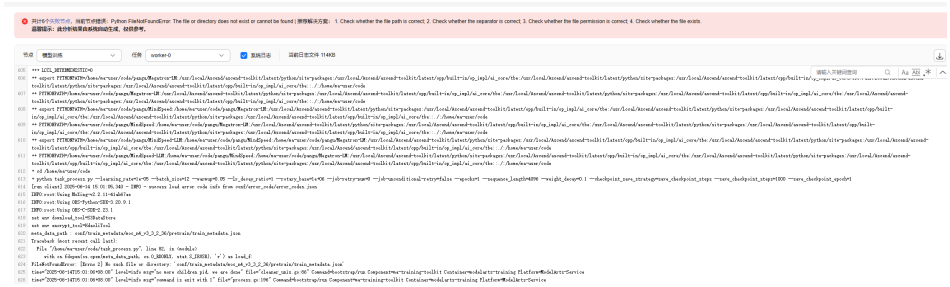
4. 当训练任务运行失败时，可查看“模型训练详情 -> 基本信息 -> 状态”，单击“查看失败节点清单”获取训练任务故障节点及退出码信息。



5. 训练作业遇到问题时，可查看“模型训练详情 -> 日志”，多数场景下的问题可以通过日志报错信息直接定位。

   如果训练作业运行失败，ModelArts Studio会自动识别导致作业失败的原因，在训练日志界面上给出提示并对部分常见训练错误给出分析建议。包含失败的可能原因、推荐的解决方案。

图 5-12 训练错误分析建议



说明

针对分布式作业，只会显示当前节点的一个分析结果，作业的失败需要综合各个节点的失败原因做一个综合判断。

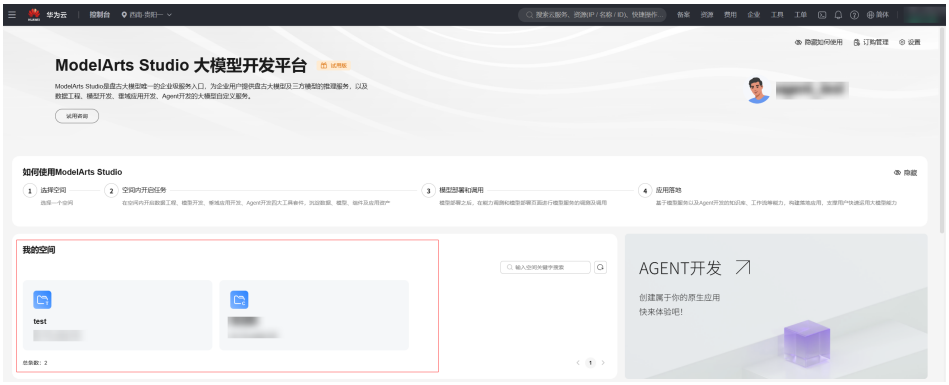
5.2.5 发布训练后的 NLP 大模型

NLP大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-13 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 5-14 训练结果页面



4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

   说明

   1. 通过预训练发布的模型，支持再次进行训练、模型部署。
   2. 通过微调发布的模型，支持再次微调、强化学习及部署。

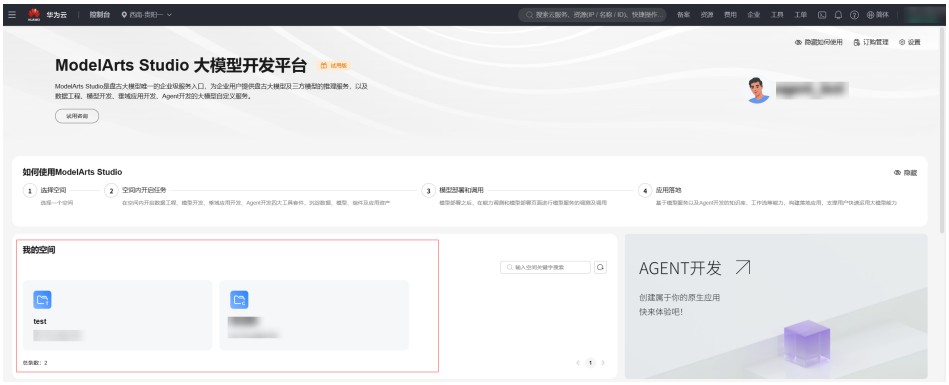
5.2.6 管理 NLP 大模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-15 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 编辑。单击操作列的“编辑”，可以修改模型的checkpoints、训练参数、训练数据以及基本信息等。
   - 克隆。单击操作列的“克隆”，参照**创建NLP大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“删除”，可以删除当前不需要的训练任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

5.2.7 NLP 大模型训练常见报错与解决方案

NLP大模型训练常见报错及解决方案请详见**表5-17**和**表5-18**。

表 5-17 NLP 大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空。	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。
训练日志提示“root: XXX valid number is 0”	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。

版权所有 © 华为云计算技术有限公司

常见报错	问题现象	原因分析	解决方案
训练日志提示“ValueError: label_map not match”	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

表 5-18 NLP 大模型训练常见错误码和建议解决方法

错误码	错误信息	说明	建议解决方法
model.50010001	当前驱动 {0} 不满足指定的驱动版本 {1}。	当前驱动 {0} 不满足指定的驱动版本 {1}。	联系oncall检查宿主机上的驱动，若不符合要求，则需要oncall重新升级驱动到错误信息中指定的版本。
model.50010002	当前固件 {0} 不满足指定的驱动版本 {1}。	当前固件 {0} 不满足指定的驱动版本 {1}。	联系oncall检查宿主机上的固件，若不符合要求，则需要oncall重新升级驱动到错误信息中指定的版本。
model.50010003	当前磁盘空间大小不满足要求。	当前磁盘空间大小不满足要求。	联系oncall扩容宿主机上的磁盘，要求挂载到/cache目录上的空间大于等于19T。
model.50010101	Obs文件下载错误。	Obs文件下载错误。	请检查OBS服务是否正常，如果正常请联系技术支持。
model.50010104	Obs文件上传错误。	Obs文件上传错误。	请检查OBS服务是否正常，如果正常请联系技术支持。
model.50010106	文件删除错误，OBS服务不可用。	文件删除错误，OBS服务不可用。	请检查OBS服务是否正常，如果正常请联系技术支持。

版权所有 © 华为云计算技术有限公司

错误码	错误信息	说明	建议解决方法
model.50010107	文件上传错误，HDFS服务不可用。	文件上传错误，HDFS服务不可用。	请检查HDFS服务是否正常，如果正常请联系技术支持。
model.50010108	文件下载错误，HDFS服务不可用。	文件下载错误，HDFS服务不可用。	请检查HDFS服务是否正常，如果正常请联系技术支持。
model.50010109	文件删除错误，HDFS服务不可用。	文件删除错误，HDFS服务不可用。	请检查HDFS服务是否正常，如果正常请联系技术支持。
model.50010110	列举文件夹下文件错误，HDFS服务不可用。	列举文件夹下文件错误，HDFS服务不可用。	请检查HDFS服务是否正常，如果正常请联系技术支持。

5.3 压缩 NLP 大模型

模型在部署前，通过模型压缩可以降低推理显存占用，节省推理资源提高推理性能。

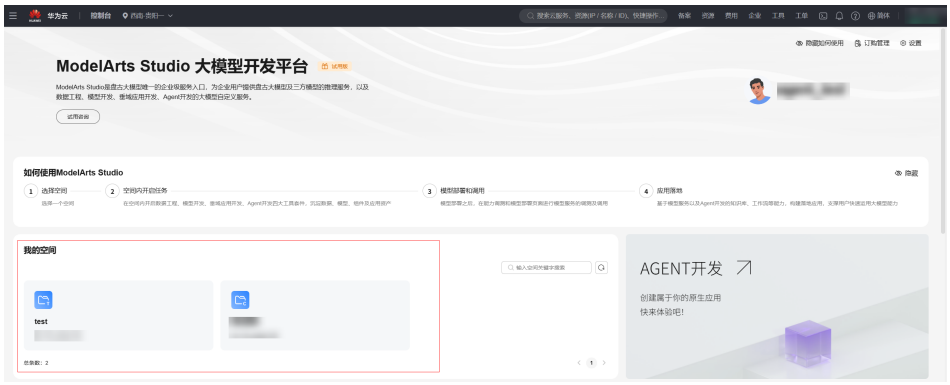
约束限制

平台当前仅可对NLP大模型进行压缩，**支持压缩的模型清单请详见**《产品介绍》>“模型能力与规格 > **盘古NLP大模型**”。

NLP 模型压缩步骤

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-16 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型压缩”，单击界面右上角“创建压缩任务”。

版权所有 © 华为云计算技术有限公司

3. 在“创建压缩任务”页面，选择需要压缩的基础模型，支持选择已发布模型或未发布模型。
4. 选择压缩策略。除INT8压缩策略外，部分模型支持INT4压缩策略，可在选择模型后，根据页面展示的策略进行选择。
   - INT8：该压缩策略将模型参数压缩至8位字节，可以有效降低推理显存占用。
   - INT4：该压缩策略与INT8相比，可以进一步减少模型的存储空间和计算复杂度。
5. 配置资源。选择计费模式并设置训练单元。
6. 可选择开启订阅提醒。开启后，系统将在本次压缩任务状态变更时，向用户发送短信/邮件提醒。
7. 选择开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
8. 填写基本信息，包括任务名称、压缩后模型名称与描述，单击“立即创建”。

   当压缩任务状态为“已完成”时，表示模型已完成压缩操作。

蒸馏 NLP 大模型

1、模型蒸馏介绍：模型蒸馏是一种模型压缩技术，通过训练一个小模型（学生模型）去模仿一个性能更强但体积更大的模型（教师模型）的行为。蒸馏过程包括：先训练好教师模型，然后用其输出作为“软标签”指导学生模型学习，以提升学生模型的泛化能力。该方法能在保持较高性能的同时，显著减少模型参数量与推理开销，广泛应用于边缘计算、移动设备部署等场景。

2、构建蒸馏数据

NLP模型蒸馏数据要求：

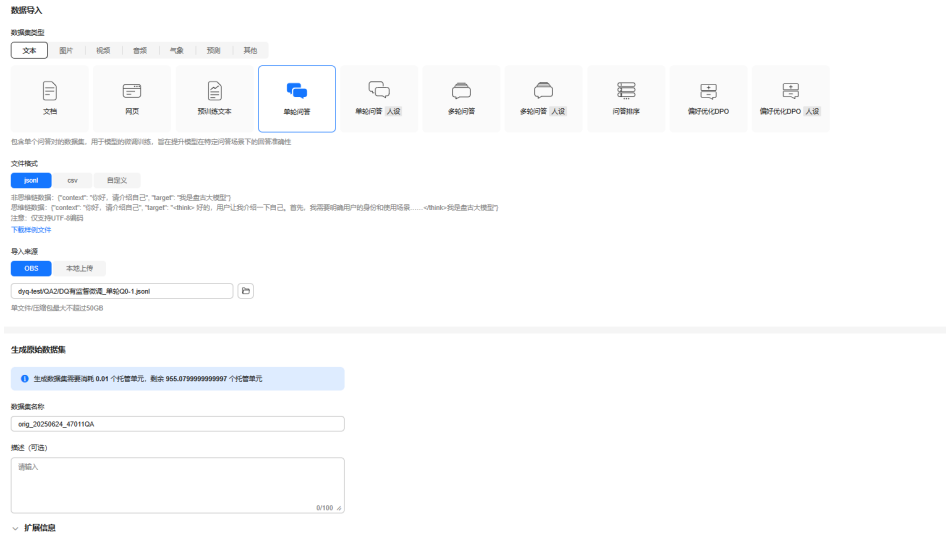
1. **高质量语料**：需使用覆盖面广、语义丰富的文本数据，有助于学生模型学习教师模型的语义表达和语言理解能力。
2. **与任务相关**：蒸馏数据应贴近目标任务（如分类、问答、翻译等），可使用原始训练集、增强数据或无标签文本。
3. **数据量充足**：虽然学生模型规模较小，但蒸馏通常需大量样本以充分吸收教师模型的知识，特别是对大模型蒸馏。
4. **一致性处理**：输入文本格式确保学生模型学习的输入与教师一致。

蒸馏数据获取：

步骤1 在ModelArts Studio中，单击“数据获取”，单击“创建导入数据”，导入有Q无A的数据，如图5-17。

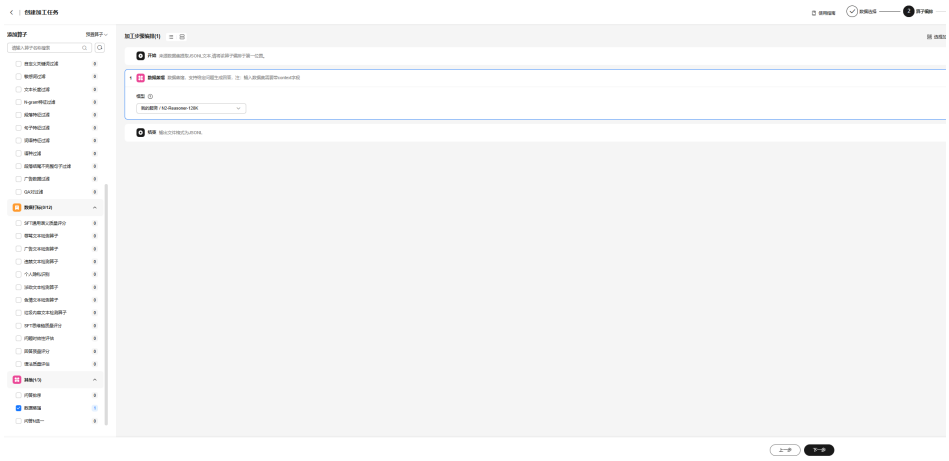
 版权所有 © 华为云计算技术有限公司

图 5-17 导入单轮问答类数据集



步骤2 单击"数据加工 > 创建加工任务”，选中导入的单轮问答类数据集，单击“下一步”，选择预置指令“数据蒸馏”如图5-18，配置模型信息，单击下一步，配置资源池信息，选择自动生成数据集，启动即可，任务执行成功后，获得了蒸馏数据集。

图 5-18 创建数据蒸馏任务



步骤3 单击“数据发布”，将生成的蒸馏数据集发布到模型训练。

----结束

使用蒸馏后的数据进行模型训练：可查看**训练NLP大模型**进行操作。

5.4 部署 NLP 大模型

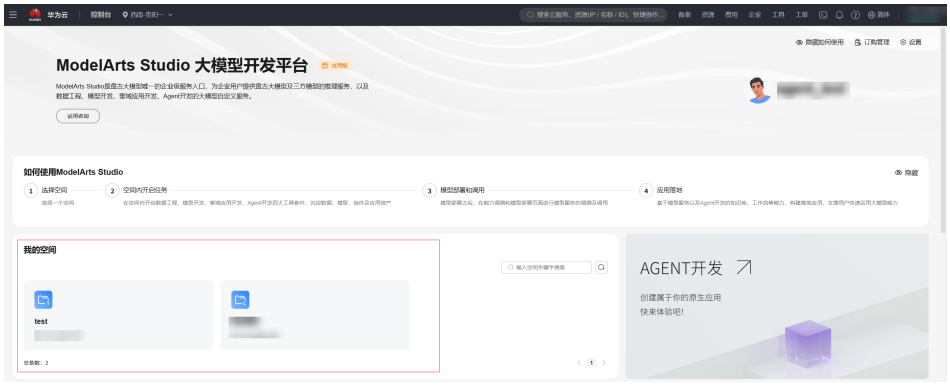
5.4.1 创建 NLP 大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-19 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表5-19**完成部署参数设置。

表 5-19 NLP 大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择需要进行部署的模型和版本。
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。 – 实例数：设置部署模型时所需的实例数。
	最大TOKEN长度	模型处理序列长度最大值。
	自定义名称	此名称是通过V2版本推理接口调用该推理服务时的唯一标识。创建后不支持修改。
安全护栏	安全护栏开关	安全护栏保障模型调用安全。
	版本选择	当前支持安全护栏基础版，内置了默认的内容审核规则。

版权所有 © 华为云计算技术有限公司

参数分类	部署参数	参数说明
资源配置（选择云上部署时）	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

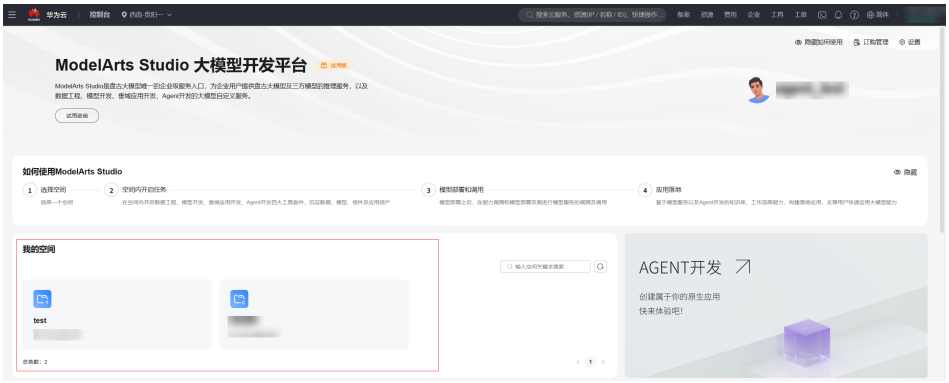
5. 参数填写完成后，单击“立即部署”。

5.4.2 查看 NLP 大模型部署任务详情

部署任务创建成功后，可以查看大模型部署任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-20 我的空间



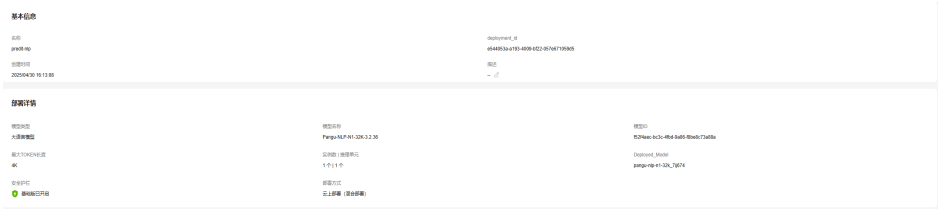
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建NLP大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

版权所有 © 华为云计算技术有限公司

图 5-21 部署详情



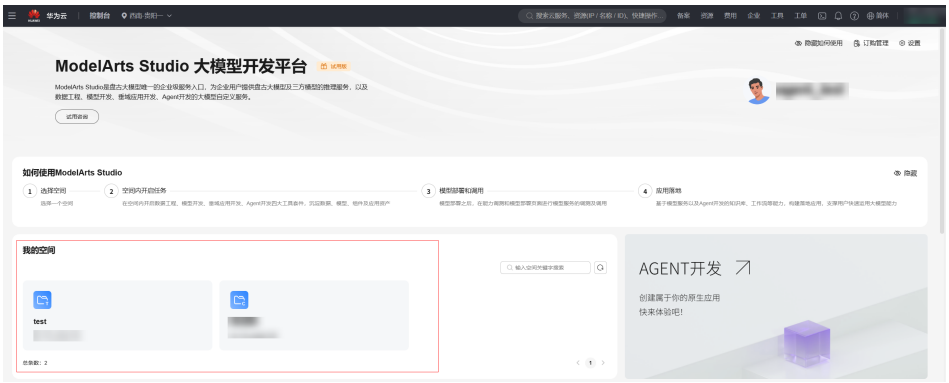
5.4.3 管理 NLP 大模型部署任务

模型更新

完成**创建NLP大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-22 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

版权所有 © 华为云计算技术有限公司

图 5-23 升级模式



说明

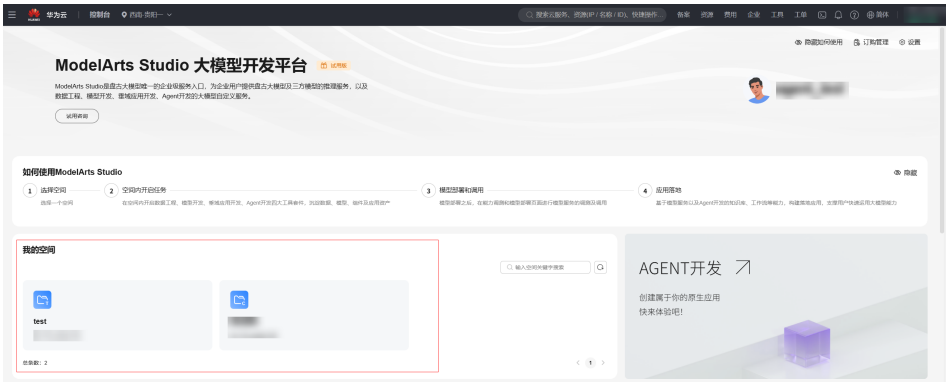
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建NLP大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-24 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

版权所有 © 华为云计算技术有限公司

图 5-25 升级模式



说明

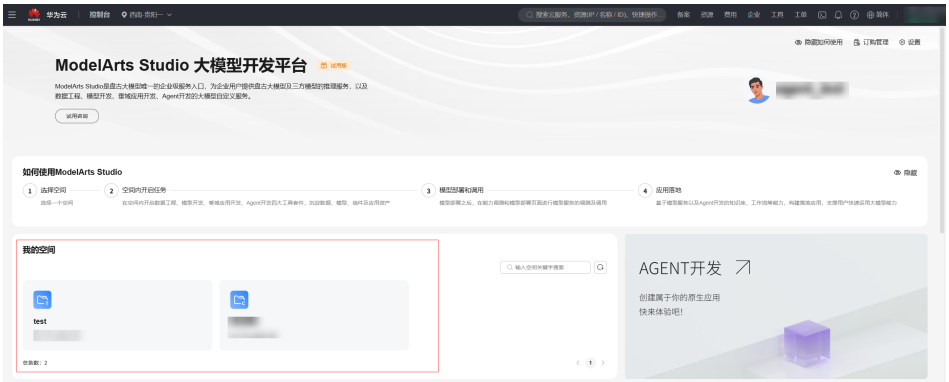
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成**创建NLP大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-26 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

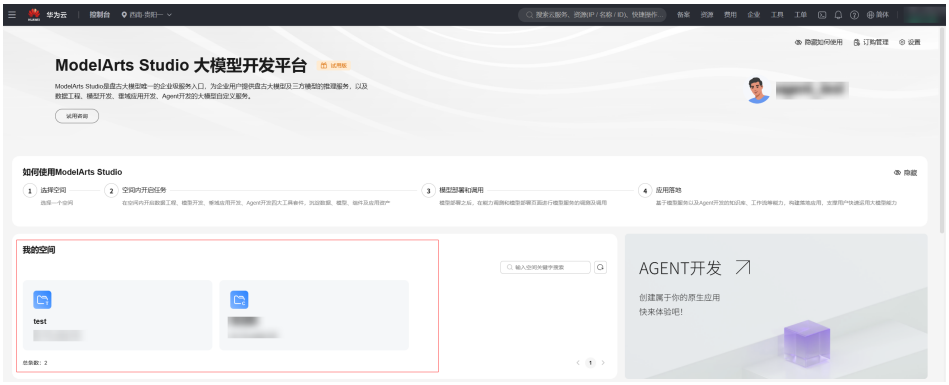
管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-27 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

5.4.4 NLP 大模型推理常见错误及解决方案

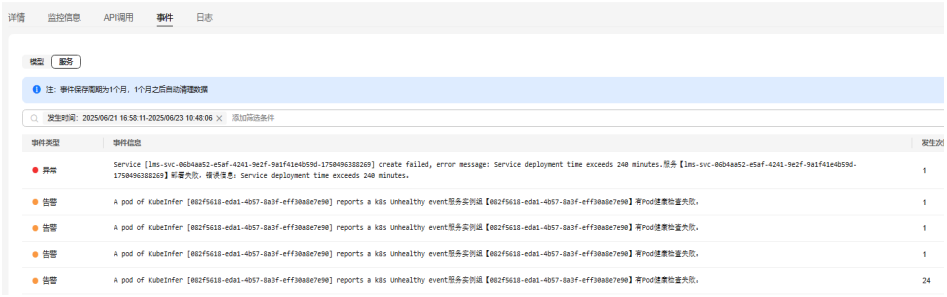
推理日志分析方法

推理部署或者推理过程失败可以查看推理的日志进行原因分析，多数场景下的问题可以通过日志报错信息直接定位。

如果推理服务部署失败或者在运行过程中异常崩溃，ModelArts Studio会自动识别导致推理服务异常的原因，可首先在服务详情-事件-模型|服务中查看对应的异常报错信息。



版权所有 © 华为云计算技术有限公司



更详细的日志，可以查看日志页签，并通过搜索过滤“ERROR”或“Error”定位关键信息。但请注意，不是所有的Error信息都是阻塞性错误，一般寻找异常出现的第一个Error日志进行定位。日志中存在错误码与错误描述信息见**表5-20**。

推理错误码

表 5-20 推理错误清单

错误码	错误原因	解决方案
Model.50020001	当前驱动 {0} 不满足指定的驱动版本 {1}。	请联系技术支持。
Model.50020002	当前固件 {0} 不满足指定的驱动版本 {1}。	请联系技术支持。
Model.50020003	当前磁盘空间大小不满足要求。	请检查容器内磁盘空间大小。
Model.50020004	当前资源（CPU/内存/NPU卡）不满足要求。	请联系技术支持。
Model.50020005	推理部署参数异常{0}。	请联系技术支持。
Model.50020101	文件下载错误，OBS路径{0}。	请检查OBS服务是否正常，OBS路径是否正确，如果正常请联系技术支持。
Model.50020102	文件下载错误，OBS服务不可用。	请检查OBS服务是否正常，OBS路径是否正确，如果正常请联系技术支持。
Model.50020103	文件上传错误，本地路径{0}，OBS路径{1}。	请检查OBS服务是否正常，OBS路径是否正确，如果正常请联系技术支持。
Model.50020104	HDFS文件下载错误，HDFS路径{0}。	请检查HDFS服务是否正常，HDFS路径是否正确，如果正常请联系技术支持。
Model.50020105	文件上传错误，本地路径{0}，HDFS路径{1}。	请检查HDFS服务是否正常，HDFS路径是否正确，如果正常请联系技术支持。
Model.50020201	文件完整性校验失败，文件{0}。	请检查文件是否正常，联系技术支持。

版权所有 © 华为云计算技术有限公司

错误码	错误原因	解决方案
Model.50020202	文件一致性校验失败，文件{0}。	请检查文件是否正常，联系技术支持。
Model.50020203	文件加密失败，文件{0}。	请联系技术支持。
Model.50020204	文件解密失败，文件{0}。	请联系技术支持。
Model.50020205	文件加解密秘钥获取失败。	请联系技术支持。
Model.50020301	模型加载报错。	请联系技术支持。
Model.50020401	推理服务启动失败。	请联系技术支持。
Model.50021001	推理请求url，或类型校验错误。	请检查请求路径，请求类型，如果正常请联系技术支持。
Model.50021002	推理请求参数格式校验错误，应为json格式。	请检查请求参数内容格式，如果正常请联系技术支持。
Model.50021003	推理请求参数内容校验错误，参数{0}。	请检查请求参数内容信息，如果正常请联系技术支持。
Model.50021101	推理请求排队超时错误。	请联系技术支持。
Model.50021102	推理请求生成超时错误。	请联系技术支持。
Model.50021201	推理请求生成过程错误。	请联系技术支持。
Model.50021202	推理请求采样过程错误。	请联系技术支持。
Model.50021203	推理请求结果返回过程错误，流式处理错误。	请联系技术支持。
Model.50021204	推理过程内部错误。	请联系技术支持。

常见推理问题定位方法

常见推理问题定位思路如下：

1. 根据事件及日志界面提示中提供的分析建议解决。
   - 参考案例解决：会提供当前故障对应的指导文档链接，请参照文档中的解决方案修复问题。
   - 事件描述：根据事件问题描述进行排查问题解决。
2. 若果第一步不能解决问题时，可以尝试分析日志中提示的错误信息，定位并解决问题。
3. 如果以上均不能解决问题，可以提工单进行人工咨询。

5.5 评测 NLP 大模型

版权所有 © 华为云计算技术有限公司

5.5.1 创建 NLP 大模型评测集

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同情景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

NLP大模型支持人工评测、自动评测两种评测模式。

- 人工评测：通过人工创建的评测数据集和评测指标项对模型生成的回答进行评测，评测时需要人工基于创建好的评测项对模型回答进行打分，评测完成后会基于打分结果生成评测报告。
- 自动评测：包含“基于规则”与“基于大模型”两种评测规则。
  - 基于规则（相似度/准确率）自动对模型生成的回答进行评测。用户可使用评测模板中预置的专业数据集进行评测，或者自定义评测数据集进行评测。支持的预置数据集如表5-21。
  - 基于大模型，使用大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景，包含评分模式与对比模式。

表 5-21 预置评测集

数据集类别	数据集	说明
通用知识与技能	常识知识	评测模型对于日常生活基本知识和信息的掌握程度，包括历史、地理、文化等多个方面的基础认知。
	数学能力	评测模型解决数学问题的能力，包括算数运算、代数方程求解、几何图形分析能力等。
	逻辑推理	评测模型根据已知信息进行合理推断和分析的能力，涉及演绎，归纳等多种逻辑思维过程。
	中文能力	评测大模型在中文语言和文化背景下的高级知识和推理能力。
领域知识	金融领域	评测大模型在金融领域的能力，包括快速理解和解释复杂的金融概念，风险预测和数据分析，投资建议以及金融决策支持等。

评测集支持从本地上传、OBS 文件和数据工程数据集导入

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“评测中心 > 评测集管理”，单击界面右上角“创建评测集”。
3. 在“创建评测集”页面选择所需要的“模型类型”、“文件格式”、“导入来源”，并单击“选择路径”上传数据文件。本地上传最大10MB，OBS默认最大100M，评测集通过数据工程导入时数据大小约束以数据工程为准，参见表5-22

版权所有 © 华为云计算技术有限公司

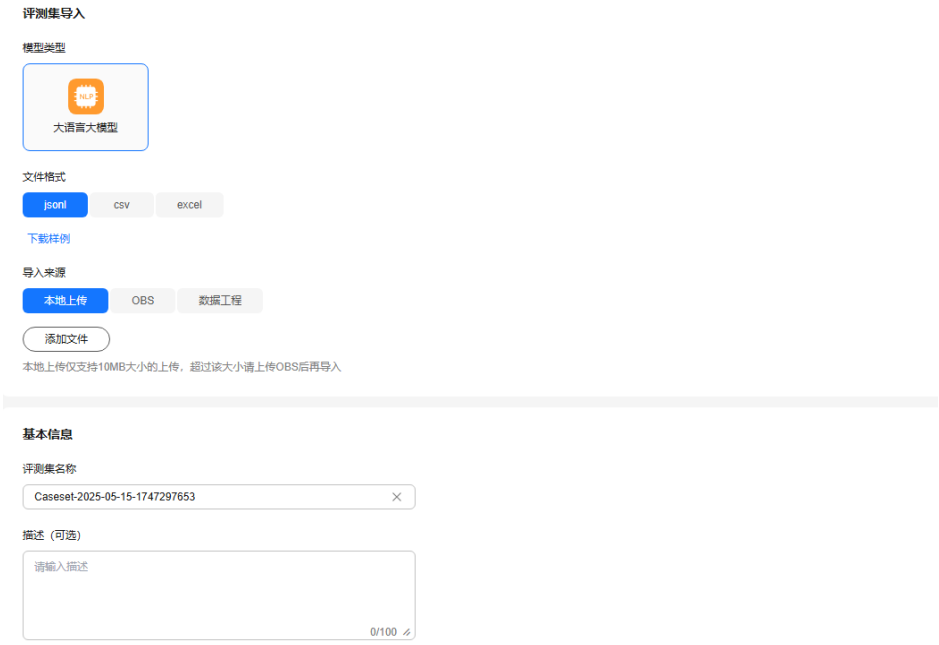


表 5-22 评测数据集参数说明

模型类型	评测数据集格式	大小
NLP大模型	json/csv/xlsx	本地上传最大10M，OBS最大100M

说明

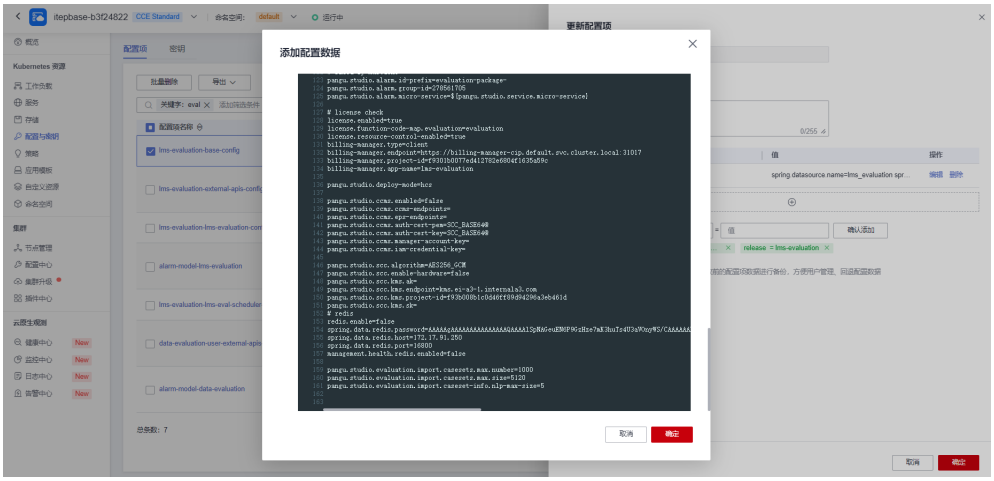
cce评测服务启动配置参数说明如下：

pangu.studio.evaluation.import.casesets.max.number: 所有评测任务数据集最大数量，默认100条，上限与数据集所在的OBS桶容量相关

pangu.studio.evaluation.import.casesets.max.size: 所有评测任务数据集文件总大小上限，默认512MB，上限与数据集所在的OBS桶容量相关

pangu.studio.evaluation.import.caseset-info.nlp-max-size: NLP单次上传最大文件尺寸，默认100MB，上限与数据集所在的OBS桶容量相关

pangu.studio.evaluation.import.caseset-info.nlp-max-data: NLP单次上传最大数据集数量，默认10000条，上限与数据集所在的OBS桶容量相关



版权所有 © 华为云计算技术有限公司

数据工程数据集创建

评测数据集的创建步骤与训练数据集一致，本章节仅做简单介绍，详细步骤请参见**使用数据工程构建NLP大模型数据集**。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面选择所需要的“数据集类型”、“文件格式”、“导入来源”，并单击“存储位置”上传数据文件。

   NLP大模型评测数据集支持的格式见**表5-23**。

   表 5-23 评测数据集格式

模型类型	评测数据集格式
NLP大模型	文本-单轮问答-jsonl格式

4. 上传数据文件后，填写“数据集名称”与“描述”，单击“立即创建”。
5. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
6. 在“创建发布任务”页面选择数据集模态并选择数据集文件。
7. 单击“下一步”，选择发布格式，填写名称，选择数据集可见性，单击“确定”。

   说明

   如果评测盘古大模型， 需要在发布数据集时，将数据集格式发布为“盘古格式”。

5.5.2 创建 API 服务（可选）

API管理模块可以对评测服务的API进行管理，在创建评测任务的时候选择评测服务即可。可以对API进行创建、编辑、删除。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

   图 5-28 我的空间



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。
3. 定义API名称、组织机构，并选择模型类型，单击“提交”。

版权所有 © 华为云计算技术有限公司

图 5-29 创建 API



4. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。
5. 单击“新增版本”，并配置版本、并发数、QPM、URL、Param、Body、Response。
   a. 版本：自定义。
   b. 并发数：可以根据模型的性能进行配置，范围1~32。
   c. QPM：可以根据模型的性能进行配置，范围1~5000。
   d. URL：模型推理地址。
   e. Param：用户可以根据接口需要配置，可选，可以定义header和请求参数信息。
   f. Body(自定义请求体)：推理请求体，可以自定义也可以选择预置的openai或者tgi结构。
   g. Response(响应体)：通过jsonpath的形式定义怎么获取响应体里面的内容。jsonpath语法的作用是从响应体的json字段中提取出所需的数据。

 版权所有 © 华为云计算技术有限公司

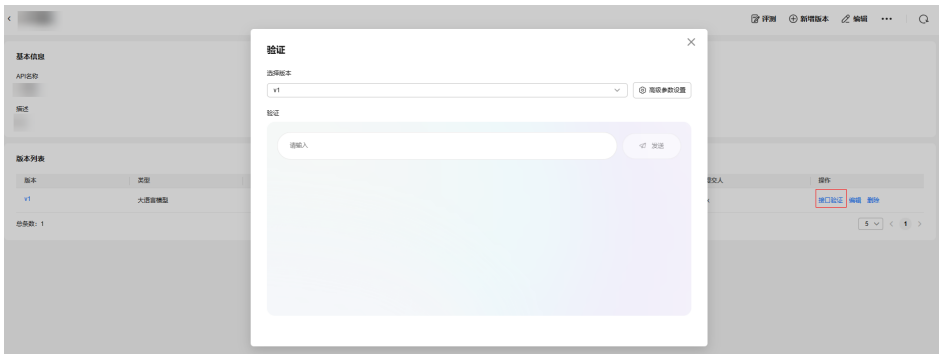
图 5-30 新增版本



6. 创建API版本后，需要验证API是否可用，只有验证通过的API才可以进行评测。在“API管理”页面单击创建好的API任务名称，单击“接口验证”输入问题后查看接口是否有响应内容输出。

 版权所有 © 华为云计算技术有限公司

图 5-31 接口验证



说明

验证API调用如果出现API调用失败，可能原因是：

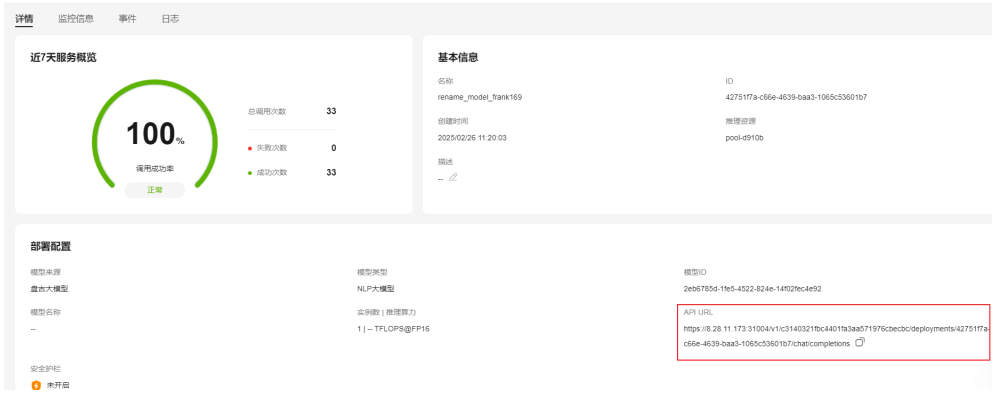
1. 接口配置错误，需要排查URL、请求参数等。
2. 如果配置没问题，可以试着输入简单的问题，比如你好，如果接口能响应，但是输入复杂问题时，接口失败，可以将max-token这类调整token大小的参数配置改小，比如30，这样模型响应就会更快。

创建 API 服务示例

本示例中使用已经部署，且正常运行的盘古NLP模型创建API服务。

1. 在“模型开发 > 模型部署”页面，找到已部署的NLP模型，单击服务名称，在详情页复制API URL。

图 5-32 获取 API URL



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。定义API名称、组织结构，模型类型选择NLP大模型。
3. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。单击“新增版本”，参考表5-24完成配置。

表 5-24 配置 API 服务

参数	填写说明
版本	可自定义。

版权所有 © 华为云计算技术有限公司

参数	填写说明
并发数	设置为1。
QPM	设置为1。
URL	填写获取到的API URL。
Param	共填写两个参数，均设置为header类型。 参数1：参数名X-AUTH-TOKEN，参数值为Token值，参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。 参数2：参数名Content-Type，参数值application/json。
Body	选择自定义，填写内容： {"model": "model","messages": [{"role": "user","content": "{{message}}"}],"max_tokens": 1024, "temperature": 0.7,"top_p": 0.95,"presence_penalty": 0.5,"frequency_penalty": 0.5,"stream": false}
Response	填写： $.choices[0].message.content

版权所有 © 华为云计算技术有限公司

图 5-33 配置 API 服务



5.5.3 创建 NLP 大模型评测任务

创建NLP大模型评测任务前，请确保已完成**创建NLP大模型评测集**操作。

预训练的NLP大模型不支持评测。

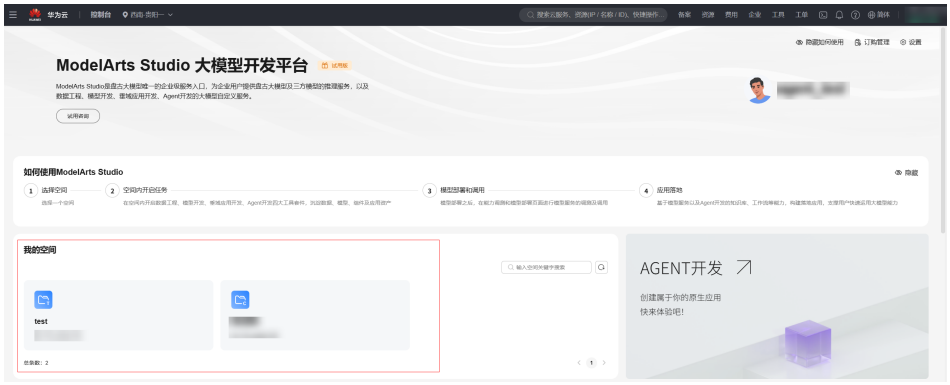
版权所有 © 华为云计算技术有限公司

创建 NLP 大模型自动评测任务（基于规则）

创建NLP大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-34 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考表5-25完成部署参数设置。

表 5-25 NLP 大模型自动评测任务参数说明（基于规则）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测规则	选择“基于规则”：基于规则自动打分，即基于相似度/准确率进行打分，对比模型预测结果与标注数据的差异，适合标准选择题或简单问答场景。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。 • 自定义评测集：由用户指定评测指标（F1分数、准确率、BLEU、Rouge）并上传评测数据集进行评测。选择“自定义评测集”时需要上传待评测数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

创建 NLP 大模型自动评测任务（基于大模型）

创建NLP大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-35 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考表5-26完成部署参数设置。

版权所有 © 华为云计算技术有限公司

表 5-26 NLP 大模型自动评测任务参数说明（基于大模型）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置（基于大模型）	评测规则	选择“基于大模型”。
	选择模式	● 评分模式：每个模型独立评测。 ● 对比模式：对比两个模型的评测效果，报告上显示的是对比结果。
	评测数据集	选择模型评测所需的数据集。
	评测结果存储位置	模型评测结果的存储位置。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
裁判员配置	裁判模型	● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
	打分规则	打分prompt模板可以选择预置，也可以选择自定义。创建自定义prompt模板需要单击“添加自定义规则->新建”，根据页面输入名称，人设，任务描述，是否包含问题，是否包含参考答案，打分策略，评测指标以及指标，最终单击“保存模板”。
基本信息	评测任务名称	填写评测任务名称。
	描述（可选）	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

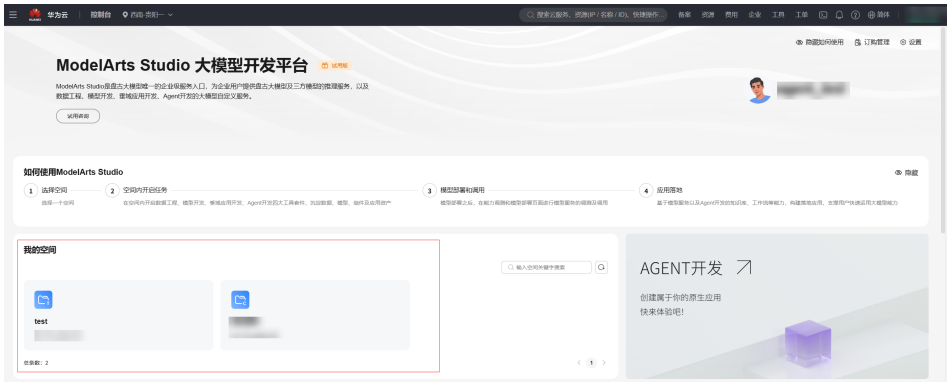
创建 NLP 大模型人工评测任务

创建NLP大模型人工评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 5-36 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建人工评测任务”。
3. 在“创建人工评测任务”页面，参考表5-27完成部署参数设置。

表 5-27 NLP 大模型人工评测任务参数说明

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测指标	由用户自定义评测指标并填写评测标准。
	评测数据集	待评测的数据集。
	评测结果存储位置	模型评测结果的存储位置。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 人工评测”页面。
5. 当状态为“待评测”时，可以单击操作列“在线评测”进入评测页面。
6. 依据页面提示完成评测，全部数据评测完成后单击“提交”。
   - 评测详情页面，单击“盲测”，会隐藏模型名称，进行盲评。
   - 单击“存疑”或者“作废”进行用例的存疑或作废，若取消存疑或作废，单击“取消存疑”或“取消作废”进行处理。
   - 单击备注下方“单击以添加备注”，可以进行新增备注。
   - 评测页面，长按鼠标左键选中需要标记的文本内容，单击“标记”可以标记成重点内容。

   图 5-37 人工评测



7. 返回“评测中心 > 评测任务 > 人工评测”页面，单击操作列“评测报告”查看模型评测结果。

   评测完成之后，进入人工评测列表页面，单击“人工复核”，进行复核评测，复核完成之后，单击“提交”，提交评估结果。

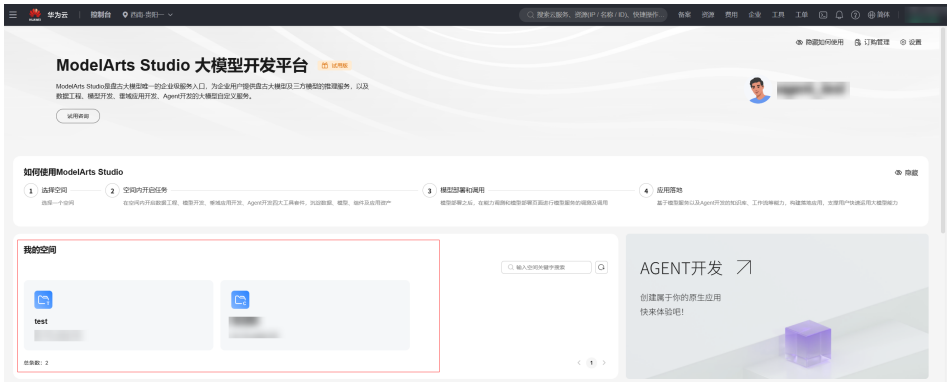
5.5.4 查看 NLP 大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 5-38 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**NLP大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 服务结果分析”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

NLP 大模型评测指标说明

NLP大模型支持自动评测与人工评测，各指标说明如**表5-28**、**表5-29**、**表5-30**。

表 5-28 NLP 大模型自动评测指标说明-不使用预置评测集

评测指标（自动评测-自定义评测集）	指标说明
准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
F1分数	精确率和召回率的调和平均数，分数越高表示模型在这两个指标上表现越好，即模型在精确率和召回率之间取得了更好的平衡。
BLEU-1	模型生成句子与实际句子在单字层面的匹配度，分数越高，表示模型效果越好。
BLEU-2	模型生成句子与实际句子在中词组层面的匹配度，分数越高，表示模型效果越好。
BLEU-4	模型生成结果和实际句子的加权平均精确率，分数越高，表示模型效果越好。
ROUGE-1	将模型生成结果和标注结果按1-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。

版权所有 © 华为云计算技术有限公司

评测指标（自动评测-自定义评测集）	指标说明
ROUGE-2	将模型生成结果和标注结果按2-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。
ROUGE-L	将模型生成结果和标注结果按最长公共子序列（longest-gram）拆分后，计算出的召回率，分数越高，表示模型效果越好。

表 5-29 NLP 大模型自动评测指标说明-使用预置评测集

评测指标（自动评测-使用评测模板）	指标说明
评测得分	每个数据集上的得分为模型在当前数据集上的通过率；评测能力项中若有多个数据集则按照数据量的大小计算通过率的加权平均数。
综合能力	综合能力是计算所有数据集通过率的加权平均数。

表 5-30 NLP 大模型人工评测指标说明

评测指标（人工评测）	指标说明
准确性	模型生成答案正确且无事实性错误。
average	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
goodcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
badcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分以下的占比。
用户自定义的指标	由用户定义的指标，如有用性、逻辑性、安全性等。

表 5-31 NLP 大模型自动评测指标说明-基于大模型

模型类型	评测指标（自动评测-基于规则-基于大模型）	指标说明
NLP大模型	裁判员模型打分	数据集中每个用例，裁判员模型给的评分值。
	平均值	数据集中所有用例得分的平均值。

版权所有 © 华为云计算技术有限公司

模型类型	评测指标（自动评测-基于规则-基于大模型）	指标说明
	中位数	数据集中所有用例得分的中位数。
	标准差	数据集中所有用例得分的标准差。
	win	对比模式下，对比模型win基准模型的个数。
	lose	对比模式下，对比模型lose基准模型的个数。
	tie	对比模式下，对比模型tie基准模型的个数。
	分位	(win+tie)/(lose+tie)
	分位（剔除tie_bad）	剔除tie_bad所得分位，(win+tie_good)/(lose+tie_good)
	分位（剔除tie_good）	剔除tie_good所得分位，(win+tie_bad)/(lose+tie_bad)

5.5.5 管理 NLP 大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-39 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。

版权所有 © 华为云计算技术有限公司

- 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

  说明

  删除属于高危操作，删除前请确保当前任务不再需要。

5.6 调用 NLP 大模型

5.6.1 使用“体验中心”调用 NLP 大模型

体验中心功能支持用户调用预置或训练后的NLP大模型。使用该功能前，请完成模型的部署操作，步骤详见**创建NLP大模型部署任务**。

使用“体验中心”调用NLP大模型可实现文本对话能力，即在输入框中输入问题，模型将基于问题输出相应的回答，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-40 我的空间



2. 在左侧导航栏中选择“体验中心”，单击“文本对话”页签。
3. 选择需要调用的服务。可从“预置服务”或“我的服务”中选择。
4. 填写系统人设。如“你是一个AI助手”，若不填写，将使用系统默认人设。
5. 在页面右侧配置参数，具体参数说明见**表5-32**。

表 5-32 NLP 大模型体验中心参数说明

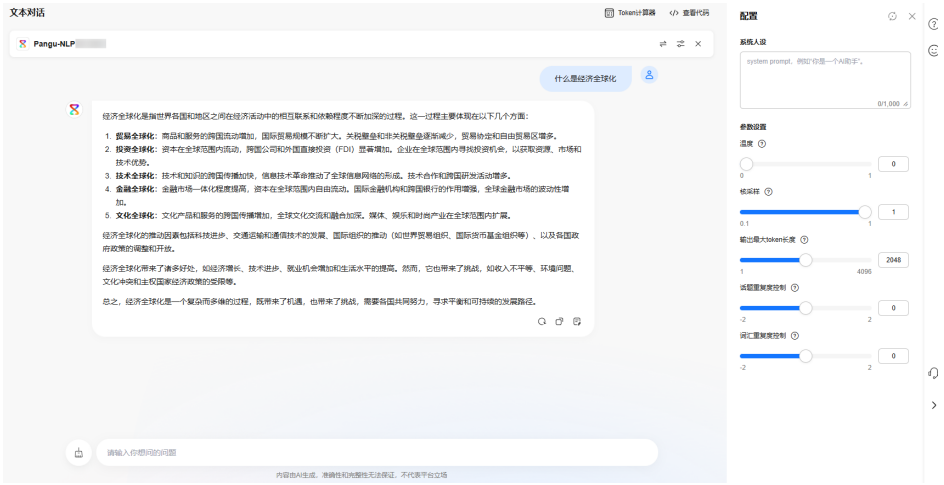
参数	说明
温度	用于控制生成文本的多样性和创造力。调高温度会使得模型的输出更多样性和创新性。
核采样	控制生成文本多样性和质量。调高核采样可以使输出结果更加多样化。
最大口令限制	用于控制聊天回复的长度和质量。

版权所有 © 华为云计算技术有限公司

参数	说明
话题重复度控制	用于控制生成文本中的重复程度。调高参数模型会更频繁地切换话题，从而避免生成重复内容。
词汇重复度控制	用于调整模型对频繁出现的词汇的处理方式。调高参数会使模型减少相同词汇的重复使用，促使模型使用更多样化的词汇进行表达。
输出最大TOKEN长度	用于控制模型生成回复的长度，一般来说，较大的max_tokens值可以生成较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较简洁的回复，但也可能导致生成不完整或不连贯的内容，请根据场景选择合适的值。

6. 如图5-41，输入对话，单击“生成”，模型将输出相应的回答。

图 5-41 调测 NLP 大模型



5.6.2 使用 API 调用 NLP 大模型

预置模型或训练后的模型部署成功后，可以使用“文本对话”API实现模型调用。NLP推理服务支持使用盘古推理接口（V1推理接口）调用，也支持使用业界通用的OpenAi格式接口（V2推理接口）调用。V1接口、V2接口的鉴权方式不同，请求体和返回体略有差异。

表 5-33 NLP 大模型推理 API 清单

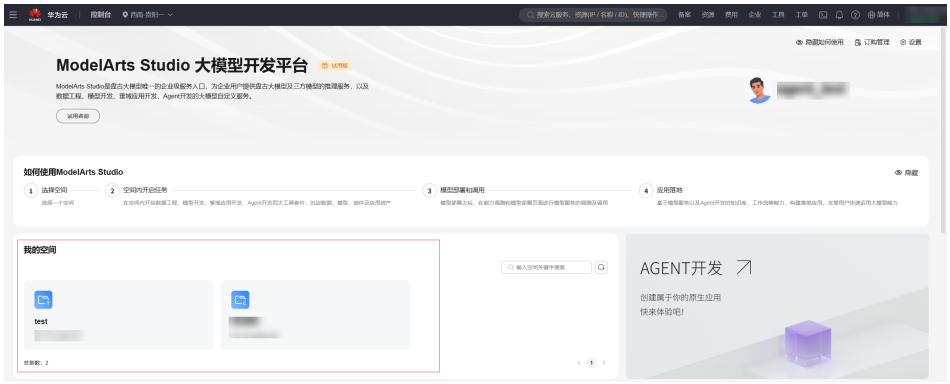
API分类	API访问路径（URI）
V1推理接口	/v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	/api/v2/chat/completions

版权所有 © 华为云计算技术有限公司

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-42 我的空间

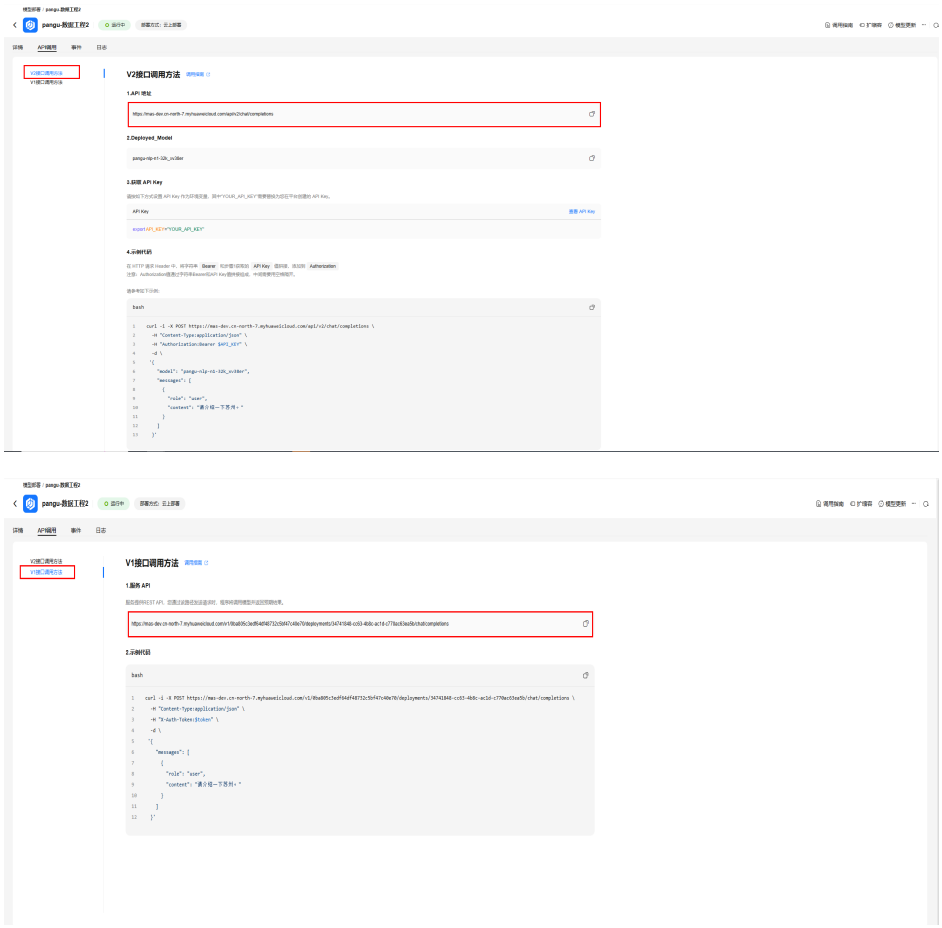


2. 获取调用路径。

   在左侧导航栏中选择“模型开发 > 模型部署”。

   - 获取已部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“API调用”页签，可获取模型调用路径，并根据页签的调用方法进行调用，如图5-43。

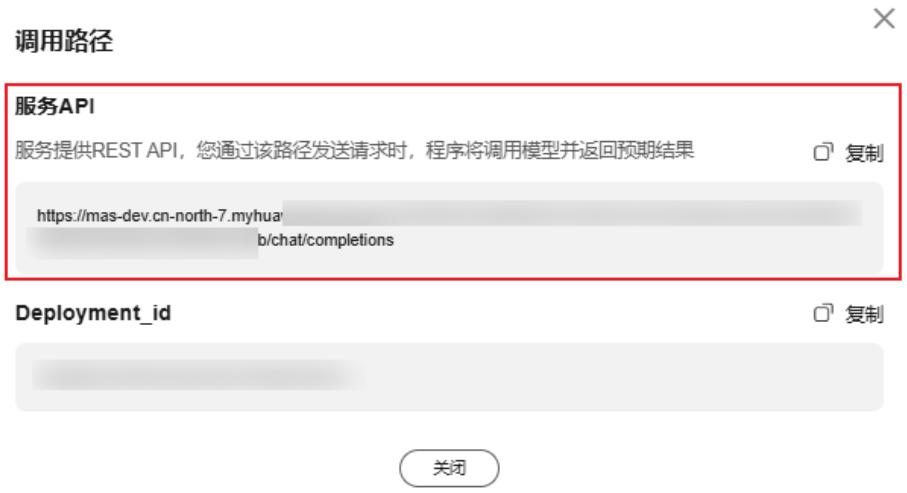
图 5-43 获取已部署模型的调用路径



版权所有 © 华为云计算技术有限公司

- 获取预置服务的调用路径。在“预置服务”页签中，选择所需调用的NLP大模型，单击“调用路径”，在“调用路径”弹窗可获取模型调用路径，如图5-44。

图 5-44 获取预置服务的调用路径



获取边缘部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“详情”页签，可获取模型调用路径。

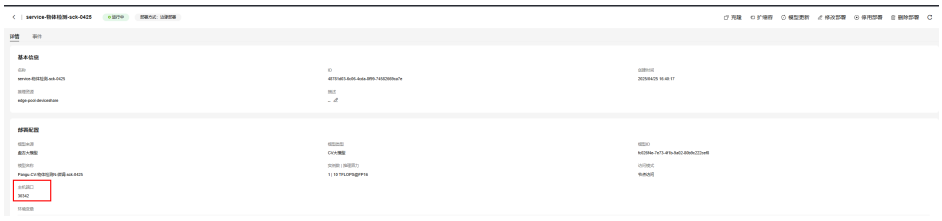
负载均衡方式：

模型路径为http://{ELB IP}:{ELB负载端口}/{API URL}/{推理的调用URL}，ELP IP需获取对应的公网IP，各个部分的获取如图所示：



节点方式：

模型路径为http://{节点所在的IP}:{主机端口}/{推理的调用URL}，节点所在IP为边缘池的工作节点的IP，各个部分的获取如图所示：



使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见获取调用路径。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，API Key认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。

版权所有 © 华为云计算技术有限公司

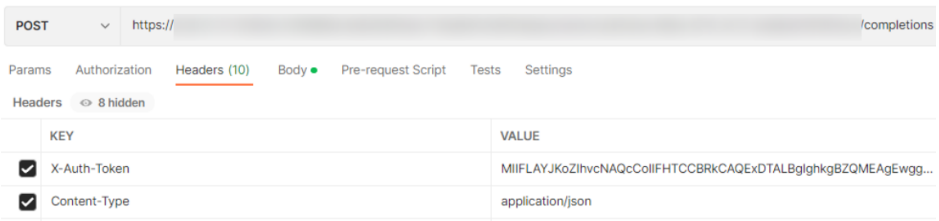
参考表5-34填写请求Header参数。

表 5-34 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
V1推理接口API Key认证	Content-Type	application/json
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。
V2推理接口API Key认证	Content-Type	application/json
	Authorization	用户创建应用接入获取的API Key，拼接“Bearer ”后的字符串。示例：Bearer d59******9C3

如图5-45，为Token认证方式的请求Header参数填写示例。

图 5-45 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "messages": [
     {
        "content": "介绍下长江，以及长江中典型的鱼类"
     }
    ],
   "temperature": 0.9,
```


版权所有 © 华为云计算技术有限公司

```
    "max_tokens": 600
}
```

4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示NLP大模型API调用成功。

API Key 鉴权说明

当用户部署的推理服务需要开放给其他用户调用时，原有Token认证无法支持，可用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为API Key值），API服务仅校验API Key，从而实现快速响应。

获取API Key步骤如下：

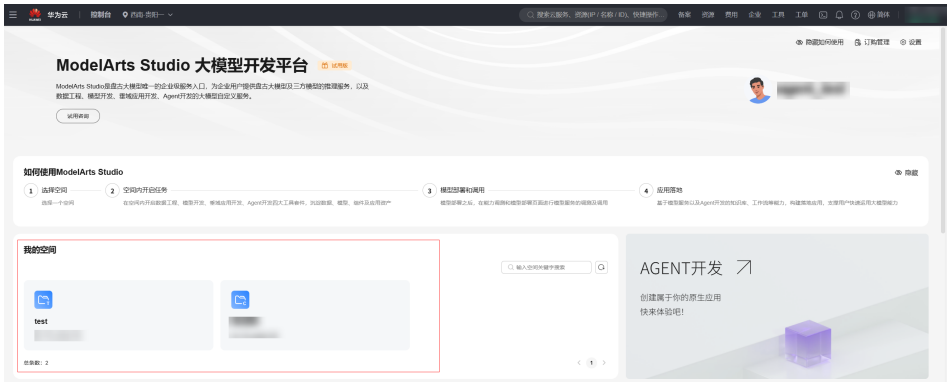
1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。
3. 在“关联服务”中，选择全部服务或者指定已部署好的推理服务，并设置应用接入名称、描述，单击“确定”。
4. 在“应用接入”列表的“API Key”列中可获取API Key值。

5.6.3 统计 NLP 大模型调用信息

针对调用的大模型，平台提供了统一的管理功能。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 5-46 我的空间



2. 单击左侧导航栏“调用统计”，选择“NLP”页签。
3. 选择当前调用的NLP大模型，可以按照不同时间跨度查看当前模型的调用总数、调用失败的次数、调用的总Tokens数、以及输入输出的Tokens数等基本信息。

   此外，该功能还提供了可视化界面，可额外查看响应时长以及安全护栏拦截次数。

   对于已经接入API Key方式鉴权的模型，可以直接在“应用接入”页面，单击操作列“调用统计”按钮跳转至“调用统计”功能页查看调用信息。

版权所有 © 华为云计算技术有限公司

图 5-47 调用统计



版权所有 © 华为云计算技术有限公司

6 开发盘古 CV 大模型

使用数据工程构建CV大模型数据集
训练CV大模型
部署CV大模型
评测CV大模型
调用CV大模型

6.1 使用数据工程构建 CV 大模型数据集

CV 大模型支持接入的数据集类型

盘古CV大模型支持接入图片类、视频类、其他类数据集，不同模型所需数据见**表6-1**，数据集格式要求请参见**图片类数据集格式要求**、**视频类数据集格式要求**、**其他类数据集格式要求**。

表 6-1 训练 CV 大模型数据集类型要求

基模型	训练场景	文件内容	文件格式
盘古-CV-图像分类	微调	图片+分类标注	图片+txt
盘古-CV-物体检测	微调	图片+检测标注	图片+xml
盘古-CV-语义分割	微调	图片+语义分割标注	图片+xml
盘古-CV-实例分割	微调	图片+实例分割标注	图片+xml

版权所有 © 华为云计算技术有限公司

训练 CV 大模型所需数据量

初期启动训练时，每种模型类别先提供**1000张已标注的图片数据**进行训练，后续根据验证结果再动态提供数据迭代。

- 图片中需要识别的目标是清晰可见的，没有遮挡、模糊等特征破坏问题。图片中的目标大小显著，目标物体在不放大图片的情况下人眼清晰可见。
- 画面光照良好，如果是在恶劣天气、户外、晚上等光照不好的场景，需要有补光设备保证良好的光照条件，需要保障在图片中人眼能清晰辨别目标。
- 图片分辨率大于640x640 px，关于拍摄角度、距离、分辨率等画面拍摄条件，需要保证训练集图片和测试部署时的图片保持一致。



构建 CV 大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程构建盘古CV大模型数据集流程见**表6-2**。

表 6-2 盘古 CV 大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	**导入数据至盘古平台**
加工图片、视频类数据集	加工图片、视频类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	**加工图片类数据集**、**加工视频类数据集**
	标注图片、视频类数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。	**标注图片类数据集**、**标注视频类数据集**
	配比图片类数据集	数据配比是将多个数据集按特定比例组合的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。	**配比图片类数据集**

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
发布图片、视频类数据集	评估图片、视频类数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。	评估图片类数据集、评估视频类数据集
	发布图片、视频类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。 • 标准格式：平台默认的格式。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。	发布图片类数据集、发布视频类数据集

6.2 训练 CV 大模型

6.2.1 CV 大模型训练流程与选择建议

CV 大模型训练流程介绍

目前，CV大模型支持微调训练。

- **预训练阶段：**通过在大规模通用数据集上训练，使模型学习到通用视觉特征，使得模型在面对不同类型的图像数据时，能够快速捕捉到关键信息，提升了模型的训练效率和泛化能力。
- **微调阶段：**微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。

CV 大模型选择建议

选择合适的CV大模型类型有助于提升训练任务的准确程度。您可以根据模型适用场景，选择合适的模型，从而提高模型的整体效果，详见表6-3。

版权所有 © 华为云计算技术有限公司

表 6-3 CV 大模型的类型

模型名称	适用场景	说明
Pangu-CV-物体检测-S-2.1.0	盘古计算机视觉物体检测大模型，该模适用的任务是找出图像中所有感兴趣的目标，确定它们的位置和类别。适用于如积水检测、占道经营检测、人员离岗检测、动植物检测、工业缺陷检测等任务。	S模型特点是小参数量，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。支持全量微调，在线推理，边缘推理，1卡起训，1卡部署。
Pangu-CV-物体检测-N-2.1.0	该模型属于物体检测模型，旨在识别图像中的所有感兴趣目标，定位其位置并确定其类别。适用于各种任务，如：积水检测、占道经营检测、人员离岗检测、动植物检测、工业缺陷检测等。	2024年12月发布的版本，支持全量微调、在线推理。 物体检测-N模型为中参数量模型，在保证计算效率的同时，具备较强的特征识别能力，提供高效的性能表现。
Pangu-CV-图像分类-2.1.0	盘古计算机视觉图像分类大模型，根据在图像信息中所反映的不同特征，把不同类别的图像区分开来的图像处理方法，利用计算机视觉技术对图像进行定量分析，把图像划归为若干个类别中的一种或多种。适用于动植物分类、车辆类型分类、车牌分类、废钢定级、零部件分类等任务。	支持全量微调，在线推理，边缘推理，1卡起训，1卡部署。
Pangu-CV-物体检测-S-3.1.0	盘古计算机视觉物体检测大模型，该模适用的任务是找出图像中所有感兴趣的目标，确定它们的位置和类别。适用于如积水检测、占道经营检测、人员离岗检测、动植物检测、工业缺陷检测等任务。	S模型特点是小参数量，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。支持全量微调，在线推理，边缘推理，1卡起训，1卡部署。
Pangu-CV-实例分割-1.1.0	盘古计算机视觉实例分割大模型，可以检测输入图像中的目标，并为每个目标像素分配类别标签。实例分割能够对前景语义类别相同的不同实例进行区分。	支持全量微调，在线推理，边缘推理，1卡起训，1卡部署。
Pangu-CV-语义分割-2.1.0	将数字图像细分为多个图像子区域，适用于车道分割、建筑分割、选煤厂筛面状态识别等任务	支持全量微调，在线推理，边缘推理，1卡起训，1卡部署。

版权所有 © 华为云计算技术有限公司

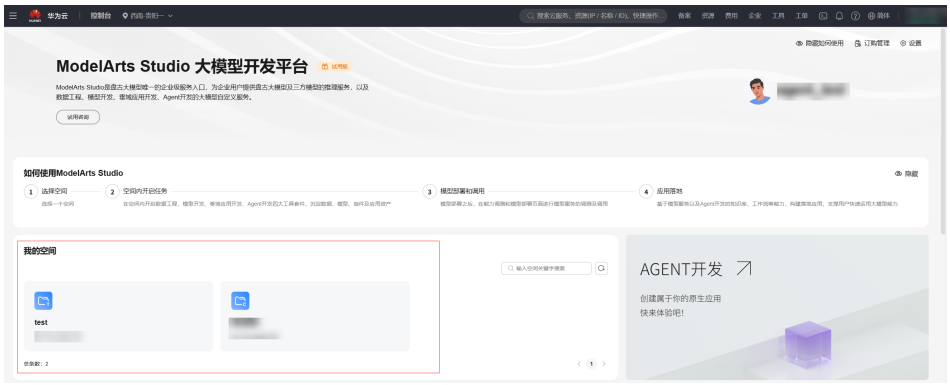
6.2.2 创建 CV 大模型训练任务

创建 CV 大模型微调任务

创建CV大模型微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台首页，在“我的空间”模块，单击进入所需空间。

图 6-1 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > CV大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表6-4完成训练参数设置。

表 6-4 CV 大模型微调参数说明

参数分类	训练参数	说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“CV大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
训练参数	训练参数	模型训练参数，参考表6-5。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	说明
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

表 6-5 CV 大模型微调参数说明

模型类型	训练参数	说明
Pangu-CV-物体检测-S-3.1.0	数据集	训练数据集。
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	单卡训练批处理大小	单卡批量处理大小，建议取值大于等于1。
	单卡验证批处理大小	单卡验证批量处理大小，建议取值大于等于1。
	图像分辨率	训练和推理的统一图像分辨率尺寸， 要求为32的倍数，不符合要求会自动转换为最近的32倍数尺寸。
	学习率	学习率决定了每次训练时模型参数更新的幅度。取值为大于0，小于等于1， 取列表中的第一个值。
	模型规格	不同模型规格，X参数量最大，L参数量适中，S参数量最小。
	混合精度训练	是否使用混合精度训练开关，开启时可降低训练显存占用但存在溢出风险，关闭时可提高训练稳定性但显存和训练速度有所增加。
	标签映射	是否增加标签映射功能，增加预训练模型的迁移效果。
Pangu-CV-物体检测-S-2.1.0	数据集	训练数据集。
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	批次大小	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	验证集图片保存最大数量	用于限制验证过程中模型预测错误（与标注不符）的图片保存数量，便于后续分析错误类型，同时避免存储空间浪费。
	学习率	学习率决定了每次训练时模型参数更新的幅度。
	权重衰减	通过在每次参数更新时对模型权重进行小幅度缩减，避免模型过度依赖特定数据。
	模型保存轮数	模型训练时保存断点的间隔轮数。
Pangu-CV-物体检测-N-2.1.0	数据集	训练数据集。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	热身轮次	表示在模型训练初期，逐步增加学习率到预设值的训练轮次，用于帮助模型在训练初期稳定收敛，避免大幅度的参数更新导致不稳定的学习过程。
	热身阶段学习率	热身轮次中使用的初始学习率。
	权重衰减	用于防止模型过拟合。在更新模型权重时，它会对模型参数施加惩罚，使得权重值趋于较小，从而提高模型的泛化性能。
	优化器	选择用于训练模型的优化算法。这里选择的“sgd”是随机梯度下降法，它是深度学习中常用的优化算法之一，适合大规模数据集训练。
	锚框的长边和短边的比例	定义检测物体锚框的长宽比。通过设置不同的长短比例，模型可以更好地适应多种尺寸和形状的物体。
	锚框大小	指锚框的初始尺寸。锚框是物体检测中的一个关键概念，通过合理设置，可以帮助模型检测出多种尺寸的目标。
	框重叠比例阈值	用于计算损失时区分正负样本的IoU阈值。当候选框与某个真实框的IoU大于等于该阈值时，该候选框被视为正样本；否则视为负样本。
	滑动平滑训练	一种训练策略，通过在模型预测的标签上添加少量噪声来避免过拟合，常用于提升模型在测试数据集上的泛化能力。
	极大值抑制阈值	在预测多个边界框时，用于去除高度重叠的边界框。此阈值控制相似的边界框保留的条件。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	类别无关极大值抑制开关	决定是否在不同类别中应用极大值抑制阈值。
	模型大小调整因子	用于模型抽取过程中，选择不用的模型因子来调节模型参数的多少。
	输入模型图像宽度	用于控制模型输入图像分辨率的宽度，以此调节适应不用场景的模型识别能力。
	输入模型图像高度	用于控制模型输入图像分辨率的高度，以此调节适应不用场景的模型识别能力。
	模型保存轮数	模型训练时保存断点的间隔轮数。
Pangu-CV-图像分类-2.1.0	数据集	训练数据集。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	每卡批次大小	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	分类模式	图像分类模式，支持单标签分类和多标签分类。
	超参搜索	是否启用超参搜索获取最优的超参用于训练。
	初始学习率	学习率决定了每次训练时模型参数更新的幅度。初始学习率是模型训练最开始阶段所设定的学习率，它是学习率的初始值。
	超参搜索训练轮数	在超参搜索阶段，设置模型会经过几轮训练以找到最优的超参数组合。
Pangu-CV-实例分割-1.1.0	数据集	训练数据集。
	每卡批大小	单张卡上每次前向传播使用的样本数量，每卡批大小*节点数*每节点卡数需要小于训练数据量。
	基础学习率	用于控制参数更新的步长,学习率过大可能会导致训练过程震荡不稳,无法收敛到最优解，学习率过小,可能会导致训练过程非常缓慢。
	是否使用自定义L1预训练模型	是否使用自定义预训练模型进行训练。
	自定义L1预训练模型目录	自定义预训练模型所在的OBS路径。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
Pangu-CV-语义分割-2.1.0	数据集	训练数据集。
	是否使用自定义L1预训练模型	是否使用自定义预训练模型进行训练，模型为用户与服务共建可联系CV算法技术支持获取。
	自定义L1预训练模型的文件	自定义预训练模型所在的OBS路径。
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	批量处理大小	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	是否忽略类别0	该参数功能是为了忽略数据集类别0，具体做法是将类别0改为255，其余类别相应编号减1，同时decode head里将255设为ignore index，即不参与loss计算。
	学习率	用于控制参数更新的步长，学习率过大可能会导致训练过程震荡不稳无法收敛到最优解，学习率过小可能会导致训练过程非常缓慢。

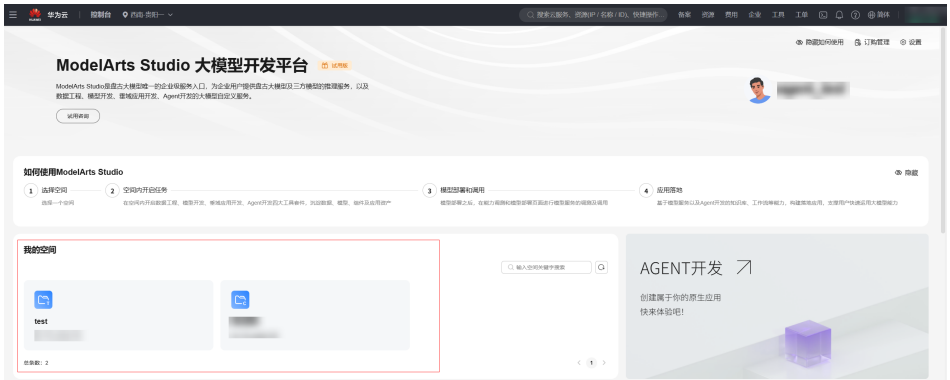
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建 CV 大模型预训练任务

创建CV大模型预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台首页，在“我的空间”模块，单击进入所需空间。

图 6-2 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > CV大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表6-6**完成训练参数设置。

表 6-6 CV 大模型预训练参数说明

参数分类	训练参数	说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“CV大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
训练参数	训练参数	模型训练参数，参考**表6-7**。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

表 6-7 CV 大模型预训练参数说明

模型类型	训练参数	说明
Pangu-CV-物体检测-S-3.1.0	数据集	训练数据集。
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	单卡训练批处理大小	单卡批量处理大小，建议取值大于等于1。

版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	单卡验证批处理大小	单卡验证批量处理大小，建议取值大于等于1。
	图像分辨率	训练和推理的统一图像分辨率尺寸, 要求为32的倍数，不符合要求会自动转换为最近的32倍数尺寸。
	学习率	学习率决定了每次训练时模型参数更新的幅度。取值为大于0，小于等于1, 取列表中的第一个值。
	模型规格	不同模型规格，X参数量最大，L参数量适中，S参数量最小。
	混合精度训练	是否使用混合精度训练开关，开启时可降低训练显存占用但存在溢出风险，关闭时可提高训练稳定性但显存和训练速度有所增加
	标签映射	是否增加标签映射功能，增加预训练模型的迁移效果。
Pangu-CV-物体检测-S-2.1.0	数据集	训练数据集。
	训练轮次	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	批次大小	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	验证集图片保存最大数量	用于限制验证过程中模型预测错误（与标注不符）的图片保存数量，便于后续分析错误类型，同时避免存储空间浪费。
	学习率	学习率决定了每次训练时模型参数更新的幅度。
	权重衰减	通过在每次参数更新时对模型权重进行小幅度缩减，避免模型过度依赖特定数据。
Pangu-CV-物体检测-N-2.1.0	数据集	训练数据集。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	热身轮次	表示在模型训练初期，逐步增加学习率到预设值的训练轮次，用于帮助模型在训练初期稳定收敛，避免大幅度的参数更新导致不稳定的学习过程。
	热身阶段学习率	热身轮次中使用的初始学习率。
	权重衰减	用于防止模型过拟合。在更新模型权重时，它会对模型参数施加惩罚，使得权重值趋于较小，从而提高模型的泛化性能。
	优化器	选择用于训练模型的优化算法。这里选择的“sgd”是随机梯度下降法，它是深度学习中常用的优化算法之一，适合大规模数据集训练。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	锚框的长边和短边的比例	定义检测物体锚框的长宽比。通过设置不同的长短比例，模型可以更好地适应多种尺寸和形状的物体。
	锚框大小	指锚框的初始尺寸。锚框是物体检测中的一个关键概念，通过合理设置，可以帮助模型检测出多种尺寸的目标。
	框重叠比例阈值	用于计算损失时区分正负样本的IoU阈值。当候选框与某个真实框的IoU大于等于该阈值时，该候选框被视为正样本；否则视为负样本。
	滑动平滑训练	一种训练策略，通过在模型预测的标签上添加少量噪声来避免过拟合，常用于提升模型在测试数据集上的泛化能力。
	极大值抑制阈值	在预测多个边界框时，用于去除高度重叠的边界框。此阈值控制相似的边界框保留的条件。
	类别无关极大值抑制开关	决定是否在不同类别中应用极大值抑制阈值。
	模型大小调整因子	用于模型抽取过程中，选择不用的模型因子来调节模型参数的多少。
	输入模型图像宽度	用于控制模型输入图像分辨率的宽度，以此调节适应不用场景的模型识别能力。
	输入模型图像高度	用于控制模型输入图像分辨率的高度，以此调节适应不用场景的模型识别能力。
	模型保存轮数	模型训练时保存断点的间隔轮数。
Pangu-CV-图像分类-2.1.0	数据集	训练数据集。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	每卡批次大小	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	分类模式	图像分类模式，支持单标签分类和多标签分类。
	超参搜索	是否启用超参搜索获取最优的超参用于训练。
	初始学习率	学习率决定了每次训练时模型参数更新的幅度。初始学习率是模型训练最开始阶段所设定的学习率，它是学习率的初始值。
	超参搜索训练轮数	在超参搜索阶段，设置模型会经过几轮训练以找到最优的超参数组合。

 版权所有 © 华为云计算技术有限公司

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

6.2.3 查看 CV 大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表6-8。

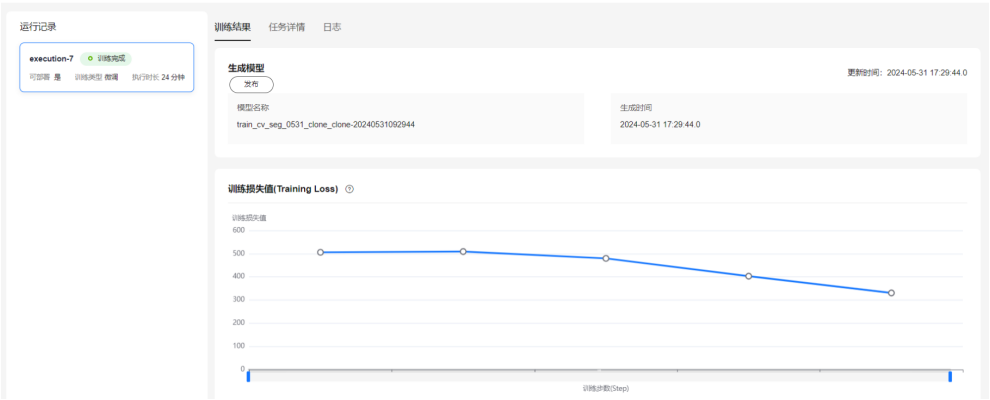
表 6-8 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见图6-3。

图 6-3 查看训练指标



版权所有 © 华为云计算技术有限公司

表 6-9 训练指标说明

模型	训练指标	指标说明
Pangu-CV-物体检测-S-2.1.0 Pangu-CV-物体检测-S-3.1.0 Pangu-CV-物体检测-N-2.1.0 Pangu-CV-实例分割-1.1.0 Pangu-CV-图像分类-2.1.0	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	精准率	精准率是指在模型预测为正类的样本中，真正类样本的比例。数值越高，表明模型在检测正类样本时的准确性越高。
	召回率	召回率是指在所有实际为正类的样本中，被模型正确预测为正类的比例。数值越高，表明模型在检测正类样本时的全面性越高。
Pangu-CV-语义分割-2.1.0	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	平均交并比	是一种常用的评估指标，主要用于衡量图像分割任务的性能。它通过计算预测结果与真实标签之间的交集与并集的比值，来评估模型的准确性。是一个综合性的评估指标，能够全面反映模型在图像分割任务中的表现。
	像素精度	是语义分割任务中的一种简单而直观的评估指标，用于衡量模型预测正确的像素占总像素的比例。在初步评估模型性能时，像素精度可以提供一个快速的参考。
	平均精度	是语义分割任务中的一种评估指标，用于衡量每个类别中预测正确的像素占该类别总像素的比例，然后对所有类别的精度取平均值。在需要平衡不同类别性能的评估中，平均精度是一个有用的指标。
	平均Dice系数	是语义分割任务中的一种评估指标，用于衡量预测结果与真实标签之间的相似度。Dice系数通过计算两个集合的交集与并集的比值来评估模型的性能。

 版权所有 © 华为云计算技术有限公司

模型	训练指标	指标说明
	F1分数	是评估分类模型性能的一种常用指标，特别是在处理类别不平衡问题时尤为有用。F1分数是精确率（Precision）和召回率（Recall）的调和平均值，能够综合反映模型在正类上的性能。
	精准率	精准率是指在模型预测为正类的样本中，真正类样本的比例。数值越高，表明模型在检测正类样本时的准确性越高。
	召回率	召回率是指在所有实际为正类的样本中，被模型正确预测为正类的比例。数值越高，表明模型在检测正类样本时的全面性越高。

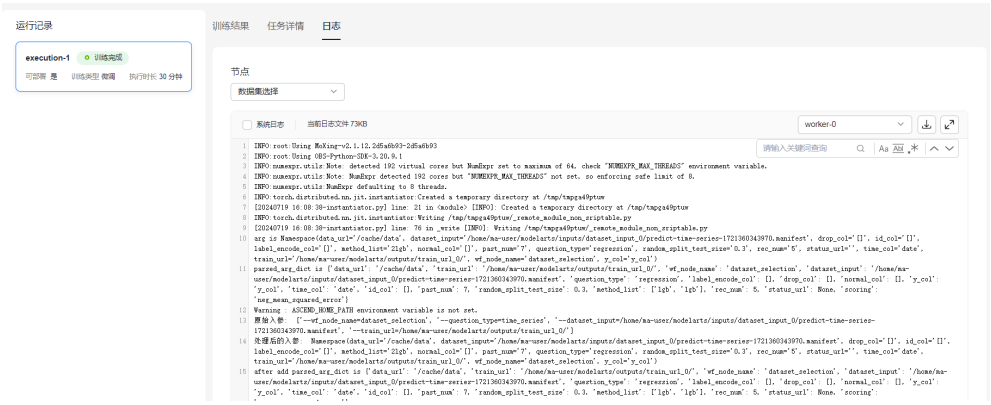
获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**CV大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 6-4 获取训练日志



训练异常定位方法

模型训练失败时，可以按照如下定位思路对训练任务做初步定位。

1. 在训练任务列表查看任务失败原因。

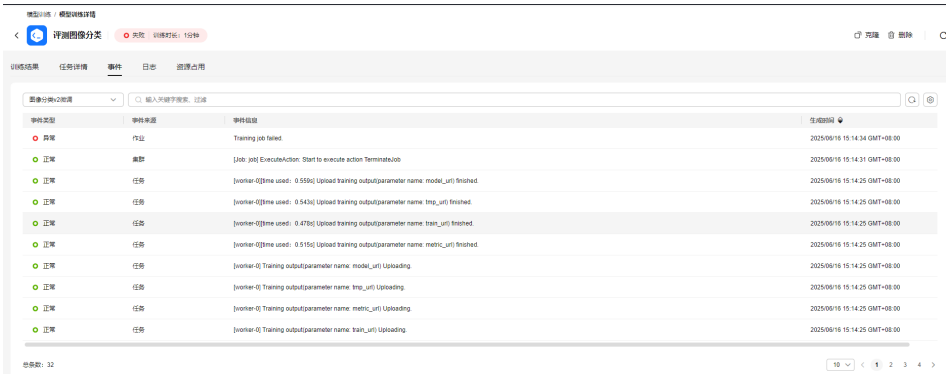
   模型启动训练后，可以在模型训练列表中查看训练任务的状态，当任务状态为“失败”时，可将鼠标放置在“失败”上，可以查看到任务失败的概要信息。

版权所有 © 华为云计算技术有限公司



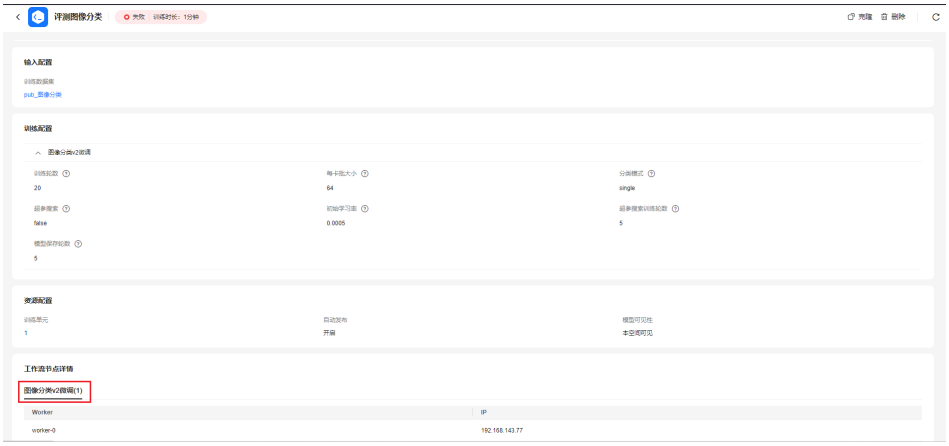
2. 单击任务名，进入任务详情页查看训练任务运行关键事件信息。

   单击“事件”页签，可查看训练任务子任务运行事件，通过事件可以观察到当前任务的运行情况。事件可以按照“正常、异常、告警”三类事件类型过滤。



3. 获取当前任务运行的worker、workerIP以及任务异常退出时的节点和退出码信息。

   在ModelArts Studio训练任务页面单击需要查看的任务名，进入“任务详情”页，在“工作流节点详情”下可以查看到当前任务所运行的节点及IP信息。



4. 当训练任务运行失败时，可查看“模型训练详情 -> 基本信息 -> 状态”，单击“查看失败节点清单”获取训练任务故障节点及退出码信息。

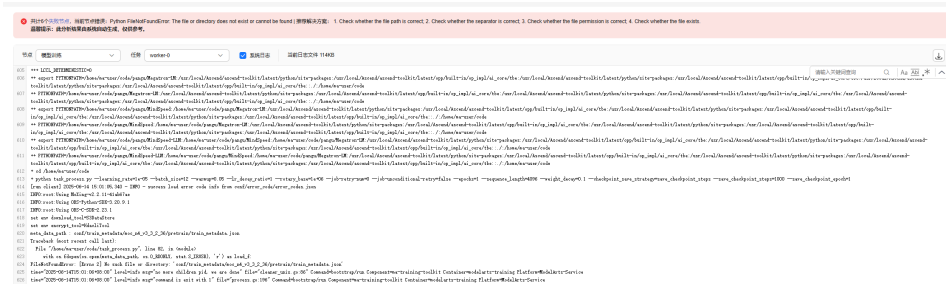


5. 训练作业遇到问题时，可查看“模型训练详情 -> 日志”，多数场景下的问题可以通过日志报错信息直接定位。

版权所有 © 华为云计算技术有限公司

如果训练作业运行失败，ModelArts Studio会自动识别导致作业失败的原因，在训练日志界面上给出提示并对部分常见训练错误给出分析建议。包含失败的可能原因、推荐的解决方案。

图 6-5 训练错误分析建议



说明

针对分布式作业，只会显示当前节点的一个分析结果，作业的失败需要综合各个节点的失败原因做一个综合判断。

6.2.4 发布训练后的 CV 大模型

CV大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-6 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

版权所有 © 华为云计算技术有限公司

图 6-7 训练结果页面



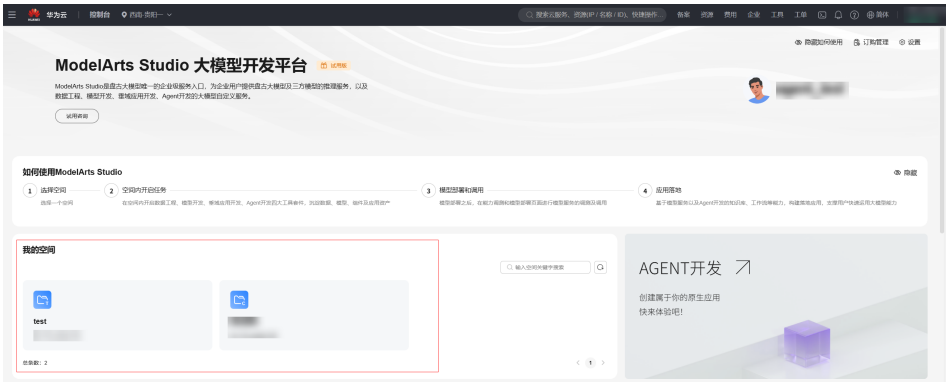
4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

6.2.5 管理 CV 大模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、启动、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-8 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 克隆。单击操作列的“更多 > 克隆”，参照**创建CV大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“更多 > 停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的训练任务。

     说明

     删除属于高危操作，删除前请确保当前任务不再需要。

6.2.6 CV 大模型训练常见报错与解决方案

CV大模型训练常见报错及解决方案请详见**表6-10**。

版权所有 © 华为云计算技术有限公司

表 6-10 CV 大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空。	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。
训练日志提示“root: XXX valid number is 0”	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。
训练日志提示“ValueError: label_map not match”	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

6.3 部署 CV 大模型

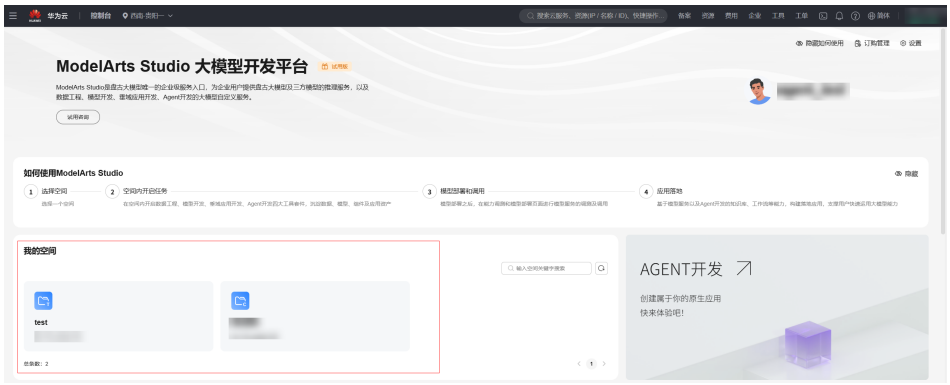
6.3.1 创建 CV 大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 6-9 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > CV大模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表6-11**完成部署参数设置。

表 6-11 CV 大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“CV大模型”，并选择需要进行部署的模型和版本。
	部署方式	选择“云上部署”。 支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 资源池。选择部署模型所需的资源池。 – CPU。部署需要使用的最小CPU值（物理核）。 – 内存。部署需要使用的最小内存值。 – Ascend。部署使用的NPU数量。 – 实例数。设置部署模型时所需的实例数。
资源配置	计费模式	包年包月计费模式。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。

版权所有 © 华为云计算技术有限公司

参数分类	部署参数	参数说明
	描述（选填）	设置部署任务的描述。

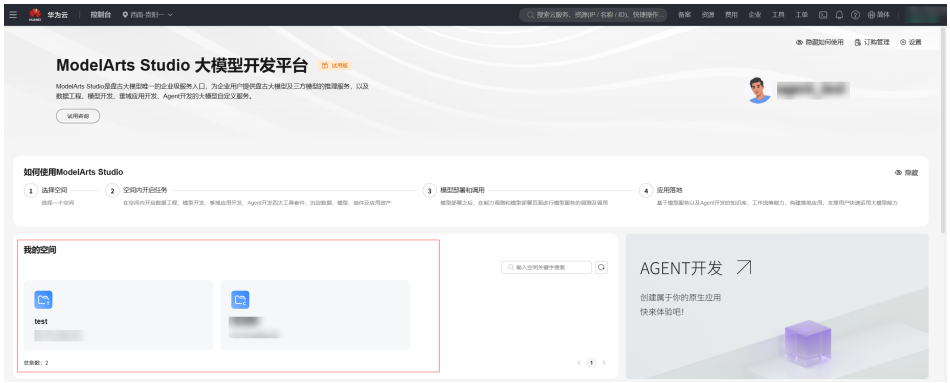
5. 参数填写完成后，单击“立即部署”。

6.3.2 查看 CV 大模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-10 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建CV大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

6.3.3 管理 CV 大模型部署任务

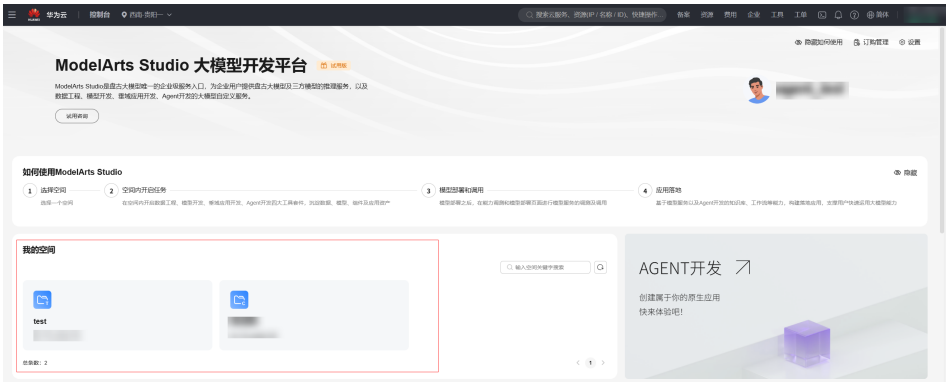
模型更新

完成**创建CV大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 6-11 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 6-12 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

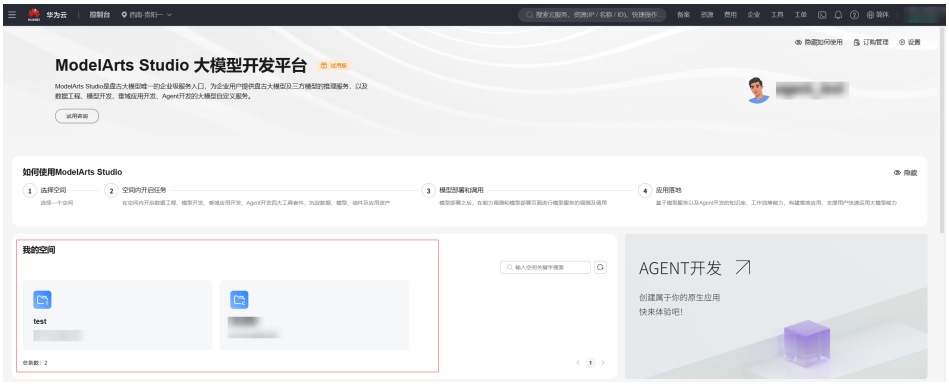
修改部署配置

完成**创建CV大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

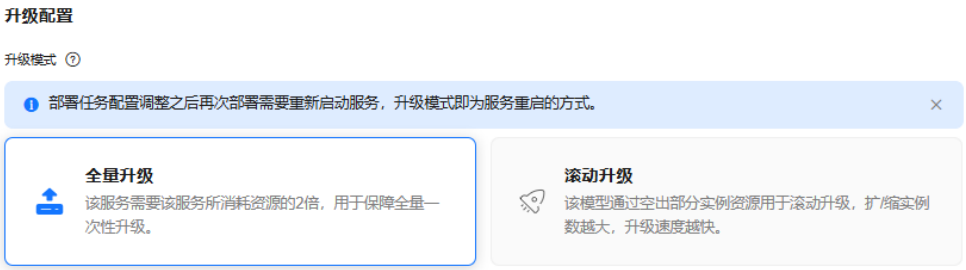
版权所有 © 华为云计算技术有限公司

图 6-13 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 6-14 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

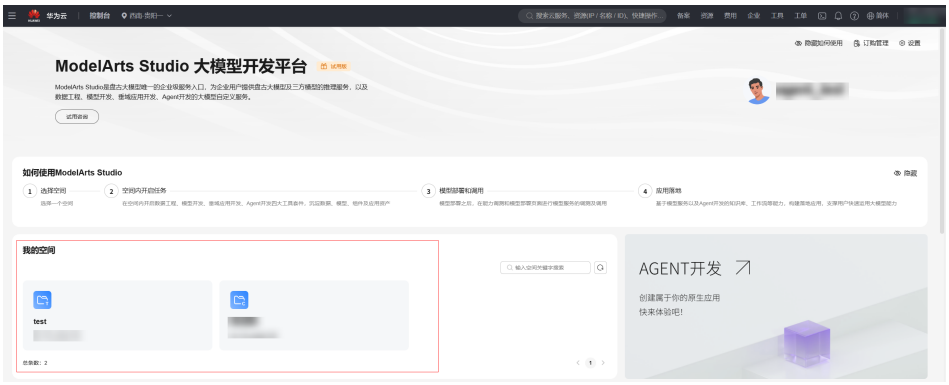
模型部署实例扩缩容

完成**创建CV大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 6-15 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

6.4 评测 CV 大模型

6.4.1 创建 CV 大模型评测数据集

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同情景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

CV大模型支持基于规则的自动评测，即基于相似度/准确率进行打分，对比模型预测结果与标注数据的差异。

评测数据集的创建步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“评测中心 > 评测集管理 ”，单击界面右上角“创建评测集”。
3. 在“创建评测集”页面选择CV大模型，选择评测模型，再选择OBS中的评测集目录（需先把评测集上传到OBS）。

   CV大模型评测数据集支持的格式见表6-12。

表 6-12 评测数据集格式

模型类型	评测数据集格式
图像分类	图片+txt
物体检测	图片+xml
语义分割	评测图片+标注图片

4. 选择好OBS路径后，填写“评测集名称”与“描述”，单击“立即创建”。

版权所有 © 华为云计算技术有限公司

5. 评测集任务创建后，返回“评测集管理”查看评测集任务。

6.4.2 创建 CV 大模型评测任务

创建CV大模型评测任务前，请确保已完成**创建CV大模型评测数据集**操作。

创建 CV 大模型自动评测任务（基于规则）

创建CV大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-16 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考**表6-13**完成部署参数设置。

表 6-13 CV 大模型自动评测任务参数说明（基于规则）

参数分类	参数名称	参数说明
选择服务	模型类型	选择“CV大模型”。
	评测模型	评测不同的CV模型。支持物体检测N，物体检测S，图像分类，语义分割等类型。
	服务来源	支持已部署服务。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 **说明：仅支持云上部署模型做自动评测。**
评测配置	评测类型	选择“自动评测”。
	评测规则	选择“基于规则”。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	数据集标注模式	选择“有标注模式”。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“模型评测 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”查看模型评测结果，包括模型详细的得分以及评测明细。

6.4.3 查看 CV 大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-17 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的报告和详情。

   其中，各评测指标说明详见**CV大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 服务结果分析”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

版权所有 © 华为云计算技术有限公司

CV 大模型评测指标说明

CV大模型支持基于规则的自动评测方式，其评测指标见表6-14。

表 6-14 CV 大模型评测指标说明

模型类型	模型场景	评测指标（自动评测）	指标说明
CV大模型	物体检测	mAP	平均精度值。
		精确率	预测为正的样本中，实际为正的比率。
		召回率	实际为正的样本，预测为正的比率。
		F1分数	精准率和召回率的调和平均数，数值越高，表明模型性能越好。
		预测正确	真正例和真负例的和。
	图像分类-单分类	准确率(Top1)	模型预测的前1个最高概率类别中包含真实标签的样本比例。
		准确率(Top5)	模型预测的前5个最高概率类别中包含真实标签的样本比例。
		预测正确	得分最高的类别和标签类别相等的样本数。
	图像分类-多分类	mAP	平均精度值。
		精确率	预测为正的样本中，实际为正的比率。
		召回率	实际为正的样本，预测为正的比率。
		F1分数	精准率和召回率的调和平均数，数值越高，表明模型性能越好。
		预测正确	预测得分大于置信度阈值的类别和对应标签包含的类别完全相等的样本数。
	语义分隔	MIoU	预测区域与真实区域的交集面积与并集面积的比值。
		精确率	预测为正的样本中，实际为正的比率。
		召回率	实际为正的样本，预测为正的比率。
		F1分数	精准率和召回率的调和平均数，数值越高，表明模型性能越好。

版权所有 © 华为云计算技术有限公司

6.4.4 管理 CV 大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 6-18 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

     说明

     删除属于高危操作，删除前请确保当前任务不再需要。

6.5 调用 CV 大模型

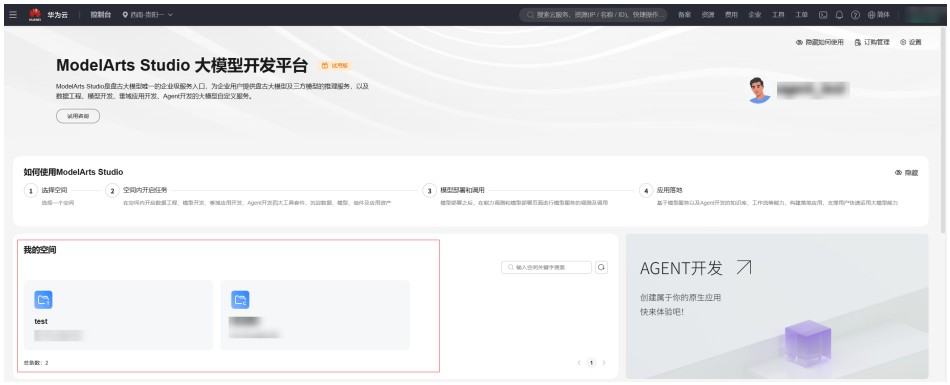
6.5.1 使用 API 调用 CV 大模型

获取 CV 大模型 API 请求地址

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 6-19 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击“模型名称”在“详情”页面获取API请求地址。

   只有处于“运行中”状态的模型才可以正常调用。

   云上部署：

图 6-20 云上部署获取 CV 大模型 API 请求地址



   边缘部署：

   - 如果是节点访问，则记录当前部署配置中的主机端口，在监控信息中获取节点IP，按照https://ip:port的格式，或者http://ip:port格式来拼接即可获取接口的调用域名。具体是使用https，还是http请在部署详情页面确认。
   - 如果是负载均衡的访问模式，获取部署配置中的ELB IP、ELB负载端口、API URL信息。按照https://ip:port+API URL的格式来拼接即可获取接口的调用域名。
   - 如果是监控网关的访问模式，获取部署配置中的ELB IP、ELB负载端口、API URL信息。按照https://ip:port+API URL的格式或者https://ip:port+API URL的格式来拼接即可获取接口的调用域名。具体是使用https，还是http请在部署详情页面确认。

3. 获取API接口域名后，结合API访问路径拼接成完整的API接口。

获取 Token

参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。

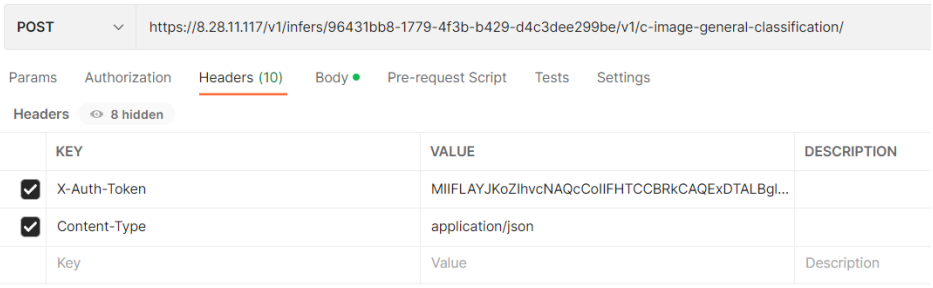
使用 API 调用 CV 大模型

1. 在Postman中新建POST请求，并填入API请求地址。

版权所有 © 华为云计算技术有限公司

2. 参考图6-21填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为获取Token中获取的Token值。

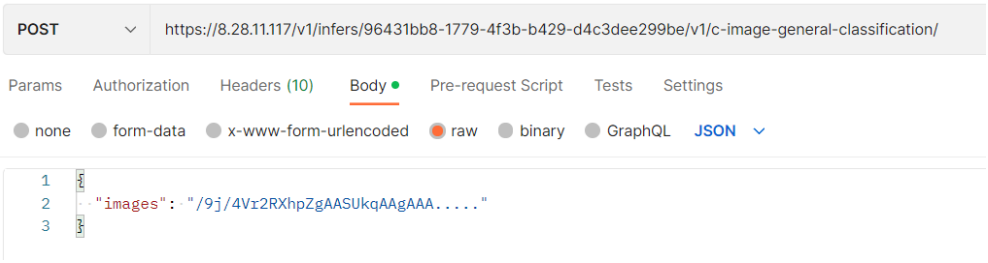
图 6-21 填写 CV 大模型 API



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。注意需要将图片转成base64编码传入。整个请求体大小建议不超过4M，建议使用JPG、PNG、JPEG、BMP格式的图片。

```
{
  "images": "/9j/4Vr2RXhpZgAASUkqAAgAAA....."
}
```

图 6-22 填写 CV 大模型请求 Body



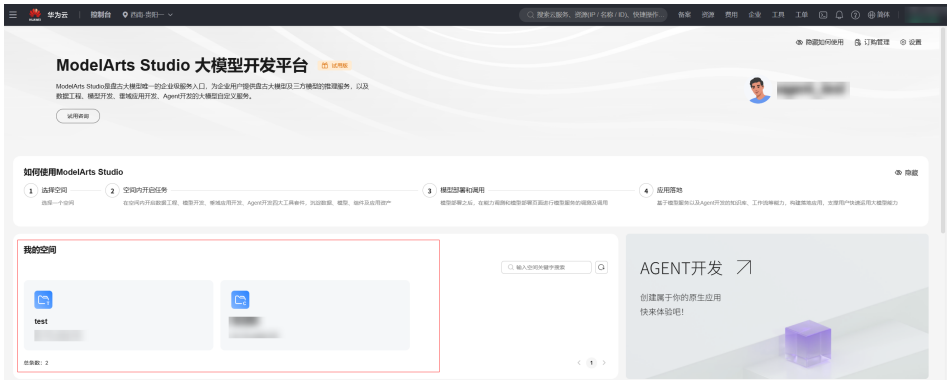
4. 单击Postman界面“Send”按钮，发送请求。

6.5.2 统计 CV 大模型调用信息

模型推理接口除了使用Token认证鉴权进行调用外，也支持使用API Key鉴权进行调用。对于API Key鉴权产生的调用信息，可以使用平台的“调用统计”功能对调用量信息、时延信息进行统计。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 6-23 我的空间



版权所有 © 华为云计算技术有限公司

2. 单击左侧导航栏“调用统计”。
3. 选择当前调用的CV大模型，可以按照不同时间跨度查看当前模型调用信息。

 版权所有 © 华为云计算技术有限公司

7 开发盘古预测大模型

使用数据工程构建预测大模型数据集
训练预测大模型
部署预测大模型
评测预测大模型
调用预测大模型

7.1 使用数据工程构建预测大模型数据集

预测大模型支持接入的数据集类型

盘古预测大模型仅支持接入预测类数据集，不同模型所需数据见**表7-1**，该数据集格式要求请参见**预测类数据集格式要求**。

表 7-1 预测大模型与数据集类型对应关系

基模型	模型分类	数据集类型	文件格式
预测大模型	统一编码时序预测分类大模型	时序分类	csv
	统一编码时序预测回归大模型	时序回归	csv
	统一编码表格预测分类大模型	结构化分类	csv
	统一编码表格预测回归大模型	结构化回归	csv
	时序异常检测大模型	时序回归	csv
	融合推荐分类大模型	结构化分类	csv

版权所有 © 华为云计算技术有限公司

基模型	模型分类	数据集类型	文件格式
	融合推荐分类大模型-极简模式	结构化分类	csv
	融合推荐回归大模型	结构化回归	csv
	融合推荐回归大模型-极简模式	结构化回归	csv
	融合推荐异常检测大模型	结构化分类	csv

图 7-1 创建数据集



训练预测大模型所需数据量

训练预测大模型时，所需的数据通常为**表格格式**，即由行和列组成的扁平化数据。具体要求如下：

- **行**：每行代表一个样本。每行与其他行具有相同的列，并且顺序相同，这些行通常按照某种特定顺序排列。
- **列**：每列表示一种特征。每列的数据类型应保持一致，不同列可以具有不同的数据类型。
- **顺序**：表格中的行通常按照特定顺序排列。
- **行数**：数据表的行数应大于5000行。
- **维度**：数据的维度（列数）应大于10维。
- **数据完整性**：必须确保数据中没有缺失值。

构建预测大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程构建盘古预测大模型数据集流程见**表7-2**。

 版权所有 © 华为云计算技术有限公司

表 7-2 盘古预测大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
发布预测类数据集	发布预测类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。	发布预测类数据集

7.2 训练预测大模型

7.2.1 预测大模型训练流程与选择建议

预测大模型训练流程介绍

目前，预测大模型支持微调训练、预训练。

- **微调阶段：**微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。针对部分预测大模型类型，开放**增量微调**，微调后的模型可以直接部署，也可以继续进行**二次微调**训练，以应对下游任务场景调整而引起的数据分布变化。
- **预训练阶段：**通过在大规模通用数据集上训练，使模型学习到通用特征，使得模型在面对不同类下游任务时，能够快速捕捉到关键信息，提升了模型的训练效率和泛化能力。

预测大模型选择建议

选择合适的预测大模型类型有助于提升训练任务的准确程度。您可以根据模型适用场景，选择合适的模型，从而提高模型的整体效果，详见**表7-3**。

 版权所有 © 华为云计算技术有限公司

表 7-3 预测大模型的类型

模型名称	适用场景
盘古融合推荐异常检测大模型（Pangu-Predict-Anom-Table-2.0.0）	**异常检测模型**用于识别数据集中的异常或离群点，常应用于安全、质量控制等领域。 ● 矿山行业：进行设备故障检测，例如监控设备运行数据，识别异常行为，防止设备故障。 ● 电力行业：进行电网异常检测，例如监控电网运行状态，识别异常情况，防止电网故障。 预测大模型提供了两种异常检测大模型，支持多种场景下的异常检测任务： 1. 盘古时序异常检测大模型，支持时序数据的异常检测任务，支持预训练、微调、以及二次微调。 a. 预训练：使用设备历史正常数据对时序异常检测大模型进行全量训练，使其更好地学习并重构正常数据的潜在特征分布。在推理阶段，利用模型对输入样本的重构误差作为异常判别依据：若重构效果显著下降，则表明该样本偏离正常模式，具有潜在异常趋势，从而实现对异常数据的有效检测与区分。 b. 微调：在时序异常检测大模型基础上引入分类头，并在训练过程中冻结模型的主干网络，仅使用历史中包含正常与异常样本的混合数据对新增分类头进行微调。通过此方式，将原始问题转化为针对特定时间点的二分类任务，判断其状态为正常或异常。 2. 盘古融合推荐异常检测大模型，支持结构化数据的异常检测二分类任务
盘古时序异常检测大模型（Pangu-Predict-Anom-TS-2.1.0）	
盘古统一编码时序预测分类大模型（Pangu-Predict-Cla-TS-2.1.0）	**分类模型**用于预测离散类别，可以应用于各种分类任务。 ● 矿山行业：进行智能洗选，例如根据煤炭的特征，将煤炭分类，提高洗选效率，降低能耗。 ● 油气行业：进行地质分层，例如基于地质数据，对不同地层进行分类，识别储层和非储层，提高勘探和开发效率。进行岩性识别，例如对不同岩石类型进行分类，帮助识别岩石的性质和特征，指导钻井和开采。进行流体识别，例如根据测井数据，识别储层中的油、气、水等流体类型。 预测大模型提供了两种分类大模型，支持多种场景下的分类任务： 1. 盘古统一编码时序预测/表格预测分类大模型，支持时序数据/结构化数据分类预测任务，支持二次微调功能。 2. 盘古融合推荐分类大模型，支持结构化数据分类预测任务。极简模式提供效率优先等极简训练配置，实现高效开发。
盘古统一编码表格预测分类大模型（Pangu-Predict-Cla-Table-2.2.0）	
盘古融合推荐分类大模型（Pangu-Predict-Cla-Table-2.0.0）	
盘古融合推荐分类大模型-极简模式（Pangu-Predict-Cla-Table-2.1.0）	

 版权所有 © 华为云计算技术有限公司

模型名称	适用场景
盘古统一编码时序预测回归大模型（Pangu-Predict-Reg-TS-2.1.0）	**回归模型**用于预测连续性数值，在工业场景中广泛用于工艺参数和性能指标的预测。 • 矿山行业：进行智慧配煤，例如预测焦炭成分质量，优化配煤比例，提高焦炭质量，降低生产成本。进行智能浓缩，例如预测焦炭成分质量，优化配煤比例，提高焦炭质量，降低生产成本。 • 油气行业：进行储层参数预测，例如预测储层的物理参数，如孔隙度、渗透率等，提高油气资源的开发利用效率。进行产能分级预测，例如预测油井的产能等级，优化油气生产计划。 • 电力行业：进行电力负荷预测，例如根据历史负荷数据，预测未来的电力负荷，优化电力生产和调度。 • 钢铁行业：进行钢水温度预测，例如预测钢水温度，提高浇注和连铸的准确性和效率。 预测大模型提供了两种回归大模型，支持不同场景下的回归任务： 1. 盘古统一编码时序预测/表格预测回归大模型，支持时序数据/结构化数据回归预测任务，支持二次微调功能。 2. 盘古融合推荐预测大模型，支持结构化数据回归预测任务。极简模式提供效率优先等极简训练配置，实现高效开发。
盘古统一编码表格预测回归大模型（Pangu-Predict-Reg-Table-2.2.0）	
盘古融合推荐回归大模型（Pangu-Predict-Reg-Table-2.0.0）	
盘古融合推荐回归大模型-极简模式（Pangu-Predict-Reg-Table-2.1.0）	

7.2.2 创建预测大模型训练任务

创建预测大模型微调任务

创建预测大模型微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-2 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 预测大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表7-4完成训练参数设置。

   其中，“训练参数”展示了各场景涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 7-4 预测大模型微调参数说明

参数分类	训练参数	说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“预测大模型”，并选择训练所用的基础模型和版本。 **说明** 二次微调的模型来源是“我的资产”。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
训练参数	训练参数	模型训练参数，参考表7-5。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

版权所有 © 华为云计算技术有限公司

表 7-5 预测大模型训练参数说明 - 微调

模型类型	训练参数	说明
盘古统一编码时序预测分类大模型（Pangu-Predict-Cla-TS-2.1.0）	数据集	选择训练所需的数据集。
	非特征列	不作为输入特征的列。此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号。
	预测目标列	预测目标的列名，指定预测目标变量列名，填写格式为中括号，时序分类任务为样本标签列。默认设置为中括号，表示选择最后一列作为预测目标变量。 **说明** 二次微调不支持分类任务的分类数发生改变，需与原资产保持一致，推理服务接口输入字段名称以二次微调为准。
	标识列	用于对连续时间段样本点的区分标识。时序分类任务会将相同ID的行数据视作同一个序列样本，因此时序分类任务必须填写。默认设置为中括号。
	类别特征列	使用LabelEncoder处理的特征列的列表，用于处理字符串类型的类别特征。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示没有需要使用LabelEncoder处理的特征。
	协变量列	用于显式指定协变量例用时序任务建模。如果不填写，则自动选择协变量列。
	训练集切分比例	训练集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。
	验证集切分比例	验证集切分比例。
	测试集切分比例	测试集切分比例。
	历史窗口大小	时序预测输入的窗口长度。数值越大建模包含的历史信息越多，但相应的显存占用会增加以及模型拟合难度也会提升。请根据实际任务特点选择合适的输入窗口，建议取值范围128至512。
	训练轮数	迭代训练的epoch数量，请根据实际情况选择，不低于1，建议范围5至10。
	数据批量大小	单卡的批处理大小，通常来说数据批量越大，梯度会越稳定。但是同时也会使用更大的显存，受硬件限制可能会OOM，并延长单步训练时长。

版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	学习率	学习率用于控制每个训练step参数更新的幅度，一般来说需要选择一个合适的学习率，否则当学习率过大的时候会导致模型难以收敛，过小的时候会收敛速度过慢。
	热身比率	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练。使用热身可以使得模型收敛速度更快，效果更佳。
	权重衰减系数	权重衰减系数，权重衰减是一种根据参数值大小对参数进行衰减的正则化方法，用于减少过拟合，权重衰减系数用于控制正则化的力度，权重衰减系数越大，正则化力度越强。
	模型保存步数	定义模型每隔多少步保存一次，注意训练集总步数必须大于该值。
盘古统一编码时序预测回归大模型（Pangu-Predict-Reg-TS-2.1.0）	数据集	选择训练所需的数据集。
	非特征列	不作为输入特征的列。此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号。
	预测目标列	预测目标的列名，指定预测目标变量列名，填写格式为中括号，时序分类任务为样本标签列。默认设置为中括号，表示选择最后一列作为预测目标变量。 **说明** 二次微调支持列名/变量数发生变化，但回归任务的目标变量数需要与之前保持一致，推理服务接口输入字段名称以二次微调为准。
	标识列	用于对连续时间段样本点的区分标识。时序分类任务会将相同ID的行数据视作同一个序列样本，因此时序分类任务必须填写。默认设置为中括号。
	类别特征列	使用LabelEncoder处理的特征列的列表，用于处理字符串类型的类别特征。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示没有需要使用LabelEncoder处理的特征。
	协变量列	用于显式指定协变量例用时序任务建模。如果不填写，则自动选择协变量列。
	训练集切分比例	训练集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。

版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	验证集切分比例	验证集切分比例。
	测试集切分比例	测试集切分比例。
	特征是否独立建模	时序变量建模模式设置。若选择独立建模，则模型只关注预测目标列的选择的数据，并不做列数据区分，统一视作出自一个序列分布。若选择非独立建模，建模会考虑列变量之间的相关性，建模的变量范围为输入的数据列除去非特征列以及标识列的剩余数据内容。
	历史窗口大小	时序预测输入的窗口长度。数值越大建模包含的历史信息越多，但相应的显存占用会增加以及模型拟合难度也会提升。请根据实际任务特点选择合适的输入窗口，建议取值范围128至512。
	预测目标窗口大小	时序预测输出的窗口长度。数值越大输出预测的时间范围越大，但相应的显存占用会增加以及模型预测精度可能会下降。请根据实际任务特点选择合适的输出窗口，建议取值范围48至96，且尽量选择预测目标周期的整数倍。注若基于已发布模型的进行增量微调，输出窗口需要与其保持一致。 **说明** 二次微调支持输入窗口可变，但输出窗口的长度需要与之前保持一致。
	训练轮数	迭代训练的epoch数量，请根据实际情况选择，不低于1，建议范围5至10。
	数据批量大小	单卡的批处理大小，通常来说数据批量越大，梯度会越稳定。但是同时也会使用更大的显存，受硬件限制可能会OOM，并延长单步训练时长。
	学习率	学习率用于控制每个训练step参数更新的幅度，一般来说需要选择一个合适的学习率，否则当学习率过大的时候会导致模型难以收敛，过小的时候会收敛速度过慢。
	热身比率	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练，使用热身可以使得模型收敛速度更快，效果更佳。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	权重衰减系数	权重衰减系数。权重衰减是一种根据参数值大小对参数进行衰减的正则化方法，用于减少过拟合，权重衰减系数用于控制正则化的力度，权重衰减系数越大，正则化力度越强。
	模型保存步数	定义模型每隔多少步保存一次，注意训练集总步数必须大于该值。
盘古统一编码表格预测分类大模型（Pangu-Predict-Cla-Table-2.2.0）	数据集	选择训练所需的数据集。
	类别特征列	使用LabelEncoder处理的特征列的列表，用于处理字符串类型的类别特征。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示没有需要使用LabelEncoder处理的特征。
	非特征列	不作为输入特征的列。此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号。
	预测目标列	预测目标的列名，指定预测目标变量列名，填写格式为中括号，时序分类任务为样本标签列。默认设置为中括号，表示选择最后一列作为预测目标变量。 **说明** 二次微调不支持分类任务的分类数发生改变，需与原资产保持一致，推理服务接口输入字段名称以二次微调为准。
	训练集切分比例	训练集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。
	验证集切分比例	验证集切分比例。
	测试集切分比例	测试集切分比例。
	训练轮数	迭代训练的epoch数量，请根据实际情况选择，不低于1，建议范围5至10。
	数据批量大小	单卡的批处理大小，通常来说数据批量越大，梯度会越稳定。但是同时也会使用更大的显存，受硬件限制可能会OOM，并延长单步训练时长。
	学习率	学习率用于控制每个训练step参数更新的幅度，一般来说需要选择一个合适的学习率，否则当学习率过大的时候会导致模型难以收敛，过小的时候会收敛速度过慢。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	热身比率	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练，使用热身可以使得模型收敛速度更快，效果更佳。
	权重衰减系数	权重衰减系数。权重衰减是一种根据参数值大小对参数进行衰减的正则化方法，用于减少过拟合，权重衰减系数用于控制正则化的力度，权重衰减系数越大，正则化力度越强。
	模型保存步数	定义模型每隔多少步保存一次，注意训练集总步数必须大于该值。
盘古统一编码表格预测回归大模型（Pangu-Predict-Reg-Table-2.2.0）	与盘古统一编码表格预测分类大模型（Pangu-Predict-Cla-Table-2.2.0）相同。	
盘古时序异常检测大模型（Pangu-Predict-Anom-TS-2.1.0）	数据集	选择训练所需的数据集。
	非特征列	不作为输入特征的列。此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号。
	预测目标列	预测目标的列名，指定预测目标变量列名，填写格式为中括号，其中列举带引号的特征列名。默认设置为中括号，表示选择除去非特征列、标识列、异常标签列以外的所有列为预测目标列。 **说明** 二次微调列名变量数需要与之前保持一致，推理服务接口输入字段名称以二次微调为准。
	异常标签列	记录每个时间点机器运行状态，为二分类列，记录正异常两个状态。默认设置为中括号，表示选择最后一列作为异常标签列。
	正常标签值	指定正常类别名称，用于最终模型评估阶段计算误报率、漏报率等指标。默认设置为0，表示正常类别名称为0。
	标识列	用于对连续时间段样本点的区分标识。时序分类任务会将相同ID的行数据视作同一个序列样本，因此时序分类任务必须填写。默认设置为中括号。
	训练集切分比例	训练集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。

版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	验证集切分比例	验证集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。
	测试集切分比例	测试集切分比例。注意：训练集切分比例、验证集切分比例、测试集切分比例三者之和为1。
	训练轮数	迭代训练的epoch数量，请根据实际情况选择，不低于1，建议范围5至10。
	数据批量大小	单卡的批处理大小，通常来说数据批量越大，梯度会越稳定。但是同时也会使用更大的显存，受硬件限制可能会OOM，并延长单步训练时长。
	学习率	学习率用于控制每个训练step参数更新的幅度，一般来说需要选择一个合适的学习率，否则当学习率过大的时候会导致模型难以收敛，过小的时候会收敛速度过慢。
	热身比率	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练，使用热身可以使得模型收敛速度更快，效果更佳。
	权重衰减系数	权重衰减系数。权重衰减是一种根据参数值大小对参数进行衰减的正则化方法，用于减少过拟合，权重衰减系数用于控制正则化的力度，权重衰减系数越大，正则化力度越强。
	模型保存步数	定义模型每隔多少步保存一次，注意训练集总步数必须大于该值。
盘古融合推荐回归大模型（Pangu-Predict-Reg-Table-2.0.0）	数据集选择	选择训练所需的数据集。
	类别特征列	指定使用LabelEncoder处理的字符串类型类别特征的列表。格式为["列名1","列名2"]，默认设置为[]，表示没有需要处理的类别特征。 LabelEncoder的作用是将类别特征转换为数值型特征，使模型能够处理这些特征。
	非特征列	列出不需要输入到模型中的特征列，用于排除冗余或无意义的特征。格式为["列名1","列名2"]，默认设置为[]，表示所有特征都用于训练。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	标准化列	指定需要进行最大最小值标准化处理的数值特征的列表。格式为["列名1","列名2"]，默认设置为[]，表示没有特征需要标准化。标准化将特征值缩放到0到1的范围，处理分布差异较大的数值特征。
	预测目标列	指定预测目标变量的列名，gatednet算法支持多目标变量预测，其余算法仅支持单目标变量预测。格式为["列名"]，默认设置为[]，表示选择最后一列作为预测目标变量。
	训练集&验证集比例	将数据集划分为训练集和验证集，填入验证集比例即可，默认设置为0.2，即训练集占0.8，验证集占0.2，可选范围为0.1、0.2、0.3、0.4。
	基模型算法池	从预定义的算法池中选择用于训练模型的算法，算法包括：["svm"， "ada"， "lgb"， "xgb"， "rf"， "et"， "gb"， "gauss"， "mlp"，"gatednet"]，其中： ● svm表示支持向量机。 ● ada表示adaboost。 ● lgb表示lightgbm。 ● xgb表示xgboost。 ● rf表示随机森林。 ● et表示extraTree。 ● gb表示梯度提升树。 ● gauss表示高斯过程，gauss适合维度小于10且数据量小于500的样本数据。 ● mlp表示多层感知机，默认设置为5lgb，多种类算法示例: 3lgb,2rf,1xgb（表示使用3个LightGBM算法、2个随机森林算法和1个XGBoost算法）。 ● gatednet表示门控自适应网络，可用于多目标预测。gatednet不支持跟其他算法同时选择，如果同时指定gatednet和其他模型，将只会执行gatednet算法。gatednet仅支持单个生效，例如，5gatednet也只会使用一个gatednet训练。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	推荐的模型个数	从推荐模型中选择的模型个数，指定推荐模型的个数，使得模型的多样性更丰富，有助于提高最终模型的性能。 推荐模型的数量参数的范围是0到20。设置为0表示不使用推荐模型。 假设基模型算法池中有5个LightGBM（lgb）模型，且推荐的模型数量设置为5。这意味着除了基模型池中的5个LightGBM模型外，系统还会再推荐5个不同的模型。因此，总共有10个模型用于训练，其中5个是LightGBM模型，另外5个是系统根据数据情况推荐的不同模型。 **说明** 当基模型算法池包含gatednet时，该参数不生效。
	训练特征重要性	是否在训练完成之后，计算模型的权重特征重要性，并在界面展示各特征的重要性分值及排序。支持权重特征重要性的模型有ada，lgb，xgb，rf，et，gb。当“基模型算法池”中至少配置以上模型中的一个，或“推荐的模型个数”至少为1时，用户可以打开此选项。否则此项无法打开，界面不展示相关信息。 融合推荐回归预测大模型提供两种特征重要性，分别为模型权重特征重要性（训练特征重要性）和排列特征重要性，对两种特征重要性的详细说明参见**排列特征重要性**部分。
盘古融合推荐异常检测大模型（Pangu-Predict-Anom-Table-2.0.0）	数据集选择	选择训练所需的数据集。
	类别特征列	指定使用LabelEncoder处理的字符串类型类别特征的列表。格式为["列名1","列名2"]，默认设置为[]，表示没有需要处理的类别特征。 LabelEncoder的作用是将类别特征转换为数值型特征，使模型能够处理这些特征。
	非特征列	列出不需要输入到模型中的特征列，用于排除冗余或无意义的特征。格式为["列名1","列名2"]，默认设置为[]，表示所有特征都用于训练。
	标准化列	指定需要进行最大最小值标准化处理的数值特征的列表。格式为["列名1","列名2"]，默认设置为[]，表示没有特征需要标准化。标准化将特征值缩放到0到1的范围，处理分布差异较大的数值特征。
	预测目标列	指定预测目标变量的列名，仅支持单目标变量预测。格式为["列名"]，默认设置为[]，表示选择最后一列作为预测目标变量。

版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	训练集&验证集比例	将数据集划分为训练集和验证集，填入验证集比例即可，默认设置为0.2，即训练集占0.8，验证集占0.2，可选范围为0.1、0.2、0.3、0.4。
	基模型算法池	从预定义的算法池中选择用于训练模型的算法，算法包括：["knn", "iforest", "loda", "ocsvm"]，其中： ● knn表示k最近邻算法。 ● iforest表示孤立森林算法。 ● loda表示Loda算法。 ● ocsvm表示单类支持向量机算法。
	推荐的模型个数	从推荐模型中选择的模型个数，指定推荐模型的个数，使得模型的多样性更丰富，有助于提高最终模型的性能。 推荐模型的数量参数的范围是0到20。设置为0表示不使用推荐模型。 假设基模型算法池中有5个LightGBM（lgb）模型，且推荐的模型数量设置为5。这意味着除了基模型池中的5个LightGBM模型外，系统还会再推荐5个不同的模型。因此，总共有10个模型用于训练，其中5个是LightGBM模型，另外5个是系统根据数据情况推荐的不同模型。
盘古融合推荐分类大模型（Pangu-Predict-Cla-Table-2.0.0）	数据集选择	选择训练所需的数据集。
	类别特征列	使用LabelEncoder处理的特征列的列表，用于处理字符串类型的类别特征。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示没有需要使用LabelEncoder处理的特征。
	非特征列	不作为输入特征的列，此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示选取全部特征用于训练。
	标准化列	使用最大最小值标准化处理的特征列的列表，用于处理分布差异较大的数值特征，填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示全部特征都不需要进行标准化。
	预测目标列	预测目标的列名，指定预测目标变量列名，仅支持单目标变量预测，填写格式为中括号，其中列举带引号的特征列名。默认设置为中括号，表示选择最后一列作为预测目标变量。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	训练集&验证集比例	数据集划分为训练集和验证集，填入验证集比例即可，默认设置为0.2，即训练集占0.8，验证集占0.2，可选范围为0.1、0.2、0.3、0.4。
	基模型算法池	从预定义的算法池中选择用于训练模型的算法，算法包括：["ada"，"lgb"，"xgb"，"rf"，"et"，"gb"，"gauss"，"mlp"]，其中： • ada表示adaboost。 • lgb表示lightgbm。 • xgb表示xgboost。 • rf表示随机森林。 • et表示extraTree。 • gb表示梯度提升树。 • gauss表示高斯过程，gauss适合维度小于10且数据量小于500的样本数据。 • mlp表示多层感知机，默认设置为5lgb，多种类算法示例: 3lgb,2rf,1xgb（表示使用3个LightGBM算法、2个随机森林算法和1个XGBoost算法）。
	推荐的模型个数	由推荐模型推荐的模型数，参数的范围为0至20。0代表不使用推荐模型。假设基模型算法池为5lgb，推荐的模型个数为5，表示基模型有5个为lgb，另外5个是由推荐模型推荐的。 假设基模型算法池中有5个LightGBM（lgb）模型，且推荐的模型数量设置为5。这意味着除了基模型池中的5个LightGBM模型外，系统还会再推荐5个不同的模型。因此，总共有10个模型用于训练，其中5个是LightGBM模型，另外5个是系统根据数据情况推荐的不同模型。
	训练特征重要性	是否在训练完成之后，计算模型的权重特征重要性，并在界面展示各特征的重要性分值及排序。支持权重特征重要性的模型有ada, lgb, xgb, rf, et, gb。当“基模型算法池”中至少配置以上模型中的一个，或“推荐的模型个数”至少为1时，用户可以打开此选项。否则此项无法打开，界面不展示相关信息。 融合推荐分类预测大模型提供两种特征重要性，分别为模型权重特征重要性（训练特征重要性）和排列特征重要性，对两种特征重要性的详细说明参见**排列特征重要性**部分。
盘古融合推荐回归大模型-极简模式（Pangu-Predict-Reg-Table-2.1.0）	非特征列	不作为输入特征的列，此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示选取全部特征用于训练。

 版权所有 © 华为云计算技术有限公司

模型类型	训练参数	说明
	类别特征列	使用LabelEncoder处理的特征列的列表，用于处理字符串类型的类别特征。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示没有需要使用LabelEncoder处理的特征。
	标准化列	使用最大最小值标准化处理的特征列的列表，用于处理分布差异较大的数值特征，填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示全部特征都不需要进行标准化。
	预测目标列	预测目标的列名，指定预测目标变量列名，仅支持单目标变量预测，填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号，表示选择最后一列作为预测目标变量。
	训练集&验证集比例	验证集比例。填入验证集比例即可。默认设置为0.2。即训练集占0.8。验证集占0.2。可选范围为0.1, 0.2, 0.3, 0.4。
	训练特征重要性	是否在训练完成之后，计算模型的训练特征重要性。此处提供的特征重要性为排列特征重要性，详细说明参见**排列特征重要性**部分。
	训练模式	代表不同的模型组合和训练策略。Efficiency First相较Default，训练用时减少，但模型效果可能降低。
	数据集	选择训练所需的数据集。
盘古融合推荐分类大模型-极简模式（Pangu-Predict-Reg-Table-2.1.0）	与盘古融合推荐回归大模型-极简模式（Pangu-Predict-Reg-Table-2.1.0）相同。	

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建预测大模型预训练任务

目前时序异常检测大模型支持预训练工作流，创建预测大模型预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 7-3 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 预测大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表7-6**完成训练参数设置。

   其中，“训练参数”展示了各场景涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 7-6 预测大模型预训练参数说明

参数分类	训练参数	说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“预测大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
训练参数	训练参数	模型训练参数，参考**表7-7**。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

版权所有 © 华为云计算技术有限公司

表 7-7 预测大模型训练参数说明 - 预训练

模型类型	训练参数	说明
盘古时序异常检测大模型（Pangu-Predict-Anom-TS-2.1.0）	数据集	选择训练所需的数据集。
	非特征列	不作为输入特征的列。此处填写的特征列名将不作为模型训练的输入特征，不要将预测目标列的列名填入。填写格式为中括号，其中列举带引号的特征列名，默认设置为中括号。
	预测目标列	预测目标的列名，指定预测目标变量列名，填写格式为中括号，其中列举带引号的特征列名。默认设置为中括号，表示选择除去非特征列、标识列以外的所有列为预测目标列。
	标识列	用于对连续时间段样本点的区分标识。时序分类任务会将相同ID的行数据视作同一个序列样本，因此时序分类任务必须填写。默认设置为中括号。
	训练集切分比例	训练集切分比例。注意：训练集切分比例、验证集切分比例两者之和为1。
	验证集切分比例	验证集切分比例。注意：训练集切分比例、验证集切分比例两者之和为1。
	训练轮数	迭代训练的epoch数量，请根据实际情况选择，不低于1，建议范围5至10。
	数据批量大小	单卡的批处理大小，通常来说数据批量越大，梯度会越稳定。但是同时也会使用更大的显存，受硬件限制可能会OOM，并延长单步训练时长。
	学习率	学习率用于控制每个训练step参数更新的幅度，一般来说需要选择一个合适的学习率，否则当学习率过大的时候会导致模型难以收敛，过小的时候会收敛速度过慢。
	热身比率	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练，使用热身可以使得模型收敛速度更快，效果更佳。
	权重衰减系数	权重衰减系数。权重衰减是一种根据参数值大小对参数进行衰减的正则化方法，用于减少过拟合，权重衰减系数用于控制正则化的力度，权重衰减系数越大，正则化力度越强。
	模型保存步数	定义模型每隔多少步保存一次，注意训练集总步数必须大于该值。

版权所有 © 华为云计算技术有限公司

创建预测大模型增量训练任务

在模型完成**创建预测大模型微调任务**后，可以对训练后的模型继续训练，该过程称为“增量训练”。

创建预测大模型增量预训练任务前，请确保有已完成预训练的预测大模型。

创建预测大模型增量预训练任务的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“我的资产 > 预测大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 其余参数配置等步骤同**创建预测大模型微调任务**。

7.2.3 查看预测大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见**表7-8**。

表 7-8 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。

版权所有 © 华为云计算技术有限公司

训练状态	训练状态含义
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见表7-9。

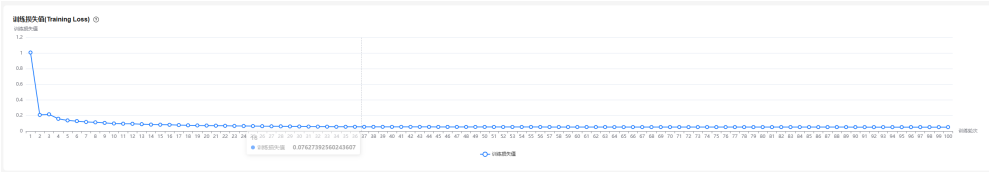
表 7-9 训练指标说明

模型	训练指标	指标说明
预测大模型	拟合度	拟合度是一种衡量模型对数据拟合程度的指标。数值范围为0到1，数值越接近1，表示模型对数据的拟合程度越好。
	均方根误差	均方根误差是预测值与真实值之间差异的平方和的均值的平方根。它用于衡量模型预测值与实际值之间的偏差，数值越小，表明模型预测的精度越高。
	平均绝对误差	平均绝对误差是预测值与真实值之间绝对误差的均值。它同样用于衡量模型预测值与实际值之间的差异，数值越小，表明模型预测的准确性越高。
	真实值和预测值	真实值和预测值在图表中的对比情况。
	准确率	模型预测结果中，所有预测正确的样本占总样本的比例。数值越高，模型效果越好。
	精准率	精准率是指在模型预测为正类的样本中，真正类样本的比例。数值越高，表明模型在检测正类样本时的准确性越高。
	召回率	召回率是指在所有实际为正类的样本中，被模型正确预测为正类的比例。数值越高，表明模型在检测正类样本时的全面性越高。
	误报率	误报率是指被错误预测为异常的正常样本在总正常样本中所占的比例。误报率关乎于不必要的维护操作，高误报率意味着系统频繁错误地预测设备故障，即使设备实际上运行正常，这种情况会导致不必要的维护活动。

版权所有 © 华为云计算技术有限公司

模型	训练指标	指标说明
	漏报率	漏报率是指被错误预测为正常的异常样本在总异常样本中所占的比例。漏报率关乎于错过真正的故障，高漏报率表示实际存在故障时，系统未能正确识别并报警，这意味着真正的故障被忽视了。这种情况下，可能会导致设备损坏加剧、意外停机甚至安全事故的发生。
	训练损失值(Training Loss)	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	详细评估	模型预测数值与给定数据实际值的对比，用于协助用户评估模型训练质量。实际值与预测值差异越少，预测模型训练效果越好。
	训练特征重要性	各特征对模型的影响大小，以便于模型调优。

图 7-5 loss 曲线示意图



获取训练日志

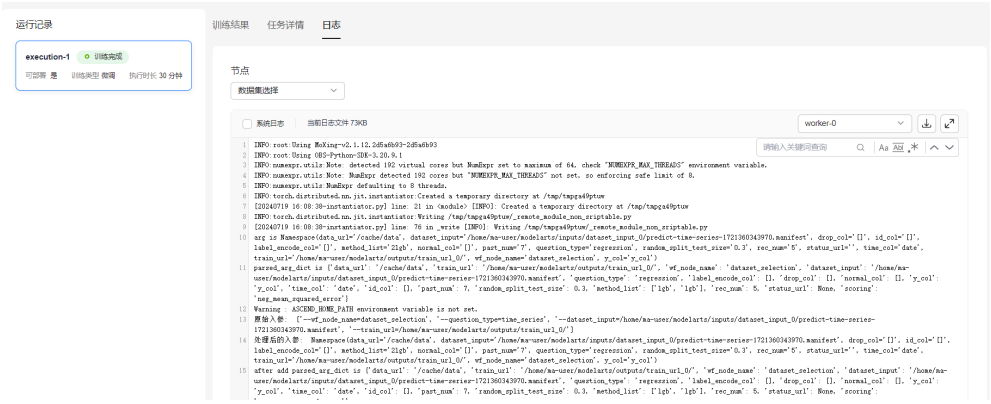
单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**预测大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

版权所有 © 华为云计算技术有限公司

图 7-6 获取训练日志



训练异常定位方法

模型训练失败时，可以按照如下定位思路对训练任务做初步定位。

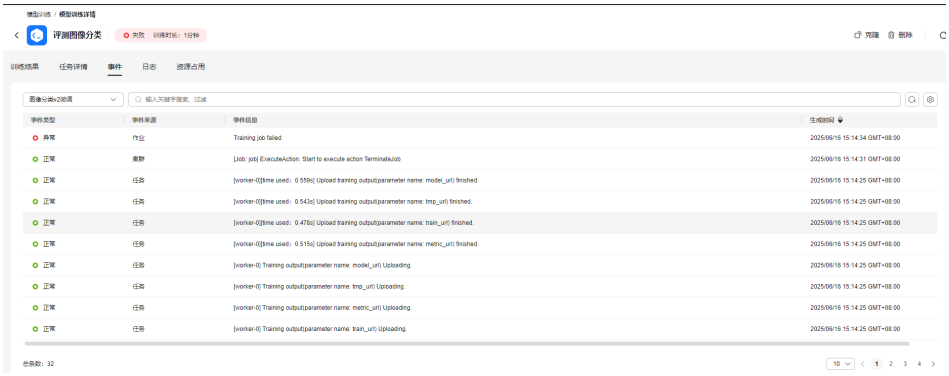
1. 在训练任务列表查看任务失败原因。

   模型启动训练后，可以在模型训练列表中查看训练任务的状态，当任务状态为“失败”时，可将鼠标放置在“失败”上，可以查看到任务失败的概要信息。



2. 单击任务名，进入任务详情页查看训练任务运行关键事件信息。

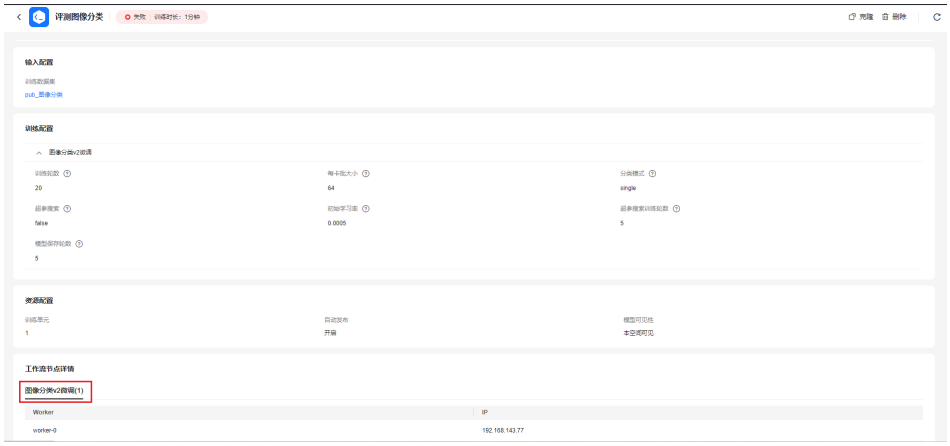
   单击“事件”页签，可查看训练任务子任务运行事件，通过事件可以观察到当前任务的运行情况。事件可以按照“正常、异常、告警”三类事件类型过滤。



3. 获取当前任务运行的worker、workerIP以及任务异常退出时的节点和退出码信息。

   在ModelArts Studio训练任务页面单击需要查看的任务名，进入“任务详情”页，在“工作流节点详情”下可以查看到当前任务所运行的节点及IP信息。

版权所有 © 华为云计算技术有限公司



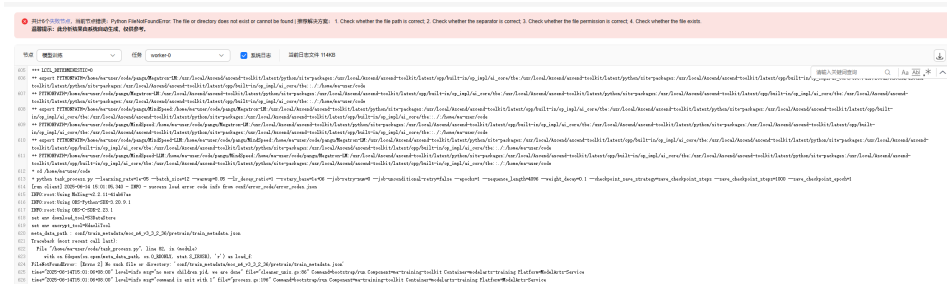
4. 当训练任务运行失败时，可查看“模型训练详情 -> 基本信息 -> 状态”，单击“查看失败节点清单”获取训练任务故障节点及退出码信息。



5. 训练作业遇到问题时，可查看“模型训练详情 -> 日志”，多数场景下的问题可以通过日志报错信息直接定位。

   如果训练作业运行失败，ModelArts Studio会自动识别导致作业失败的原因，在训练日志界面上给出提示并对部分常见训练错误给出分析建议。包含失败的可能原因、推荐的解决方案。

图 7-7 训练错误分析建议



说明

针对分布式作业，只会显示当前节点的一个分析结果，作业的失败需要综合各个节点的失败原因做一个综合判断。

7.2.4 发布训练后的预测大模型

预测大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 7-8 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 7-9 训练结果页面



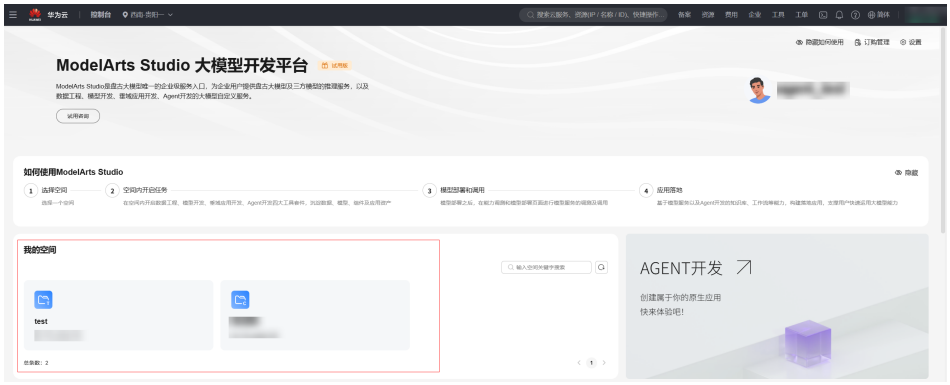
4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

7.2.5 管理预测大模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、启动、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-10 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 克隆。单击操作列的“克隆”，参照**创建预测大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“删除”，可以删除当前不需要的训练任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

7.2.6 预测大模型训练常见报错与解决方案

预测大模型训练常见报错及解决方案请详见**表7-10**。

表 7-10 预测大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
训练后推理服务部署失败，报错“KeyError”	训练后推理报错“KeyError”： KeyError: '0.0'	训练类别特征列选择为数值类型列。	类别特征列需要填字符串的特征或者是需要特征编码的列名，不能填数值类型列。
训练任务报错“ValueError: could not convert string to float”	训练任务报错“ValueError: could not convert string to float”： ValueError: could not convert string to float: '2016/6/27 1:55	训练时选择的特征列包含了非数值类型列。	训练时选择的特征列只能选择数值类型列，非数值类型列需要放到非特征列。
训练日志提示“ValueError: label_map not match”	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。
训练后推理服务部署失败，报错NPU不可用	推理服务启动失败，报错NPU不可用： get device count of NPU failed,torch.npu.is_available() is False	NPU卡未及时释放或者被其他应用占用，导致推理服务器上无推理卡可用。	释放推理服务器上的NPU卡。

版权所有 © 华为云计算技术有限公司

7.3 部署预测大模型

7.3.1 创建预测大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-11 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 预测大模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表7-11**完成部署参数设置。

表 7-11 预测大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“我的资产”。 ● 类型：选择“预测大模型”，并选择需要进行部署的模型和版本。

版权所有 © 华为云计算技术有限公司

参数分类	部署参数	参数说明
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。部署参数说明如下： ● 实例数：设置部署模型时所需的实例数。 ● 若选择“边缘部署”，部署参数说明如下： – 本地挂载路径（可选）： 在容器内部将卷挂载的本地路径。挂载后，容器中的应用程序可以通过这个路径访问宿主机上的数据。 – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

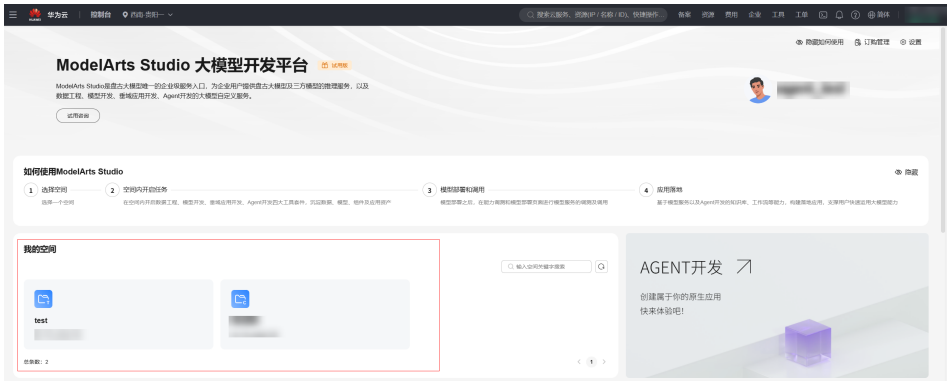
5. 参数填写完成后，单击“立即部署”。

7.3.2 查看预测大模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-12 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建预测大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

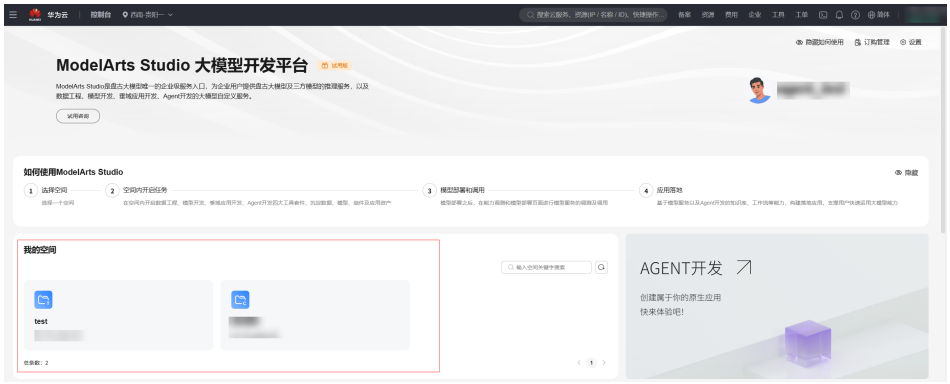
7.3.3 管理预测大模型部署任务

模型更新

完成**创建预测大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

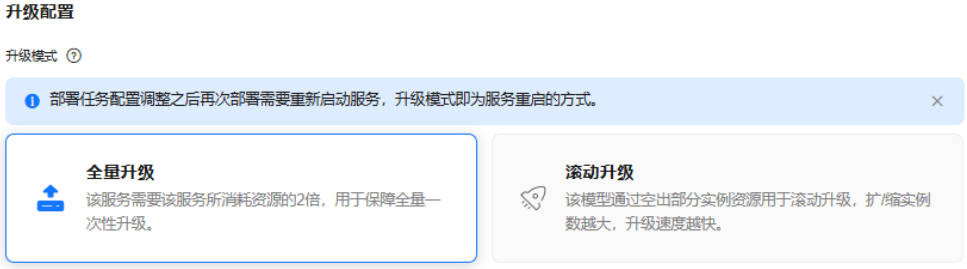
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 7-13 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

   图 7-14 升级模式



 版权所有 © 华为云计算技术有限公司

说明

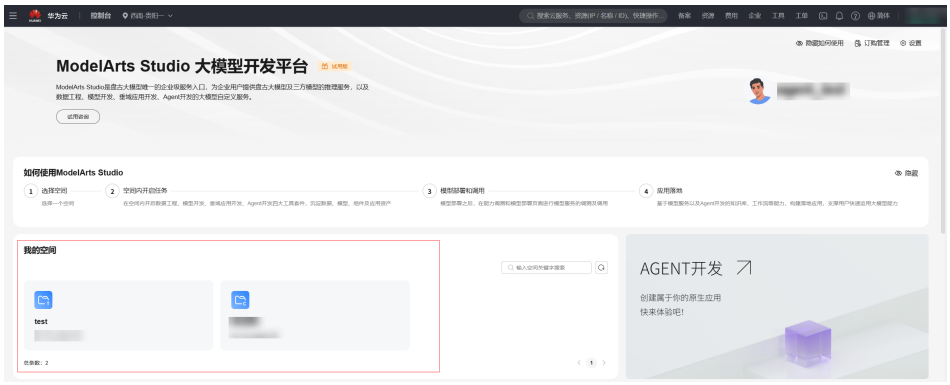
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建预测大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

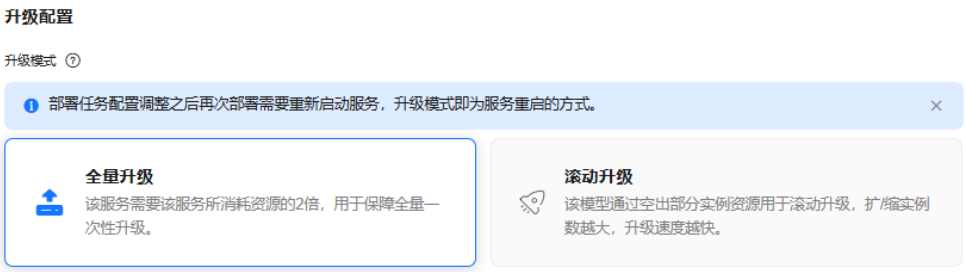
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-15 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 7-16 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

版权所有 © 华为云计算技术有限公司

模型部署实例扩缩容

完成**创建预测大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-17 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-18 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。

版权所有 © 华为云计算技术有限公司

- 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

  说明

  删除属于高危操作，删除前请确保当前任务不再需要。

7.4 评测预测大模型

7.4.1 创建预测大模型评测数据集

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同情景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

预测大模型支持基于规则的自动评测，即基于相似度/准确率进行打分。

评测数据集的创建步骤与训练数据集一致，本章节仅做简单介绍，详细步骤请参见**使用数据工程构建预测大模型数据集**。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

   图 7-19 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面选择所需要的“数据集类型”、“文件格式”、“导入来源”，并上传数据文件。

   预测大模型评测数据集支持的格式见**表7-12**。

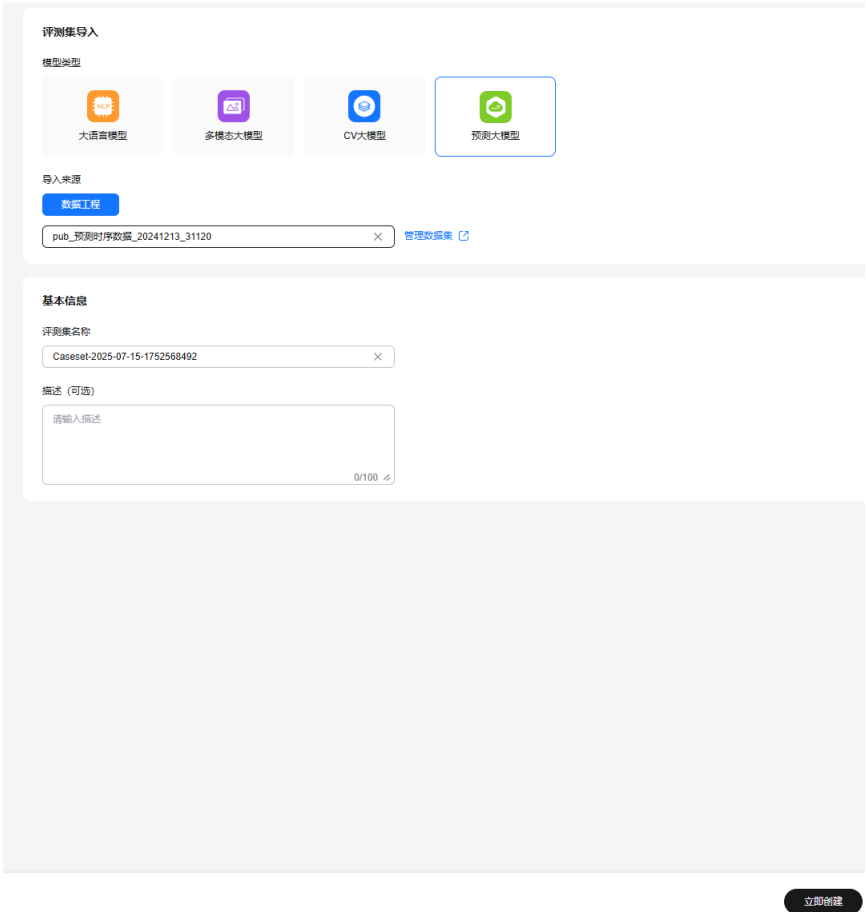
   表 7-12 评测数据集格式

模型类型	评测数据集格式
预测大模型	分类、回归分类格式

4. 上传数据文件后，填写“数据集名称”与“描述”，单击“立即创建”。
5. 在左侧导航栏中选择“数据工程 > 数据发布 > 数据发布”，单击界面右上角“创建发布任务”。

版权所有 © 华为云计算技术有限公司

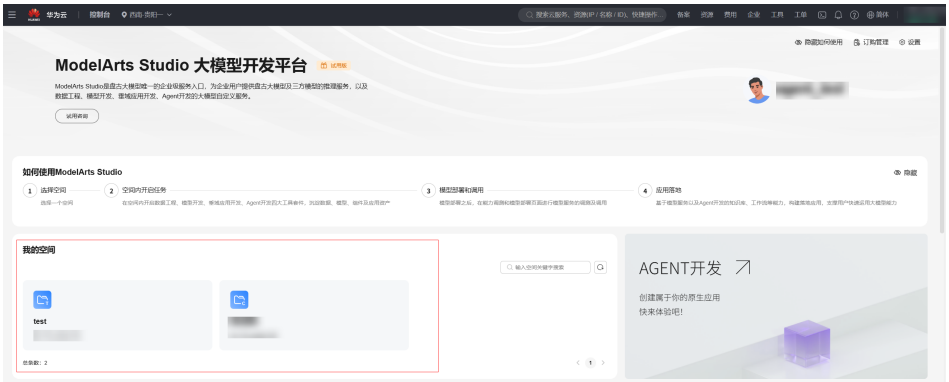
6. 在“创建发布任务”页面选择数据集模态并选择数据集文件。
7. 单击“下一步”，填写名称，选择数据集可见性等信息，单击“确定”。
8. 在左侧导航栏中选择“评测中心 > 评测集管理”，单击界面右上角“创建评测集”。
9. 在“创建评测集”页面选择所需要的“模型类型”、“导入来源”，并选择数据集，数据集来源为数据工程发布的预测数据集，填写评测集名称，最后单击“立即创建”。



7.4.2 创建预测大模型评测任务

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 7-20 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面选择“预测大模型”，参考表7-13完成评测任务参数设置。

表 7-13 创建预测大模型评测任务参数说明

参数分类	参数名称	参数说明
选择服务	评测类型	选择“预测大模型”。
	评测模型	当前支持预测大模型如下模型场景： • 回归 • 分类 • 异常检测
	服务来源	当前仅支持已部署服务 • 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。
评测配置	评测规则	评测规则仅支持“基于规则”。
	评测数据集	待评测的数据集。
	评测指标	待评测模型的评测的指标。根据评测模型的不同，各评测指标也不相同。至少需要选择一项指标评测。具体评测指标可参考表7-14。
	评测结果存储位置	模型评测结果的存储位置。
	命中浮动偏差值	设置命中指标的上下浮动范围。如果某待预测指标在命中指标加减该值的范围，则为命中。当“评测指标”选定“命中率”时需要设置该指标。
	数据集配置	预测目标列：指定预测目标变量列名（注：当前仅支持单变量预测）格式：[""]。如：["result"]，[]表示默认最后一列。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 评测任务参数设置完成后单击“立即创建”。
5. 返回“评测任务 > 自动评测”页面，可查看创建好的评测任务，此时任务状态显示为“初始化”，等到自动评测完成，状态显示为“已完成”时，可以单击操作列“评测报告”查看模型评测结果。在“评测报告”页面可以查看详细的得分以及评测数据信息。

版权所有 © 华为云计算技术有限公司

图 7-21 预测大模型多模型评测效果对比

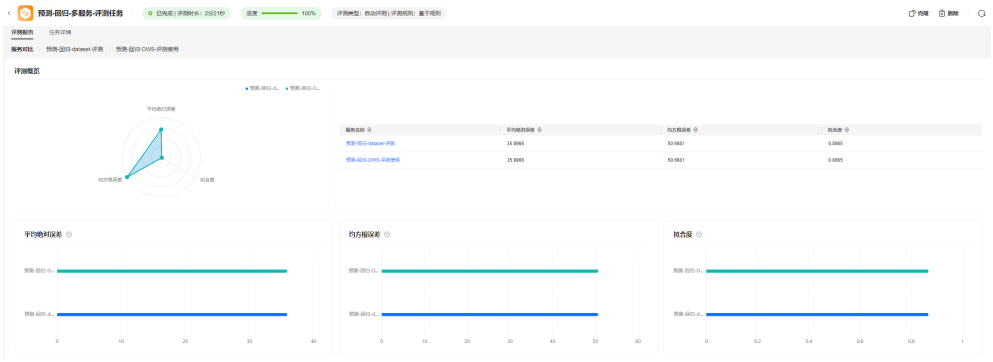
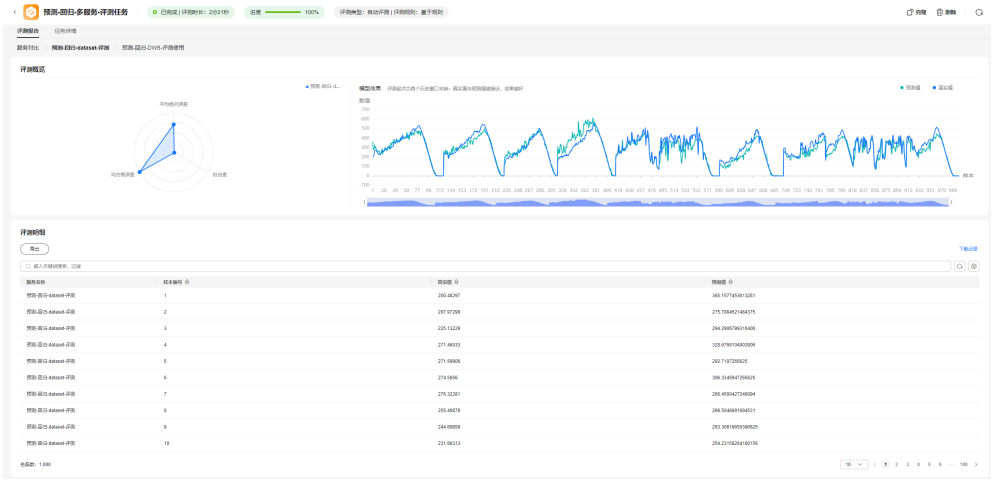


图 7-22 预测大模型回归场景模型评测报告



7.4.3 查看预测大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 7-23 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告及详情。

版权所有 © 华为云计算技术有限公司

其中，各评测指标说明详见**预测大模型评测指标说明**。

4. 导出评测报告。
   a. 在“评测报告 > 服务结果分析”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

预测大模型评测指标说明

表 7-14 预测大模型评测指标说明

模型类型	模型场景	评测指标（自动评测）	指标说明
预测大模型	回归	拟合度	预测值对真实值变动的解释程度，越接近于1，说明模型对真实值的拟合越好。
		平均绝对误差	预测值与真实值之间差异的平均绝对值。它衡量了模型预测的平均偏差程度，值越小表示模型的预测越准确。
		均方根误差	预测值与真实值之间差异的平方和的平均值的平方根。它与MAE类似，但对大误差给予更高的惩罚。RMSE在一定程度上反映了模型预测的波动程度，值越小表示模型的预测越准确。
		命中率	表示预测结果误差在可接受范围内的样本所占的比例，命中率接近1，表示效果越好。
	分类	F1分数	精确率和召回率的调和平均数，分数越高表示模型在这两个指标上表现越好，即模型在精确率和召回率之间取得了更好的平衡。
		准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
		精确率	预测为正的样本中，实际为正的比率。
		召回率	实际为正的样本，预测为正的比率。
	异常检测	F1分数	精确率和召回率的调和平均数，分数越高表示模型在这两个指标上表现越好，即模型在精确率和召回率之间取得了更好的平衡。

版权所有 © 华为云计算技术有限公司

模型类型	模型场景	评测指标（自动评测）	指标说明
		准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
		精确率	预测为正的样本中，实际为正的比率。
		召回率	实际为正的样本，预测为正的比率。
		误报率	衡量检测系统或分类模型在负样本中错误判断为正样本的比例。其核心意义在于反映系统对正常数据的误判程度。

7.4.4 管理预测大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 7-24 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

版权所有 © 华为云计算技术有限公司

7.5 调用预测大模型

表 7-15 预测大模型 API 清单

API分类	API访问路径（URI）
推理	/
特征重要性	/perm-feat-importance

7.5.1 使用 API 调用预测大模型

获取预测大模型 API 请求地址

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 7-25 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击“模型名称”在“详情”页面获取API请求地址。

只有处于“运行中”状态的模型才可以正常调用。

云上部署：

图 7-26 云上部署获取预测大模型 API 请求地址



版权所有 © 华为云计算技术有限公司

边缘部署：

边缘部署-负载均衡访问方式，获取预测大模型API请求地址由三部分组成

图2 边缘部署负载均衡访问时，api请求地址获取

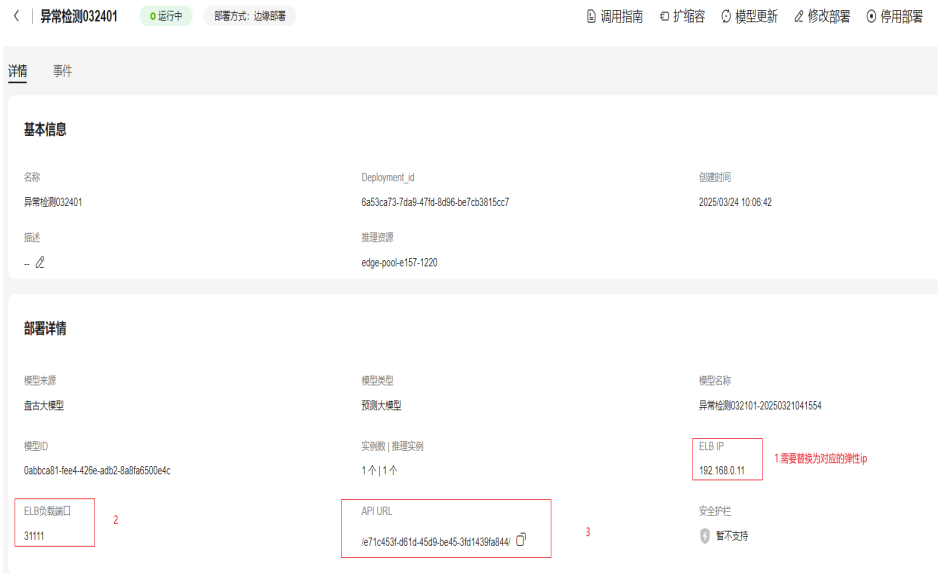


图3 在华为云前端界面搜索ECS，进入弹性云服务器界面

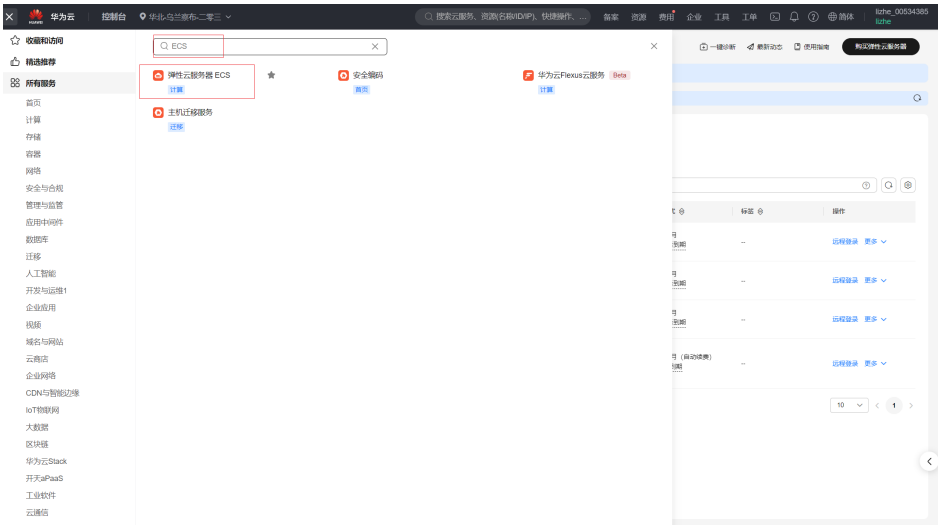
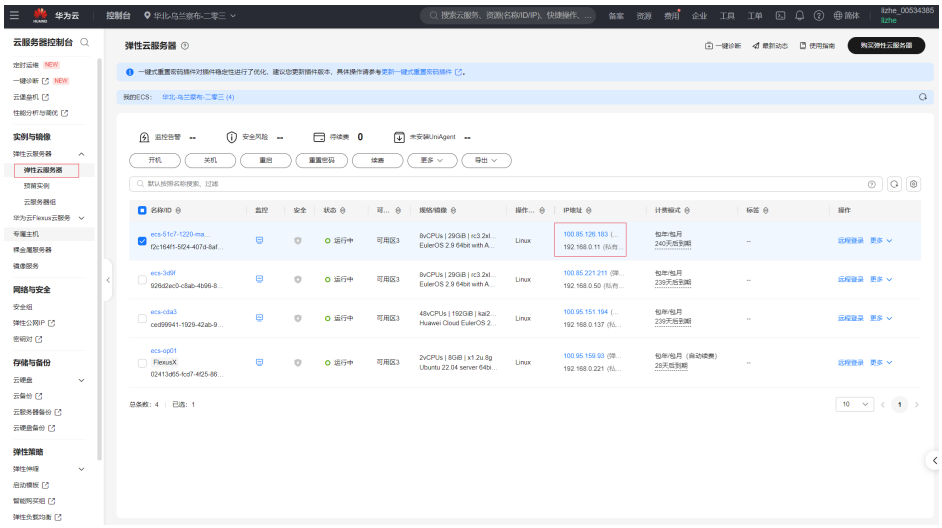


图4 在弹性云服务器界面下，寻找到私有ip对应的弹性ip，如192.168.0.11->100.85.126.183

版权所有 © 华为云计算技术有限公司



拼接弹性ip、图2中的负载端口、图2中的api url，得到请求url地址。

获取 Token

参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。

使用 API 调用预测大模型

1. 在Postman中新建POST请求，并填入API请求地址。
2. 参考如图填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**获取Token**中获取的Token值。

   图 7-27 填写预测大模型 API

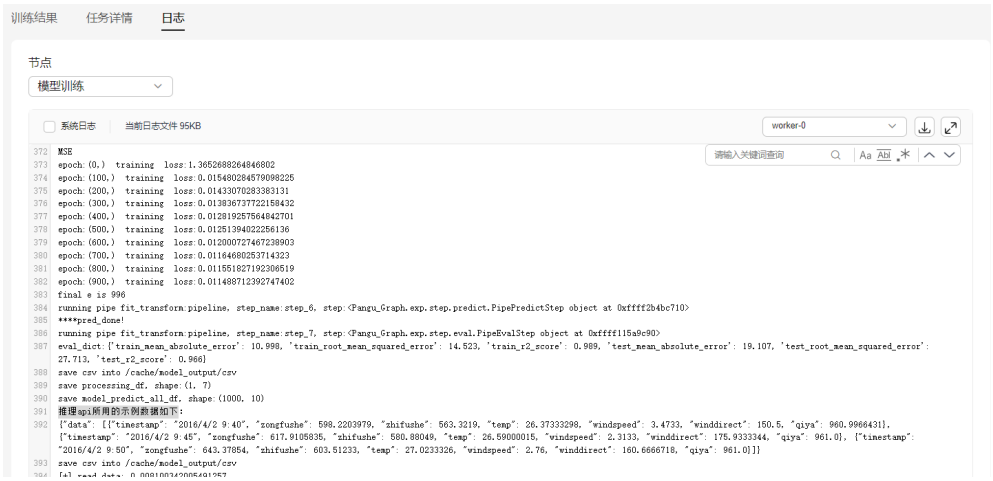


3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。feature表示待预测数据中的特征名称，需要与训练数据中的特征名称保持一致，同时特征数量需要与训练数据保持一致。

   预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。

版权所有 © 华为云计算技术有限公司

图 7-28 获取推理 api 所用的示例数据




```
{
 "data": [
  {
   "feature_1": xx,
   "feature_2": xx,
   ...
   "feature_n": xx
  },
  ...
  {
   "feature_1": xx,
   "feature_2": xx,
   ...
   "feature_n": xx
  }
 ]
}
```

4. 单击Postman界面“Send”按钮，发送请求。

排列特征重要性

盘古融合推荐回归大模型和盘古融合推荐分类大模型在训练阶段提供训练特征重要性（模型**权重特征重要性**），还支持在推理阶段计算排列特征重要性，详见《API参考》“API > 模型推理接口 > 预测大模型 >盘古融合推荐分类大模型 > 排列特征重要性”和“API > 模型推理接口 > 预测大模型 > 盘古融合推荐回归大模型 > 排列特征重要性”章节。（注意，这两种特征重要性功能独立，即训练阶段的权重特征重要性是否打开，不影响推理阶段的排列特征重要性计算）

对于模型权重特征重要性（训练特征重要性）和排列特征重要性这两种特征重要性，说明如下：

1. 模型权重特征重要性（训练特征重要性）

   使用方式：创建预测大模型训练任务时，打开“训练特征重要性”开关，训练完成后在“训练结果”>“训练特征重要性”查看。

   原理简述：对于随机森林、XGBoost等基于树模型的模型，衡量每个特征对决策树构建的参与程度（具体而言，指该特征作为分裂特征的次数，以及作为分裂节点时的信息增益）。

   方法特点：计算快、效率高，使用简单。当特征为连续的数值数据，或为取值较多的类别特征数据时，获得的重要性分数可能偏高。此外，仅有部分基模型（ada，lgb，xgb，rf，et，gb）支持此重要性计算，最终重要性值为这些基模型

 版权所有 © 华为云计算技术有限公司

的重要性均值，因此，当用户配置了较多其他基模型时，重要性值的准确度可能会下降。

2. 排列特征重要性

   使用方式：准备一批包含目标列真值的数据，调用推理服务的排列特征重要性接口，在返回结果中查看。

   原理简述：衡量当某个特征的值被随机打乱后，模型效果下降的程度。模型效果下降越多，表明特征对模型效果影响越大，特征越重要。

   方法特点：可靠性更高，且不受特征是否是连续或多类别的影响，但如果特征之间存在高度相关性，可能会导致特征的重要性被低估。

⚠ 注意

盘古融合推荐回归/分类大模型-极简模式在训练阶段提供的训练特征重要性为排列特征重要性，从用户数据中随机采样数据进行计算。在推理阶段不提供特征重要性的API接口。

预测置信度

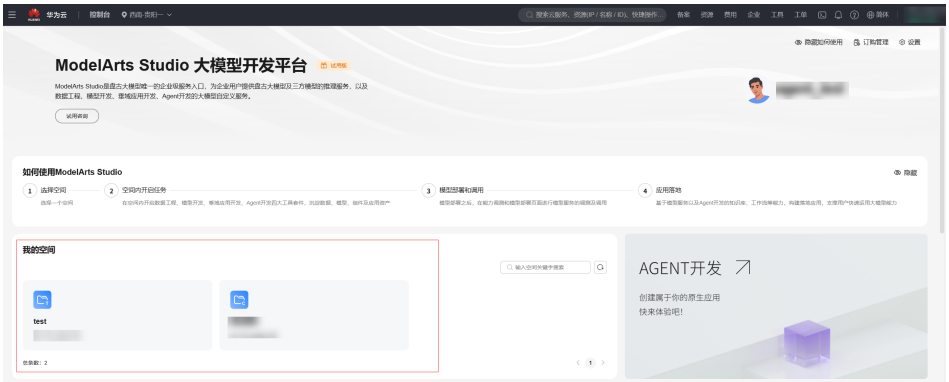
盘古融合推荐异常检测大模型和盘古融合推荐分类大模型支持预测时输出置信度（即每种结果的概率值），详见《API参考》“API > 模型推理接口 > 预测大模型 > 盘古融合推荐异常检测大模型”和“API > 模型推理接口 > 预测大模型 > 盘古融合推荐分类大模型”。

7.5.2 统计预测大模型调用信息

模型推理接口除了使用Token认证鉴权进行调用外，也支持使用API Key鉴权进行调用。对于API Key鉴权产生的调用信息，可以使用平台的“调用统计”功能对调用量信息、时延信息进行统计。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

   图 7-29 我的空间



2. 单击左侧导航栏“调用统计”。
3. 选择当前调用的预测大模型，可以按照不同时间跨度查看当前模型调用信息。

版权所有 © 华为云计算技术有限公司

8 开发盘古科学计算大模型

气象/海洋数据获取指南
使用数据工程构建科学计算大模型数据集
训练科学计算大模型
部署科学计算大模型
调用科学计算大模型

 版权所有 © 华为云计算技术有限公司

8.1 气象/海洋数据获取指南

表 8-1 科学计算大模型训练数据要求

数据类型	常用特征/层次	水平分辨率	数据范围	数据获取方式
ERA5	地表变量：海平面气压，10米u风，10米v风，2m温度，相对湿度 海表变量：有效波高，平均波周期，平均波向 高空变量：重力位势、u风、v风、比湿、温度 高空层次：1000、925、850、700、600、500、400、300、250、200、150、100、50hPa共13高空层次。	1/4°	全球范围 纬度：90N~-90S 经度：0W~360E	训练数据一般可通过公开数据集获取，例如ERA5。ERA5是由欧洲中期天气预报中心（ECMWF）提供的全球气候的第五代大气再分析数据集，它覆盖从1940年1月至今的时间段，提供每小时的大气、陆地和海洋气候变量的估计值。 • ERA5数据下载官方指导：**https://confluence.ecmwf.int/display/CKB/How+to+download+ERA5** • 高空变量数据下载链接：**https://cds.climate.copernicus.eu/datasets**，查找名称中包含ERA5和pressure levels的数据集。 • 表面变量数据下载链接：**https://cds.climate.copernicus.eu/datasets**，查找名称中包含ERA5和single levels的数据集。

版权所有 © 华为云计算技术有限公司

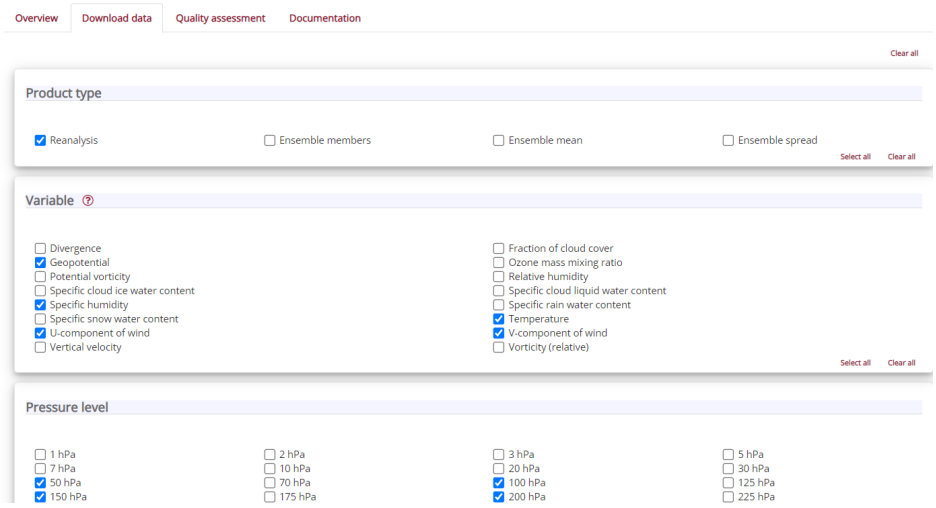
数据类型	常用特征/层次	水平分辨率	数据范围	数据获取方式
EC0	地表变量：海平面气压，10米u风，10米v风，2m温度，相对湿度 海表变量：有效波高，平均波周期，平均波向 高空变量：重力位势、u风、v风、比湿、温度 高空层次：1000、925、850、700、600、500、400、300、250、200、150、100、50hPa共13高空层次。	1/4°	全球范围 纬度：90N~-90S 经度：0W~360E	EC0数据是指ECMWF实时预测中的零时刻数据，作为模型的初始场进行推理。 • EC0数据详细解释：**ECMWF open data: real-time forecasts from IFS and AIFS - Data and Charts - ECMWF Confluence Wiki**
GLORYS12V1	海表变量：海平面高度 深海变量：海温，海盐，海流经向速率，海流纬向速率 深海层次：*0.49 m→5727.92 m*共50层	1/12°	全球范围 纬度：90N~-80S 经度：0W~360E	GLORYS12V1是由哥白尼海洋环境监测服务（Copernicus Marine Service）发布的全球海洋物理再分析数据集。它是目前分辨率最高的全球海洋再分析数据之一。覆盖1993年1月至2021年6月的时间段，提供逐日的海洋变量再分析数据。 • GLORYS12V1数据下载官方指导：**Copernicus Marine Toolbox CLI - Subset | Copernicus Marine Help Center** • GLORYS12V1数据详细解释：**CMEMS-GLO-PUM-001-030.pdf (copernicus.eu)**

版权所有 © 华为云计算技术有限公司

ERA5 数据获取指导

1. 示例一：以下载2021年7月16日高空变量数据为例，下载内容为高空变量（重力位势、u风、v风、比湿、温度，1000、925、850、700、600、500、400、300、250、200、150、100、50hPa高空层次）0点、6点、12点、18点时刻的数据文件，下载步骤示例如下：
   a. 注册并登录数据下载平台，在高空变量数据下载链接中：
      - Product type选择Reanalysis。
      - Variable新选择Geopotential、Specific humidity、Temperature、U-component of wind、V-component of wind。
      - Pressure level选择1000hPa、925hPa、850hPa、700hPa、600hPa、500hPa、400hPa、300hPa、250hPa、200hPa、150hPa、100hPa、50hPa。
      - Year选择2021，Month选择July，Day选择16。
      - Time选择00:00、06:00、12:00、18:00。
      - Geographical area选择Whole available region。
      - Format选择NetCDF(experimental)。
   b. 数据准备好后，单击“Submit Form”，基于页面提示单击“Download”下载数据。

   图 8-1 下载高空变量数据



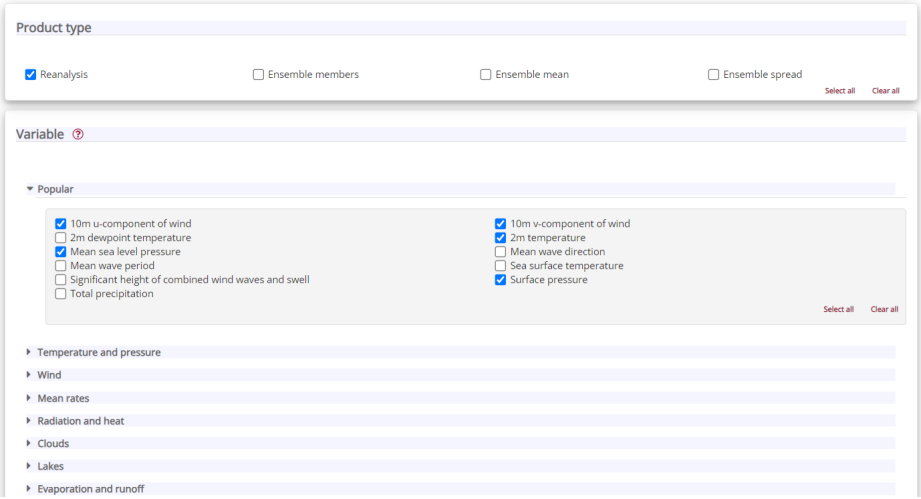
2. 示例二：以下载2021年7月16日表面变量数据为例，下载内容为表面变量（10m u风、10m v风、2米温度、海平面气压）0点、6点、12点、18点时刻的数据文件，下载步骤示例如下：
   a. 注册并登录数据下载平台，在表面变量数据下载链接中：
      - Product type选择Reanalysis。
      - Popular选择10m u-component of wind、10m v-component of wind、2m temperature、Mean sea level pressure，Surface pressure。

 版权所有 © 华为云计算技术有限公司

- Year选择2021，Month选择July，Day选择16。
- Time选择00:00、06:00、12:00、18:00。
- Geographical area选择Whole available region。
- Format选择NetCDF(experimental)。

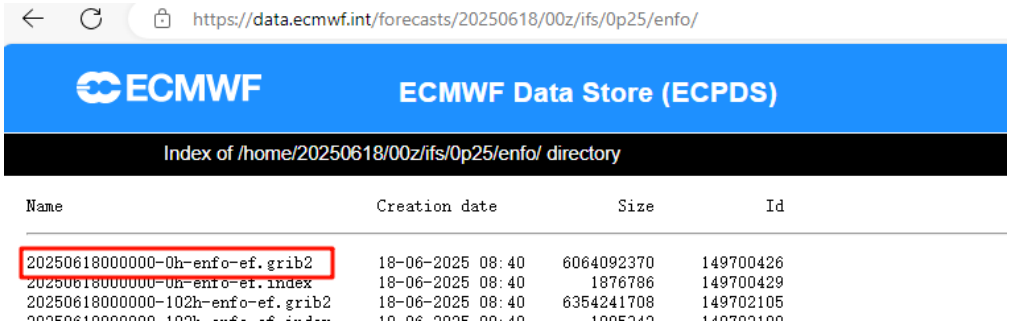
b. 数据准备好后，单击“Submit Form”，基于页面提示单击“Download”下载数据。

图 8-2 下载表面变量数据



EC0 数据获取指导

1. 进入**ECMWF Data Store (ECPDS)**后，选取对应文件夹，依次为起报日期/起报时刻/预报方法（选择ifs）/预报变量（选择enfo，海表变量选择wave）

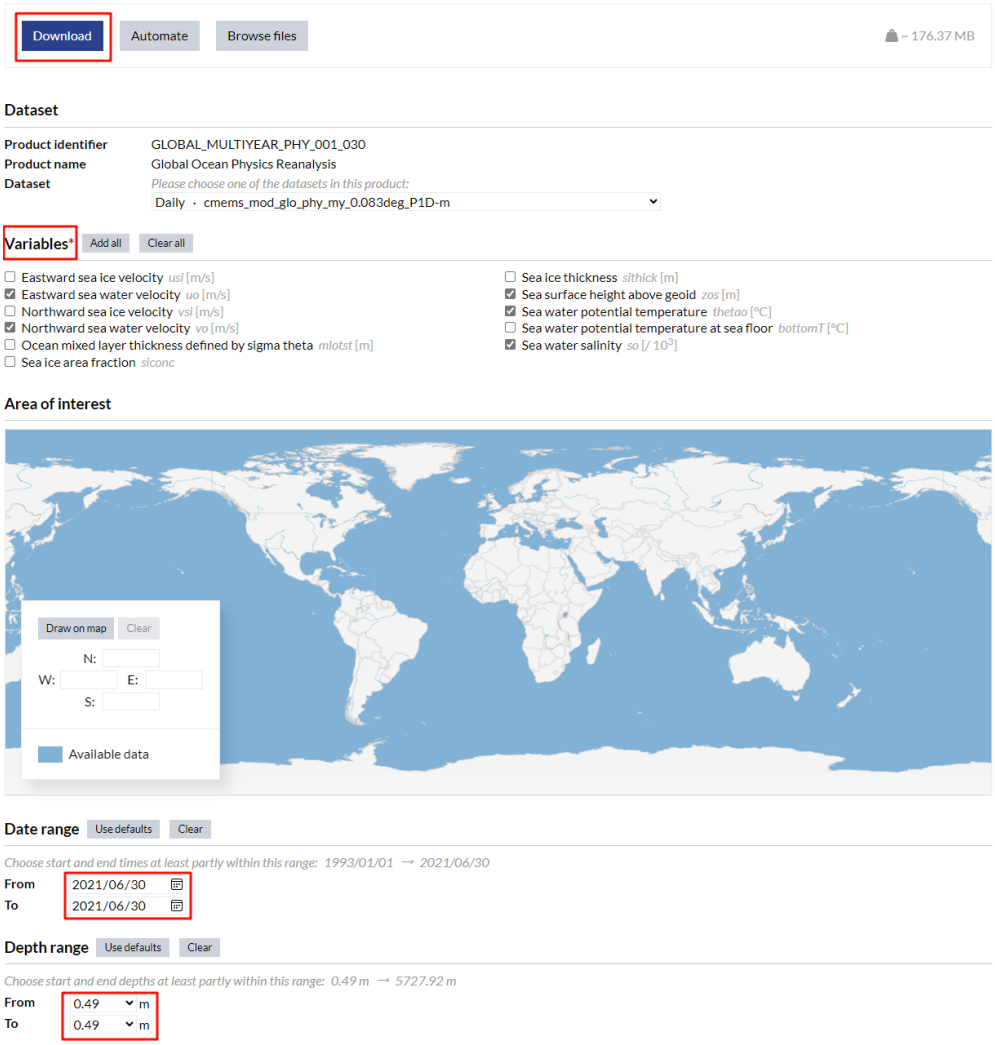


Mecator 数据获取指导

1. 方法一：以下载2021年6月30日深海变量数据为例，下载内容为深海变量（海温，海盐，海流经向速率，海流纬向速率，*0.49 m*），海表变量（海平面高度），下载步骤示例如下：
   - Depth range：选择需要的深海层，此处为0.49m。
   - Date range：选择需要的时间范围，此处为2021年6月30日。
   - Variables：选择需要的变量，此处为Eastward sea water velocity，Northward sea water velocity，Sea surface height above geoid，Sea water potential temperature，Sea water salinity。

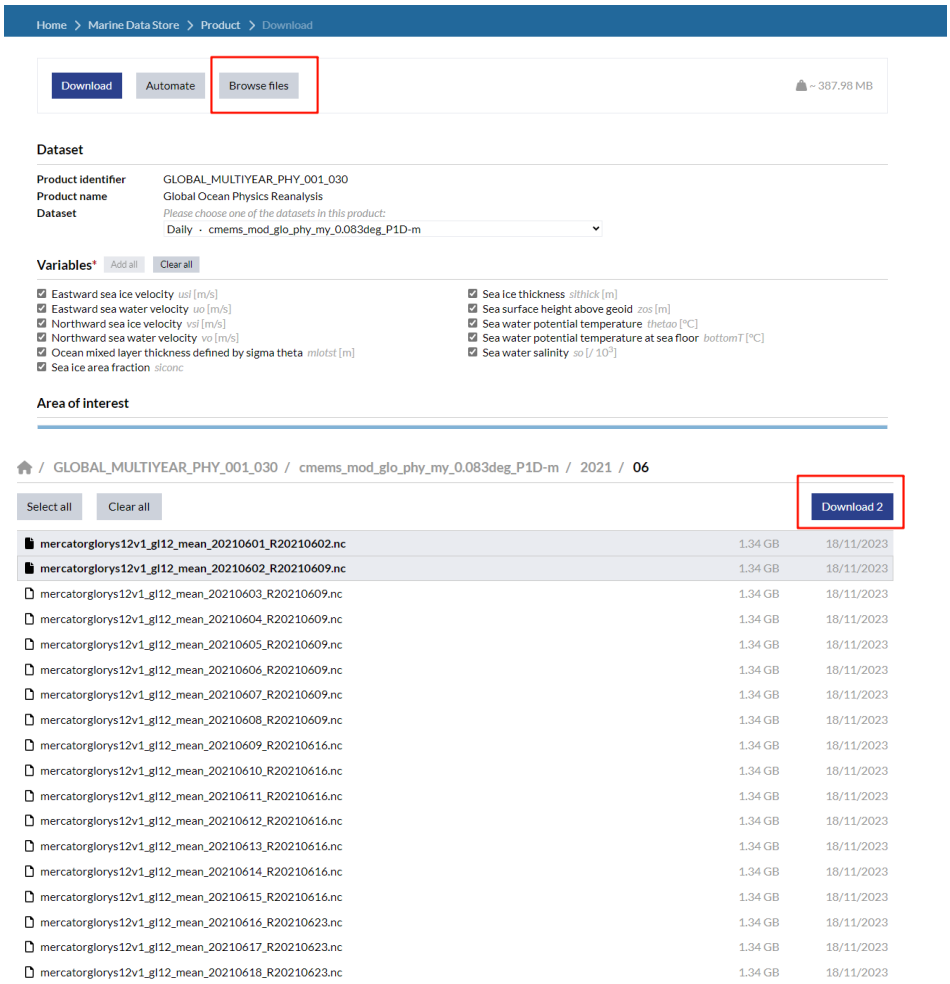
版权所有 © 华为云计算技术有限公司

- 选择完后单击Download即可下载，若文件大小超过限制，则推荐使用方法二，三。



2. 方法二：以下载2021年6月30日深海变量数据为例，下载内容为深海变量（海温，海盐，海流经向速率，海流纬向速率，*0.49 m*），海表变量（海平面高度），下载步骤示例如下：
   - 按照方法一选取所需要的数据。
   - 单击Automate，复制对应的Python API，使用python进行数据下载。
3. 方法三：以下载2021年6深海变量数据为例，下载内容为深海变量（海温，海盐，海流经向速率，海流纬向速率，*0.49 m→5727.92 m*共50层，海表变量（海平面高度），下载步骤示例如下：
   - 单击Browse files进入数据目录，选择对应的数据日期。
   - 选取对应的日期文件后单击Download进行下载。

 版权所有 © 华为云计算技术有限公司



8.2 使用数据工程构建科学计算大模型数据集

科学计算大模型支持接入的数据集类型

盘古科学计算大模型仅支持接入气象类数据集，该数据集格式要求请参见**气象类数据集格式要求**。

训练科学计算大模型训练数据要求所需数据量

构建科学计算大模型进行训练的数据要求见**表8-2**。

版权所有 © 华为云计算技术有限公司

表 8-2 科学计算大模型训练数据要求

模型类别	数据要求	特征要求	水平分辨率要求	区域范围要求	时间要求	训练数据获取方式	推理数据获取方式
气象模型	天气数据	需包含4个表面层特征（10m u风、10m v风、2米温度、海平面气压），13高空层次（1000、925、850、700、600、500、400、300、250、200、150、100、50hPa）的5个高空层特征（重力位势、u风、v风、比湿、温度）。	25km*25km。	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	参考**ERA5数据获取指导**	参考**EC0数据获取指导**

版权所有 © 华为云计算技术有限公司

模型类别	数据要求	特征要求	水平分辨率要求	区域范围要求	时间要求	训练数据获取方式	推理数据获取方式
污染物模型	天气数据	1、全球气象数据，需包含4个表面层特征（10m u风、10m v风、2米温度、海平面气压），13高空层次（1000、925、850、700、600、500、400、300、250、200、150、100、50hPa）的5个高空层特征（重力位势、u风、v风、比湿、温度）。 2、区域污染物数据，需包含6个表面层特征（PM10、PM2.5、NO_2、O_3、SO_2、CO）	0.0625°*0.0625°	中国区域，经度72.25°E~136.0°E,纬度17.25°N~55.0°N	训练集和验证集均推荐使用>1个月的历史数据。	1、全球气象数据，可参考第一行气象/降水模型中的数据获取方式。 2、污染物观测数据由客户自行提供。	-

 版权所有 © 华为云计算技术有限公司

模型类别	数据要求	特征要求	水平分辨率要求	区域范围要求	时间要求	训练数据获取方式	推理数据获取方式
海气模型	天气+海洋数据	1、全球气象数据，需包含4个表面层特征（10m u风、10m v风、2米温度、海平面气压），13高空层次（1000、925、850、700、600、500、400、300、250、200、150、100、50hPa）的5个高空层特征（重力位势、u风、v风、比湿、温度）。 2、全球海洋数据，需包含1个表面层特征（海表高度），15个深海层次（0.5m,6.4m,9.6m,21.6m,29.4m,47.4m,65.8m,92.3m,130.7m,155.9m,186.1m,266m,318.1m,380.2m,541.1m）的4个深海层特征（海盐、海洋流速u、海洋流速	天气：1/4° 海洋：1/12°	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	参考**Mecator数据获取指导**	-

版权所有 © 华为云计算技术有限公司

模型类别	数据要求	特征要求	水平分辨率要求	区域范围要求	时间要求	训练数据获取方式	推理数据获取方式
		v、温度）。					
海浪模型	海洋数据	1、全球海洋数据，需包含3个表面层特征（有效波高，平均波周期，平均波向）。 2、全球海洋数据，需包含1个表面层特征（海表高度），15个深海层次（0.5m,6.4m,9.6m,21.6m,29.4m,47.4m,65.8m,92.3m,130.7m,155.9m,186.1m,266m,318.1m,380.2m,541.1m）的4个深海层特征（海盐、海洋流速u、海洋流速v、温度）。	1/4°。	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	参考**ERA5数据获取指导**	-

版权所有 © 华为云计算技术有限公司

模型类别	数据要求	特征要求	水平分辨率要求	区域范围要求	时间要求	训练数据获取方式	推理数据获取方式
海洋生态模型	海洋数据	1、全球海洋数据，需包含9个表面层特征（总叶绿素浓度、叶绿素浓度、硅藻浓度、颗石藻浓度、蓝澡浓度、铁浓度、硝酸盐浓度 、混合层深度、磷酸盐浓度）。 2、全球海洋数据，需包含1个表面层特征（海表高度），15个深海层次（0.5m,6.4m,9.6m,21.6m,29.4m,47.4m,65.8m,92.3m,130.7m,155.9m,186.1m,266m,318.1m,380.2m,541.1m）的4个深海层特征（海盐、海洋流速u、海洋流速v、温度）。	1/4°。	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	-	-

构建科学计算大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程创建盘古科学计算大模型数据集流程见**表8-3**。

版权所有 © 华为云计算技术有限公司

表 8-3 盘古科学计算大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
加工气象类数据集	加工气象类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	加工气象类数据集
发布气象类数据集	发布气象类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。	发布气象类数据集

8.3 训练科学计算大模型

8.3.1 科学计算大模型训练流程与选择建议

科学计算大模型训练流程介绍

科学计算大模型的训练主要分为两个阶段：预训练与微调。

- **预训练阶段**：预训练是模型学习基础知识的过程，基于大规模通用数据集进行。例如，在区域海洋要素预测中，可以重新定义深海变量、海表变量，调整深度层、时间分辨率、水平分辨率以及区域范围，以适配自定义区域的模型场景。此阶段需预先准备区域的高精度数据。
- **微调阶段**：在预训练模型的基础上，微调利用特定领域的数据进一步优化模型，使其更好地满足实际任务需求。例如，区域海洋要素预测的微调是在已有模型上添加最新数据，不改变模型结构参数或引入新要素，以适应数据更新需求。

在实际流程中，通过设定训练指标对模型进行监控，以确保效果符合预期。在微调后，评估用户模型，并进行最终优化，确认其满足业务需求后，进行部署和调用，以便实际应用。

科学计算大模型选择建议

科学计算大模型支持训练的模型类型有：全球中期天气要素模型、区域中期海洋智能预测模型。

- **全球中期天气要素预测模型选择建议**：

  科学计算大模型的全球中期天气要素预测模型，可以对未来一段时间的天气进行预测，具备以下优势：

  - 高时间精度：全球中期天气要素预测模型可以预测未来1、3、6、24小时的天气情况，降水模型可预测未来6小时的降水情况。高时间精度对于农业、交通、能源等领域的决策和规划非常重要。

 版权所有 © 华为云计算技术有限公司

- 全球覆盖：全球中期天气要素预测模型和降水模型能够在全球范围内进行预测，不仅仅局限于某个地区。它的分辨率相当于赤道附近每个点约25公里*25公里的空间。
- 数据驱动：全球中期天气要素预测模型和降水模型使用历史天气数据来训练模型，从而提高预测的准确性。这意味着它可以直接利用过去的观测数据，而不仅仅依赖于数值模型。

全球中期天气要素预测模型、降水模型信息见**表 全球中期天气要素预测模型信息表**。

表 8-4 全球中期天气要素预测模型信息表

模型	预报层次	预报高空变量	预报表面变量	时间分辨率	水平分辨率	区域范围
全球中期天气要素预测模型	13层（1000hpa，925hpa，850hpa，700hpa，600hpa，500hpa，400hpa，300hpa，250hpa，200hpa，150hpa，100hpa，50hpa）	T：温度 Q：比湿 Z：重力位势 U：U风 V：V风	MLSP：海平面气压 U10：10米U风，经度方向 V10：10米V风，纬度方向 T2M：2米温度	1、3、6、24小时	0.25°*0.25°	全球

支持训练的模型清单见**表 中期天气要素预测模型的类型**，您可根据具体使用场景选择合适的模型。例如天气基础要素预测，需要时间分辨率为1小时的场景下，您可以选择**Pangu-AI4S-Global-Weather-v1.1.0-1h**模型。

版权所有 © 华为云计算技术有限公司

表 8-5 中期天气要素预测模型

模型名称	模型版本及时间分辨率	说明
Pangu-AI4S-Global-Weather	v1.1.2-1h	2025年3月发布的版本，用于天气基础要素预测，时间分辨率为1小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Global-Weather	v1.2.2-3h	2025年3月发布的版本，用于天气基础要素预测，时间分辨率为3小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Global-Weather	v1.3.2-6h	2025年3月发布的版本，用于天气基础要素预测，时间分辨率为6小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Global-Weather	v1.4.2-24h	2025年3月发布的版本，用于天气基础要素预测，时间分辨率为24小时，1个训练单元起训及1个实例部署。

- **中期海洋智能预测模型选择建议**：

  科学计算大模型的中期海洋智能预测模型，可以对未来一段时间海洋要素进行预测。可为海上防灾减灾，指导合理开发和保护渔业等方面有着重要作用。中期海洋智能预报可以训练的模型为区域海洋要素模型，信息见**表8-6**。

表 8-6 中期海洋智能预测模型信息

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
区域海洋要素模型	0m，6m，10m，20m，30m，50m，70m，100m，125m，150m，200m，250m，300m，400m，500m	T：海温(℃) S：海盐(PSU) U：海流经向速率(ms-1) V：海流纬向速率(ms-1)	MLSP：海平面气压 U10：10米U风，经度方向 V10：10米V风，纬度方向 T2M：2米温度 SSH：海表高度(m)	24h	0.25°*0.25°	特定区域

版权所有 © 华为云计算技术有限公司

支持训练的模型清单见表8-7，您可根据具体使用场景选择合适的模型。例如区域海洋基础要素预测场景下，您可以选择**Pangu-AI4S-Regional-Ocean-v1.1.0-24h**模型。

表 8-7 区域中期海洋智能预测模型的类型

模型名称	模型版本及时间分辨率	说明
Pangu-AI4S-Regional-Ocean	v1.1.2-24h	2025年3月发布的版本，用于区域海洋基础要素预测，1个训练单元起训及1个实例部署。

- **区域大气污染物预测模型选择建议**：

  科学计算大模型的区域大气污染物预测模型，可以对特定区域未来一段时间污染物浓度进行预测，在城市大气污染治理、环境治理方面具有重要作用。信息见表8-8。

表 8-8 区域大气污染物预测模型信息

模型	预报污染物变量	时间分辨率	水平分辨率	区域范围
区域大气污染物要素模型	PM10、PM2.5、NO_2、O_3、SO_2、CO	1h	0.0625°*0.0625°	特定区域

支持训练的模型清单见表8-9，您可根据具体使用场景选择合适的模型。

表 8-9 区域大气污染物预测模型的类型

模型名称	模型版本及时间分辨率	说明
Pangu-AI4S-Regional-Pollution	v1.1.2-1h	2025年3月发布的版本，用于区域污染物基础要素预测，1个训练单元起训及1个实例部署。

- **区域新能源风预测模型选择建议**：

  科学计算大模型的区域新能源风预测模型，可以对未来一段时间的天气要素以及风速进行预测。

  区域新能源风模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域新能源风模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

版权所有 © 华为云计算技术有限公司

- 预训练：可以在重新指定层次、高空变量、表面变量、以及表面观测、时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
- 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

区域新能源风，信息见**表 区域新能源风预测模型信息表**。

表 8-10 区域新能源风预测模型信息表

模型	层次	高空变量	表面变量	时间分辨率	水平分辨率	区域范围
区域新能源风模型	1, 0.994065, 0.986518, 0.976983, 0.96503, 0.931987, 0.853381, 0.562185, 0.338478	P：压强 U：U风 V：V风 Q：比湿 T：温度	MLSP：海平面气压 U10：10米U风，经度方向 V10：10米V风，纬度方向 T2M：2米温度	1小时	0.03°*0.03°	特定区域

支持训练的模型清单见**表 区域新能源风预测模型**，

表 8-11 区域新能源风预测模型

模型名称	模型版本及时间分辨率	说明
Pangu-AI4S-Regional-Wind	v1.1.3-1h	2025年5月发布的版本，用于新能源风要素预测，如百米风，时间分辨率为1小时，支持预训练、微调、部署、在线推理特性，基于Snt9B3推理。

- **区域新能源光预测模型选择建议**：

  科学计算大模型的区域新能源光预测模型，可以对未来一段时间的光辐照进行预测。

  区域新能源光模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域新能源光模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

 版权所有 © 华为云计算技术有限公司

- 预训练：可以在重新指定表面变量、以及时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
- 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

区域新能源风，信息见**表 区域新能源光预测模型信息表**。

表 8-12 区域新能源光预测模型信息表

模型	表面变量	时间分辨率	水平分辨率	区域范围
区域新能源光模型	短波辐射	1小时	0.03°*0.03°	特定区域

支持训练的模型清单见**表 区域新能源光预测模型**，

表 8-13 区域新能源光预测模型

模型名称	模型版本及时间分辨率	说明
Pangu-AI4S-Regional-Photovoltaics	v1.1.0-1h	2025年5月发布的版本，用于区域新能源光要素预测，如短波辐射、时间分辨率为1小时，支持预训练、微调、部署、在线推理特性，基于Snt9B3推理。

科学计算大模型训练类型选择建议

目前，全球中期天气要素模型提供训练功能和推理功能，降水模型仅提供推理功能。

- **全球中期天气要素预测模型的训练类型选择建议：**

  全球中期天气要素预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的中期天气要素预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

  - 预训练：训练用于添加新的高空层次、高空变量或表面变量。如果您需要在现有模型中引入新要素，需要使用训练（重新训练模型）。在重训配置参数时，您可以选择新要素进行训练。请注意，所选的数据集必须包含您想要添加的新要素。此外，您还可以通过训练更改所有的模型参数，以优化模型性能。
  - 微调：微调是将新数据应用于已有模型的过程。它适用于不改变模型结构参数和引入新要素的情况。如果您有新的观测数据，可以使用微调来更新模型的权重，以适应新数据。

- **中期海洋智能预测模型的训练类型选择建议**：

  中期海洋智能预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域中期海洋智能预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

 版权所有 © 华为云计算技术有限公司

  - 预训练：可以在重新指定深海变量、海表变量、以及深海层深、时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
  - 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

- **区域大气污染物预测模型的训练类型选择建议**：

  区域大气污染物预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域大气污染物预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

  - 预训练：可以在重新指定大气污染物要素、以及时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
  - 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

- **区域新能源风预测模型的训练类型选择建议**：

  区域新能源风预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域新能源风预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

  - 预训练：可以在重新指定层次、高空变量、表面变量、以及表面观测、时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
  - 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

- **区域新能源光预测模型的训练类型选择建议**：

  区域新能源光预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域新能源光预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

  - 预训练：可以在重新指定表面变量、以及时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
  - 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

8.3.2 创建科学计算大模型训练任务

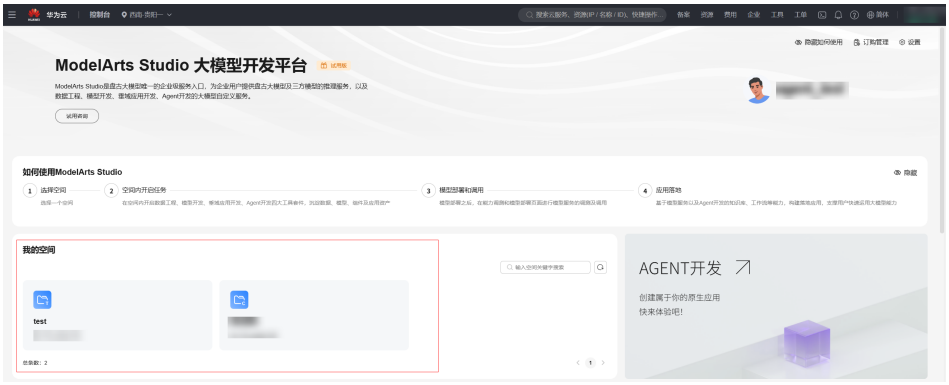
创建科学计算大模型中期天气要素预测微调任务

创建科学计算大模型中期天气要素预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 8-3 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-14完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-14 科学计算大模型中期天气要素预测微调训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	层次	设置训练数据的层次信息。在“预训练”场景中，可以添加或去除高空层次，训练任务将根据配置的层次信息重新训练模型。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	高空变量	设置训练数据的高空变量信息。在“预训练”场景中，可以添加或去除新的高空变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的高空变量重新训练模型。
	表面变量	设置训练数据的表面变量信息。在“预训练”场景中，可以添加或去除新的表面变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的表面变量重新训练模型。
	表面静态量	表面静态量默认包括地形高度、LAND_MASK和SOIL_TYPE，用于初始化模型状态并提供地表特性信息。当前不支持添加或去除这些静态量。 • LAND_MASK：一个二维数组，表示模型网格中每个单元格是否是陆地。 • SOIL_TYPE：表示地表土壤分类，影响土壤的物理和化学特性，如水分保持能力、热容量和导热性。
模型输出控制参数	表面变量相对高空变量的权重	指在模型训练过程中对表面变量相对于深海层变量赋予的权重，总Loss=高空Loss+surface_loss_weight*表面Loss。取值范围：(0.05, 10)。
	其他通用参数	请参考**表8-25**进行配置。
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	描述	训练任务描述。

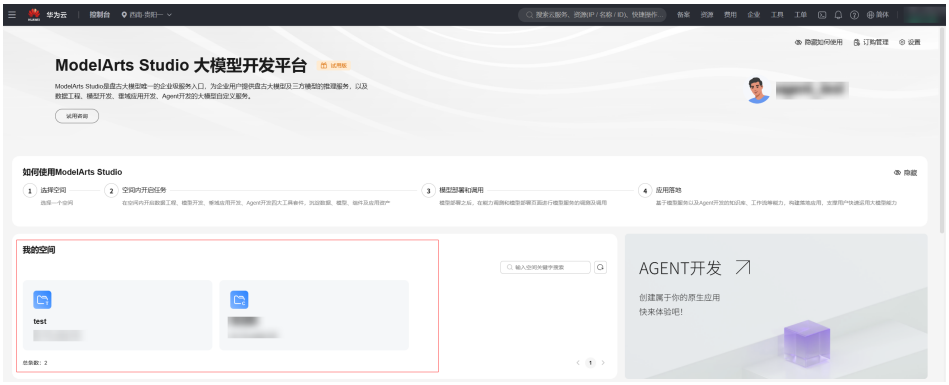
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型中期天气要素预测预训练任务

创建科学计算大模型中期天气要素预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-15完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-15 科学计算大模型中期天气要素预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	层次	设置训练数据的层次信息。在“预训练”场景中，可以添加或去除高空层次，训练任务将根据配置的层次信息重新训练模型。
	高空变量	设置训练数据的高空变量信息。在“预训练”场景中，可以添加或去除新的高空变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的高空变量重新训练模型。
	表面变量	设置训练数据的表面变量信息。在“预训练”场景中，可以添加或去除新的表面变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的表面变量重新训练模型。
	表面静态量	表面静态量默认包括地形高度、LAND_MASK和SOIL_TYPE，用于初始化模型状态并提供地表特性信息。当前不支持添加或去除这些静态量。 • LAND_MASK：一个二维数组，表示模型网格中每个单元格是否是陆地。 • SOIL_TYPE：表示地表土壤分类，影响土壤的物理和化学特性，如水分保持能力、热容量和导热性。
模型输出控制参数	表面变量相对高空变量的权重	指在模型训练过程中对表面变量相对于深海层变量赋予的权重，总Loss=高空Loss+surface_loss_weight*表面Loss。取值范围：(0.05, 10)。
	其他通用参数	请参考**表8-25**进行配置。
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域中期海洋智能预测微调任务

创建科学计算大模型区域中期海洋智能预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-5 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-16完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

版权所有 © 华为云计算技术有限公司

表 8-16 科学计算大模型区域中期海洋智能预测微调参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	深海层深	海深层深是指海洋模型将整个水柱（从海面到海底）按一定深度间隔划分成多个层次，每个深度值代表模型在这个深度层进行计算和模拟。例如，"0m"代表海平面，"6m"代表在海平面以下6米处的一层，以此类推。范围包括：0m、6m、10m、20m、30m、50m、70m、100m、125m、150m、200m、250m、300m、400m、500m。
	深海变量	深海变量是用于模拟和描述海洋状态的关键物理量。 T：15层：海温(℃) S：15层：海盐(PSU) U：15层：海流经向速率 (ms-1) V：15层：海流纬向速率 (ms-1)
	海表变量	海表变量用于描述海洋表层和其上方大气的状态的关键物理量。它们主要用于模拟和分析海洋表面的风速、温度、和气压等特征。 U10：1层：海表面10m经向风速(ms-1) V10：1层：海表面10m纬向风速(ms-1) T2m：1层：海表面2m温度 (℃) MSL：1层：平均海平面气压 (Pa) SP：1层：海表面气压 (Pa)

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
模型输出控制参数	海表变量相对深海变量的权重	指在模型训练过程中对海表变量相对于深海层变量赋予的权重，总Loss=深海层Loss+surface_loss_weight*海表Loss。取值范围：(0.05, 10)。
	其他通用参数	请参考**表8-25**进行配置。
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

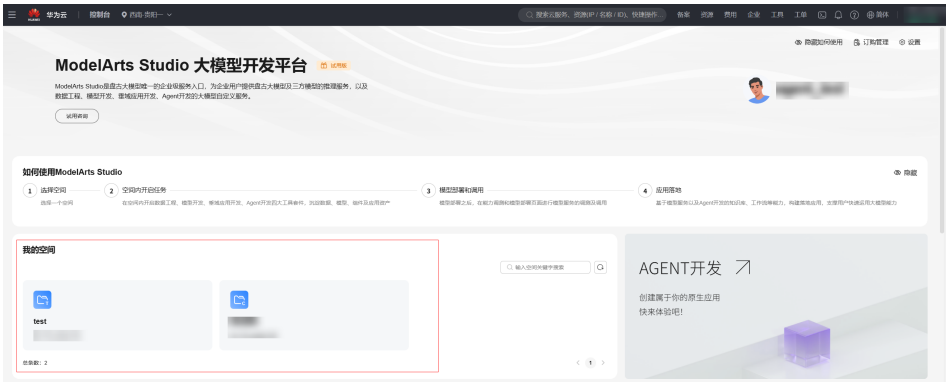
创建科学计算大模型区域中期海洋智能预测预训练任务

创建科学计算大模型区域中期海洋智能预测预训练任务步骤如下：

 版权所有 © 华为云计算技术有限公司

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-6 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-17完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-17 科学计算大模型区域中期海洋智能预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
模型数据配置	深海层深	海深层深是指海洋模型将整个水柱（从海面到海底）按一定深度间隔划分成多个层次，每个深度值代表模型在这个深度层进行计算和模拟。例如，"0m"代表海平面，"6m"代表在海平面以下6米处的一层，以此类推。范围包括：0m、6m、10m、20m、30m、50m、70m、100m、125m、150m、200m、250m、300m、400m、500m。
	深海变量	深海变量是用于模拟和描述海洋状态的关键物理量。 T：15层：海温(℃) S：15层：海盐(PSU) U：15层：海流经向速率 (ms-1) V：15层：海流纬向速率 (ms-1)
	海表变量	海表变量用于描述海洋表层和其上方大气的状态的关键物理量。它们主要用于模拟和分析海洋表面的风速、温度、和气压等特征。 U10：1层：海表面10m经向风速(ms-1) V10：1层：海表面10m纬向风速(ms-1) T2m：1层：海表面2m温度 (℃) MSL：1层：平均海平面气压 (Pa) SP：1层：海表面气压 (Pa)
	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
模型输出控制参数	海表变量相对深海变量的权重	指在模型训练过程中对海表变量相对于深海层变量赋予的权重，总Loss=深海层Loss+surface_loss_weight*海表Loss。取值范围：(0.05, 10)。
	其他通用参数	请参考**表8-25**进行配置。
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域大气污染物预测微调任务

创建科学计算大模型区域大气污染物预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-7 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-18完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

 版权所有 © 华为云计算技术有限公司

表 8-18 科学计算大模型区域大气污染物预测微调参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
	污染物变量	污染物变量是用于模拟和描述大气污染物浓度的关键物理量。 PM10（$\mu g/m^3$） PM2.5（$\mu g/m^3$） NO_2：二氧化氮（$\mu g/m^3$） O_3：臭氧（$\mu g/m^3$） SO_2：二氧化硫（$\mu g/m^3$） CO：一氧化碳（mg/m^3）
模型输出控制参数	请参考**表8-25**进行配置。	

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
正则化参数	请参考表8-25进行配置。	
优化器参数	请参考表8-25进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

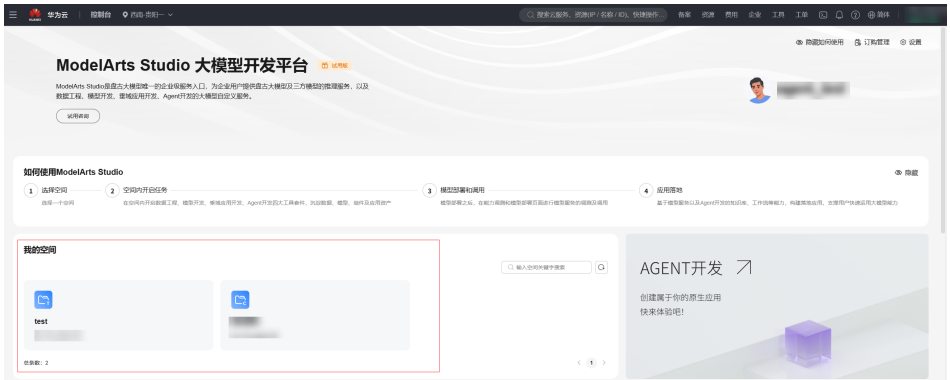
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域大气污染物预测预训练任务

创建科学计算大模型区域大气污染物预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-8 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表8-19**完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-19 科学计算大模型区域大气污染物预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
模型数据配置	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
	污染物变量	污染物变量是用于模拟和描述大气污染物浓度的关键物理量。 PM10（$\mu g/m^3$） PM2.5（$\mu g/m^3$） NO_2：二氧化氮（$\mu g/m^3$） O_3：臭氧（$\mu g/m^3$） SO_2：二氧化硫（$\mu g/m^3$） CO：一氧化碳（mg/m^3）

 版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
模型输出控制参数	请参考表8-25进行配置。	
正则化参数	请参考表8-25进行配置。	
优化器参数	请参考表8-25进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域新能源风预测微调任务

创建科学计算大模型区域新能源风预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 8-9 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表8-20**完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-20 科学计算大模型区域新能源风预测微调参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	层次	设置训练数据的模式层信息，在“预训练”的场景中也支持您添加或去除新的模式层，训练任务会根据您配置的模式层对模型重新进行训练。
	高空变量	设置训练数据的高空变量信息，同时在“预训练”的场景中也支持您添加或去除新的高空变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的高空变量对模型重新进行训练。
	表面变量	设置训练数据的表面变量信息，同时在“预训练”的场景中也支持您添加或去除新的表面变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面变量对模型重新进行训练。
	表面观测	设置训练数据的表面观测信息，同时在“预训练”的场景中也支持您添加或去除新的表面观测变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面观测变量对模型重新进行训练。
	表面静态量	在预训练场景下可以根据需要勾选表面静态量用于模型训练，详见**表面静态量参数含义**。
模型输出控制参数	请参考**表8-25**进行配置。	
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

版权所有 © 华为云计算技术有限公司

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域新能源风预测预训练任务

创建科学计算大模型区域新能源风预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-10 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-21完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-21 科学计算大模型区域新能源风预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
模型数据配置	层次	设置训练数据的模式层信息，在“预训练”的场景中也支持您添加或去除新的模式层，训练任务会根据您配置的模式层对模型重新进行训练。
	高空变量	设置训练数据的高空变量信息，同时在“预训练”的场景中也支持您添加或去除新的高空变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的高空变量对模型重新进行训练。
	表面变量	设置训练数据的表面变量信息，同时在“预训练”的场景中也支持您添加或去除新的表面变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面变量对模型重新进行训练。
	表面观测	设置训练数据的表面观测信息，同时在“预训练”的场景中也支持您添加或去除新的表面观测变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面观测变量对模型重新进行训练。
	表面静态量	在预训练场景下可以根据需要勾选表面静态量用于模型训练，详见**表面静态量参数含义**。
模型输出控制参数	请参考**表8-25**进行配置。	
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

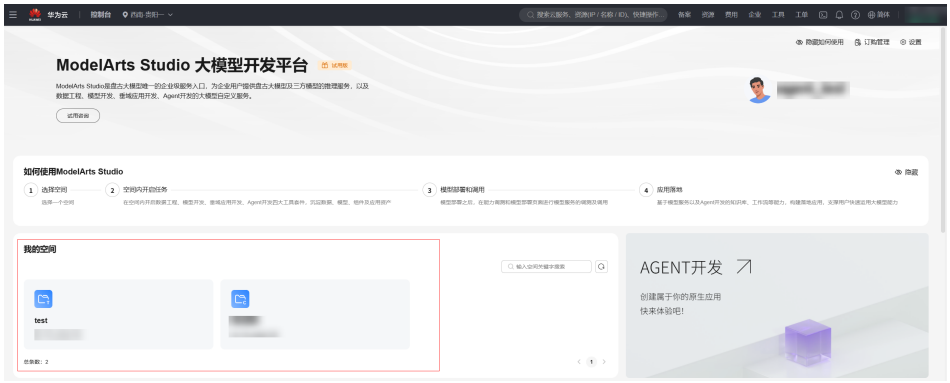
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域新能源光预测微调任务

创建科学计算大模型区域新能源光预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-11 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-22完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-22 科学计算大模型区域新能源光预测微调参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	训练目标	选择“全量微调”。 • 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	表面变量	设置训练数据的表面变量信息，同时在“预训练”的场景中也支持您添加或去除新的表面变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面变量对模型重新进行训练。
	表面静态量	在预训练场景下可以根据需要勾选表面静态量用于模型训练，详见**表面静态量参数含义**。
模型输出控制参数	请参考**表8-25**进行配置。	
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

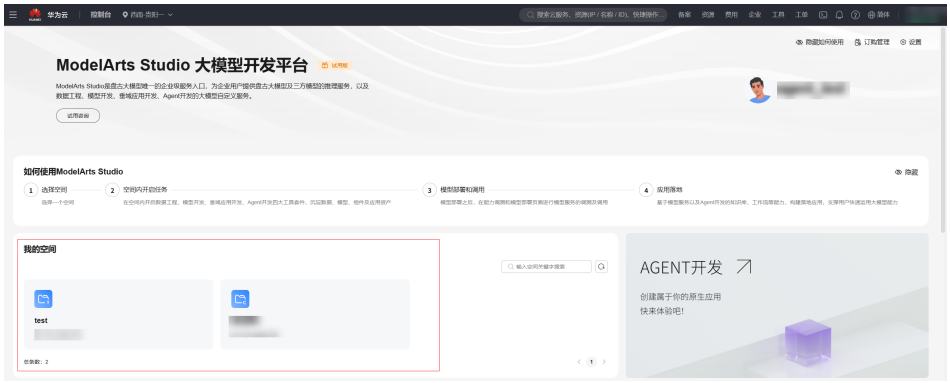
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域新能源光预测预训练任务

创建科学计算大模型区域新能源风预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-12 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表8-23完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 8-23 科学计算大模型区域新能源光预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“科学计算大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	模型时间分辨率	模型预报的时间间隔，类型为整数，可取值 1,3,6,24，单位为小时。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
模型数据配置	表面变量	设置训练数据的表面变量信息，同时在“预训练”的场景中也支持您添加或去除新的表面变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面变量对模型重新进行训练。
	表面静态量	在预训练场景下可以根据需要勾选表面静态量用于模型训练，详见**表面静态量参数含义**。
模型输出控制参数	请参考**表8-25**进行配置。	
正则化参数	请参考**表8-25**进行配置。	
优化器参数	请参考**表8-25**进行配置。	
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

版权所有 © 华为云计算技术有限公司

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

表 8-24 表面静态量参数含义

参数名	区域范围	水平分辨率	参数含义
TERRAIN_HEIGHT_025	全球	0.25	地形高度。
SOIL_TYPE_025	全球	0.25	地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
LAND_MASK_025	全球	0.25	二维数组，用于表示模型网格中每个单元格是否是陆地。
OCEAN_MASK_025	全球	0.25	二维数组，用于表示模型网格中每个单元格是否是海洋。
SOIL_TYPE_1_120	全球	1/120	地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
LAND_MASK_1_120	全球	1/120	二维数组，用于表示模型网格中每个单元格是否是陆地。
TERRAIN_HEIGHT_1_120	中国区域	1/120	地形高度。
DAY_PROGRESS	全球	0.25	日变化。
YEAR_PROGRESS	全球	0.25	季节变化。
TISR	全球	0.25	天顶辐射功率。
TISR_1H	全球	0.25	天顶辐射1小时累计能量。
TISR_3H	全球	0.25	天顶辐射3小时累计能量。
TISR_6H	全球	0.25	天顶辐射6小时累计能量。
TISR_24H	全球	0.25	天顶辐射24小时累计能量。

 版权所有 © 华为云计算技术有限公司

通用参数

表 8-25 科学计算大模型训练通用参数

参数分类	参数名称	参数说明
模型输出控制参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。取值范围：[1-1000]。
	损失类型	用来衡量模型预测结果与真实结果之间的差距的函数，提供MAE（平均绝对误差）、MSE（均方误差）两种损失函数。 ● MSE对于异常值非常敏感，因为它会放大较大的误差。因此，如果您数据中没有异常值，或者希望模型对大的误差给予更大的惩罚，可选择MSE。 ● 如果数据中存在异常值，或者希望模型对所有的误差都一视同仁，可选择MAE。
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率，定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输入数据加噪音的尺度	给输入数据加噪音的尺度，定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	给输出数据加噪音的概率	给输出数据加噪音的概率，定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	给输出数据加噪音的尺度，定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器参数	优化器种类	优化器种类。优化器是用于更新模型参数的算法，目前支持ADAM优化器。

 版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	第一个动量矩阵的指数衰减率(beta1)	数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	第二个动量矩阵的指数衰减率(beta_2)	用于定义ADAM优化器中的二阶矩估计的指数衰减率。二阶矩估计相当于RMSProp，可以调整学习率。取值范围：(0,1)。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂，取值需≥0。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。 预训练时，默认值为：0.00001，范围为[0, 0.001]。
	学习率调整策略	用于选择学习率调度器的类型。学习率调度器可以在训练过程中动态地调整学习率，以改善模型的训练效果。目前支持CosineDecayLR调度器。

8.3.3 查看科学计算大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表8-26。

表 8-26 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。

版权所有 © 华为云计算技术有限公司

训练状态	训练状态含义
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于已完成训练，训练状态是“训练完成”状态的任务，单击任务名称，可在“训练结果”页面查看训练指标，不同模型的训练指标介绍请参见表8-27。

图 8-13 查看训练指标-Loss

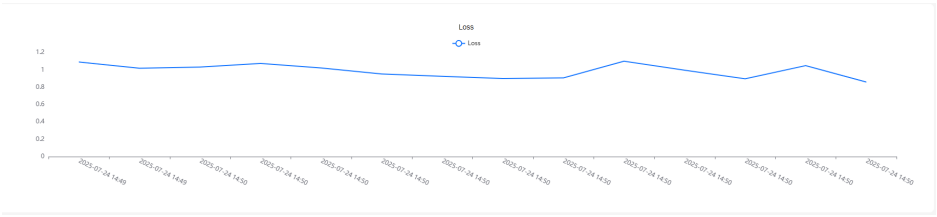


图 8-14 查看训练指标-验证 Loss

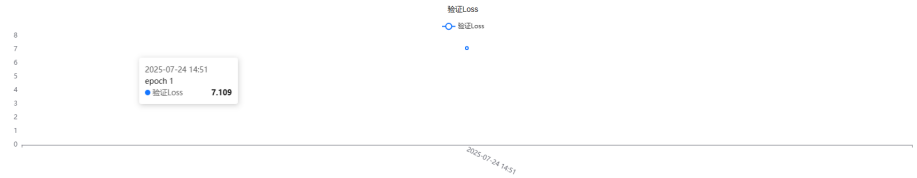


图 8-15 查看训练指标-高空 Loss

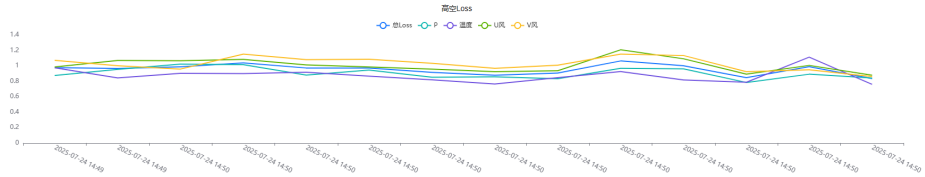
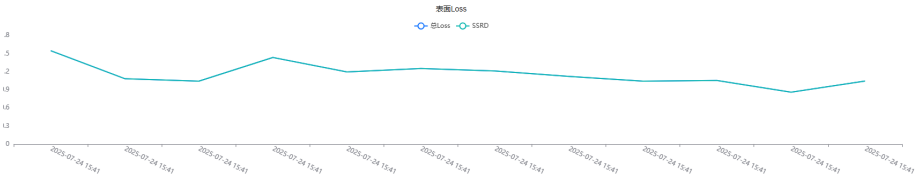


图 8-16 查看训练指标-表面 Loss



 版权所有 © 华为云计算技术有限公司

图 8-17 查看训练指标-评价指标

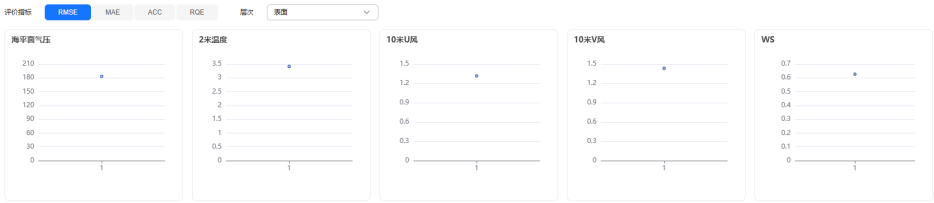


表 8-27 训练指标说明

模型	训练指标	指标说明
科学计算大模型	Loss	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。这里代表高空Loss（深海Loss）和表面Loss（海表Loss）的综合Loss。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	高空Loss（深海Loss）	高空Loss（深海Loss）是衡量模型在高空层次变量或在深海变量预测结果与真实结果之间差距的指标。该值越小，表示模型在高空（深海）变量的预测精度越高。
	表面Loss（海表Loss）	表面Loss（海表Loss）是衡量模型在表面层次变量或在海表变量预测结果与真实结果之间差距的指标。该值越小，表示模型在表面（海表）变量的预测精度越高。
	RMSE	均方根误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的平方误差的平均值的平方根。该值越小，代表模型性能越好。
	MAE	平均绝对误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的绝对误差的平均值。该值越小，代表模型性能越好。

版权所有 © 华为云计算技术有限公司

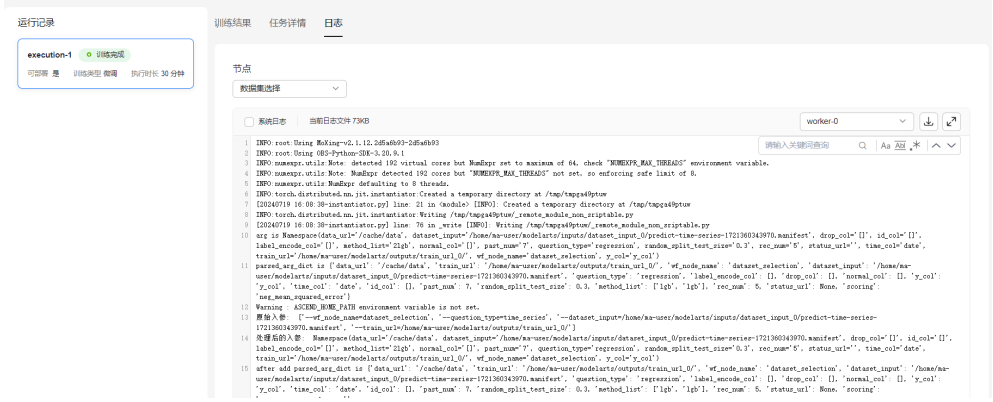
模型	训练指标	指标说明
	ACC	ACC（异常相关系数，距平相关系数，Anomaly Correlation Coefficient）是一个重要的统计指标，用于衡量预报系统的质量。它通过计算预报值与观测值之间的相关性来评估预报的准确性。ACC的计算涉及到预报值、观测值和气候平均值的差异，其值范围从-1到+1，值越接近+1表示预报与观测的一致性越好，值为0表示没有相关性，而负值则表示反向相关。
	RQE	衡量预测值与真实值之间差距的指标。它是所有单个观测的相对误差的平方和。该值越小，代表模型性能越好。
	验证loss	验证损失是用于衡量模型在验证集上预测结果和真实结果之间的差距的指标，通常情况越小越好，是一个综合所有损失的指标，可以用于调整超参数，并用于判断是否过拟合， 一般来说，一个正常的valid_loss应该是单调递减的，即随着训练的进行，valid_loss值不断减小，直到收敛到一个较小的值，需要和训练集loss协同观察。

获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。对于训练异常或失败的任务也可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**科学计算大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 8-18 获取训练日志



版权所有 © 华为云计算技术有限公司

训练异常定位方法

模型训练失败时，可以按照如下定位思路对训练任务做初步定位。

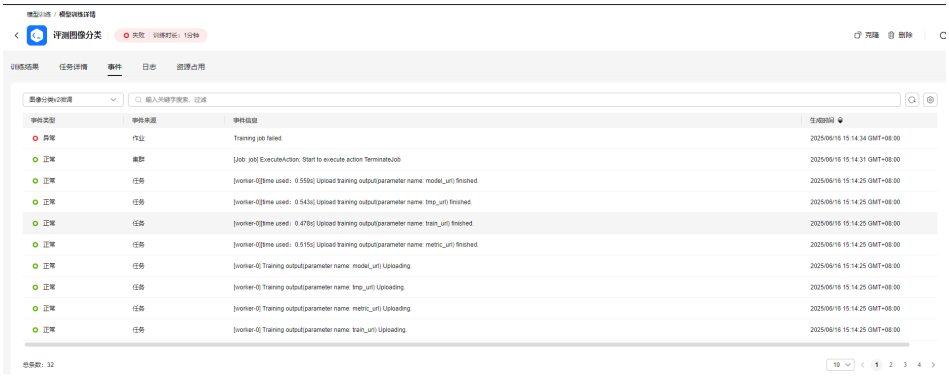
1. 在训练任务列表查看任务失败原因。

   模型启动训练后，可以在模型训练列表中查看训练任务的状态，当任务状态为“失败”时，可将鼠标放置在“失败”上，可以查看到任务失败的概要信息。



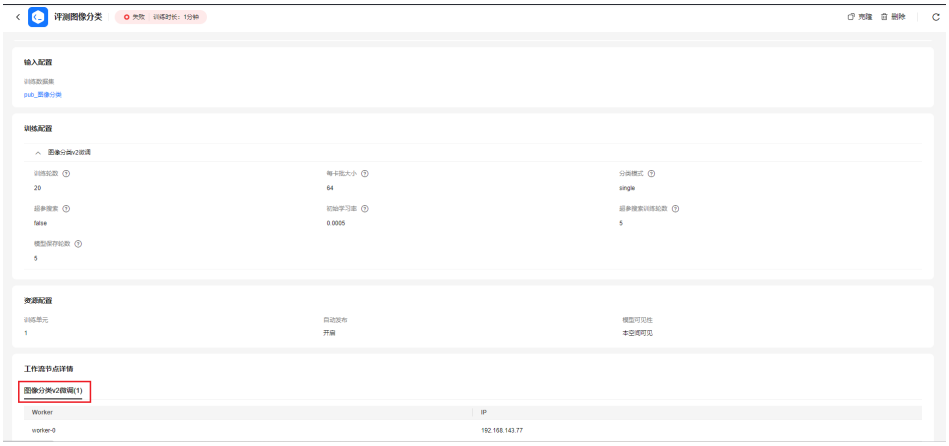
2. 单击任务名，进入任务详情页查看训练任务运行关键事件信息。

   单击“事件”页签，可查看训练任务子任务运行事件，通过事件可以观察到当前任务的运行情况。事件可以按照“正常、异常、告警”三类事件类型过滤。



3. 获取当前任务运行的worker、workerIP以及任务异常退出时的节点和退出码信息。

   在ModelArts Studio训练任务页面单击需要查看的任务名，进入“任务详情”页，在“工作流节点详情”下可以查看到当前任务所运行的节点及IP信息。



4. 当训练任务运行失败时，可查看“模型训练详情 -> 基本信息 -> 状态”，单击“查看失败节点清单”获取训练任务故障节点及退出码信息。

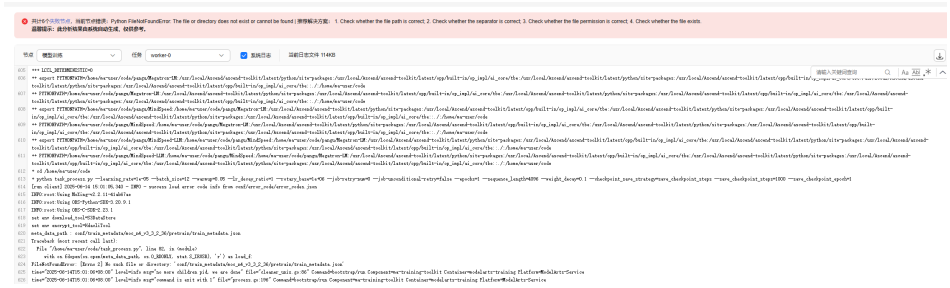
版权所有 © 华为云计算技术有限公司



5. 训练作业遇到问题时，可查看“模型训练详情 -> 日志”，多数场景下的问题可以通过日志报错信息直接定位。

   如果训练作业运行失败，ModelArts Studio会自动识别导致作业失败的原因，在训练日志界面上给出提示并对部分常见训练错误给出分析建议。包含失败的可能原因、推荐的解决方案。

图 8-19 训练错误分析建议



说明

针对分布式作业，只会显示当前节点的一个分析结果，作业的失败需要综合各个节点的失败原因做一个综合判断。

8.3.4 发布训练后的科学计算大模型

科学计算大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-20 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。

版权所有 © 华为云计算技术有限公司

3. 单击进入“训练结果”页签，单击“发布”。

图 8-21 训练结果页面



4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

8.3.5 管理科学计算大模型训练任务

在训练任务列表中，任务创建者可以对创建好的任务进行编辑、启动、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，进入模型训练页面，可进行如下操作：
   - 编辑。单击操作列的“编辑”，可以修改模型的训练参数、训练数据以及基本信息等。
   - 克隆。单击操作列的“克隆”，参照**创建科学计算大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“删除”，可以删除当前不需要的训练任务。

8.3.6 科学计算大模型训练常见报错与解决方案

科学计算大模型训练常见报错及解决方案请详见**表8-28**。

版权所有 © 华为云计算技术有限公司

表 8-28 科学计算大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
训练日志提示“TrainErrorCodes.NotEnoughTrainData”或“TrainErrorCodes.NotEnoughValidData”报错	日志提示“TrainErrorCodes.NotEnoughTrainData”，表示训练集没有有效样本，“TrainErrorCodes.NotEnoughValidData”，表示验证集没有有效样本。	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据文件特征与模型输入要求不符。 • 数据的时间跨度未满足模型训练要求。	请检查数据是否符合模型要求。

8.4 部署科学计算大模型

8.4.1 创建科学计算大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 科学计算大模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考表8-29完成部署参数设置。

表 8-29 科学计算大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： • 来源：选择“模型广场”。 • 类型：选择“科学计算大模型”，并选择需要进行部署的模型和版本。推荐的推理场景与模型组合请见表8-30。

版权所有 © 华为云计算技术有限公司

参数分类	部署参数	参数说明
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 本地挂载路径（选填）：在容器内部将卷挂载的本地路径。挂载后，容器中的应用程序可以通过这个路径访问宿主机上的数据。 – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。 – 实例数：设置部署模型时所需的实例数。
	作业输入方式	选择 “OBS”表示从OBS中读取数据。
	作业输出方式	选择 “OBS”表示将输出结果存储在OBS中。
	作业配置参数	设置模型部署参数信息，平台已给出默认值。
资源配置	计费模式	包年包月计费模式。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

表 8-30 科学计算气象模型部署场景

场景	模型组合
天气推理	Pangu-AI4S-Global-Weather (1h + 3h + 6h + 24h)

版权所有 © 华为云计算技术有限公司

场景	模型组合
降水推理	Pangu-AI4S-Global-Weather (6h + 24h) Pangu-AI4S-Global-Precip 6h
海洋推理	Pangu-AI4S-Global-Ocean + Pangu-AI4S-Regional-Ocean + Pangu-AI4S-Global-Ecology + Pangu-AI4S-Global-Swell
污染物推理	Pangu-AI4S-Global-Weather (1h + 3h + 6h + 24h) + 自行训练出的污染物区域模型（推荐时间分辨率 1h + 3h + 6h + 24h）
新能源风推理	Pangu-AI4S-Global-Weather (1h + 3h + 6h + 24h) + 自行训练出的区域新能源风模型
新能源光推理	Pangu-AI4S-Global-Weather (1h + 3h + 6h + 24h) + 自行训练出的区域新能源风模型

5. 参数填写完成后，单击“立即部署”。

8.4.2 查看科学计算大模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-22 我的空间



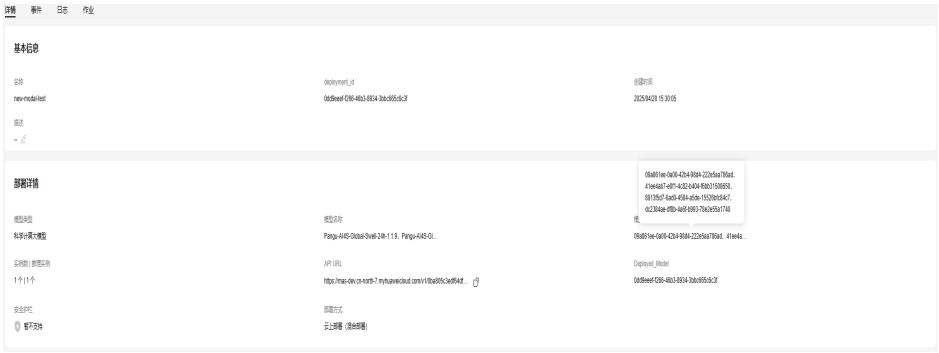
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建科学计算大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

版权所有 © 华为云计算技术有限公司

图 8-23 部署详情



8.4.3 管理科学计算大模型部署任务

模型更新

完成**创建科学计算大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-24 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

版权所有 © 华为云计算技术有限公司

图 8-25 升级模式



说明

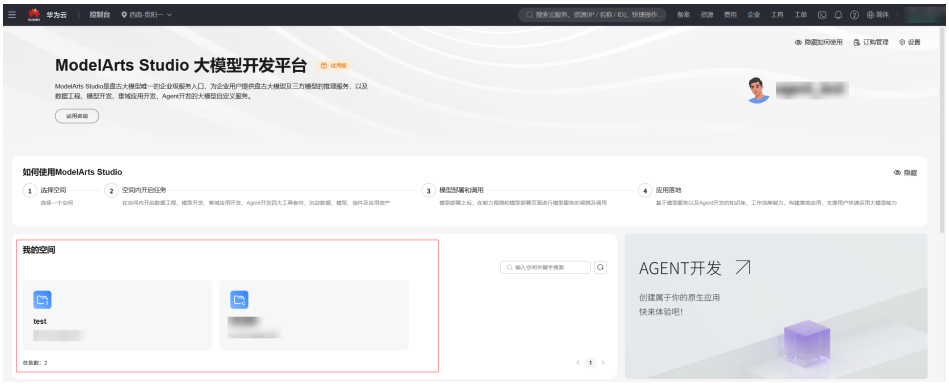
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建科学计算大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-26 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

版权所有 © 华为云计算技术有限公司

图 8-27 升级模式



说明

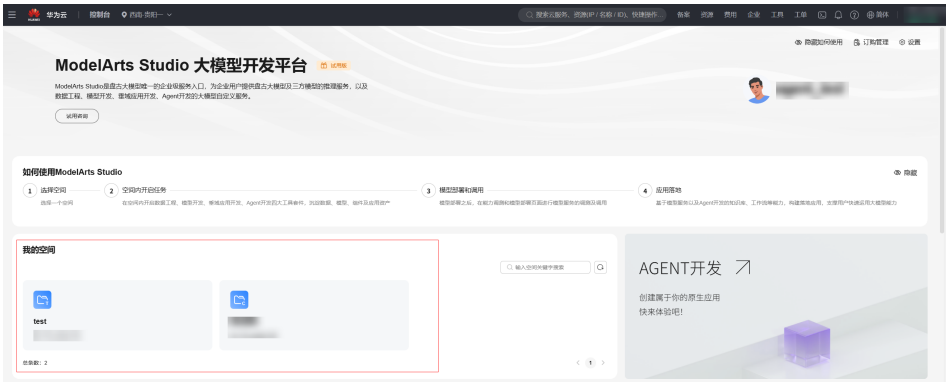
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成**创建科学计算大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-28 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 8-29 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

8.5 调用科学计算大模型

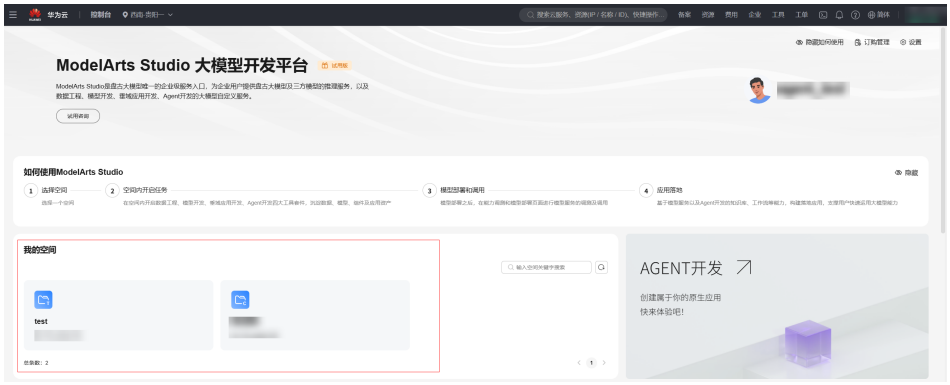
8.5.1 使用“体验中心”调用科学计算大模型

体验中心功能支持用户调用预置或训练后的科学计算大模型。使用该功能前，请完成模型的部署操作，步骤详见**创建科学计算大模型部署任务**。

使用“体验中心”调用科学计算大模型可实现包括全球中期天气要素预测、全球中期降水预测、全球海洋要素、区域海洋要素、全球海洋生态、全球海浪高度场景的预测能力。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-30 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“体验中心”，单击“科学计算”页签。
3. 选择部署的服务。
4. 完成页面参数配置。请参考**表1 科学计算大模型体验中心参数说明**。

表 8-31 科学计算大模型体验中心参数说明

参数	说明
结果存储路径	用于存放模型推理结果的OBS路径。
输入数据	支持根据不同的模型选择相应的OBS文件路径。例如：天气模型选择ERA5数据，海洋模型选择深海数据文件夹，污染物模型选择污染物观测数据文件夹等。
预报天数	支持选择以起报时间点为开始，进行预报的天数，范围为1~30天。
起报时间	支持选择多个起报时间作为推理作业的开始时间，每个起报时间需为输入数据中存在的时间点。
图片输出	用于选择是否输出图片结果。
集合预报	用于选择是否开启集合预报。 在气象预报中，集合预报是指对初始场加入一定程序的扰动，使其生成一组由不同初始场预报的天气预报结果，从而提供对未来天气状态的概率信息。这种方法可以更好地表达预报的不确定性，从而提高预报的准确性和可靠性。
集合成员数	用于选择生成预报的不同初始场的数量，取值为2~10。
扰动类型	用于选择生成集合预报初始场的扰动类型，仅支持perlin加噪方式。 • Peilin噪音通过对输入数据（比如空间坐标）进行随机扰动，让模拟出的天气接近真实世界中的变化。
集合预报的Perlin加噪scale	用于选择集合预报的Perlin加噪强度。
集合预报的Perlin加噪x	用于选择集合预报的Perlin加噪x经度方向的尺度。
集合预报的Perlin加噪octave	用于选择集合预报的Perlin加噪octave。Perlin噪音的octave指的是噪音的频率，在生成Perlin噪音时，可以将多个不同频率的噪音叠加在一起，以增加噪音的复杂度和细节。每个频率的噪音称为一个octave，而叠加的octave数越多，噪音的复杂度也就越高。
集合预报的Perlin加噪y	用于选择集合预报的Perlin加噪y纬度方向的尺度。

版权所有 © 华为云计算技术有限公司

图 8-31 科学计算参数设置 1

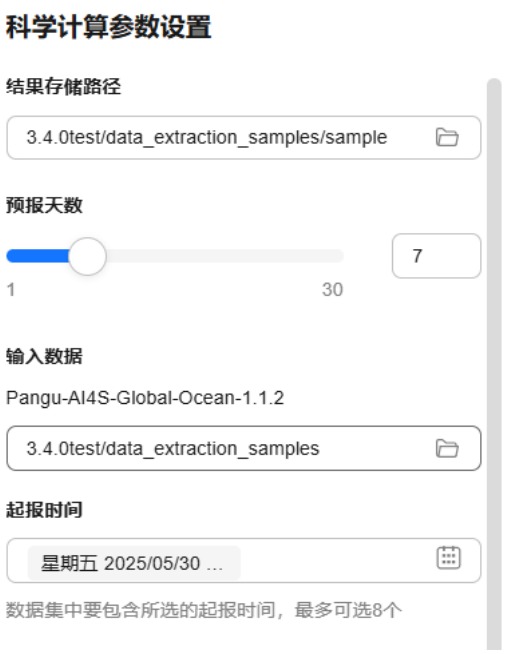


图 8-32 科学计算参数设置 2



 版权所有 © 华为云计算技术有限公司

8.5.2 使用 API 调用科学计算大模型

预置模型或训练后的模型部署成功后，可以使用API调用科学计算大模型。

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 8-33 我的空间



2. 获取调用路径。

   在左侧导航栏中选择“模型开发 > 模型部署”。

   - 获取已部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“详情”页签，可获取模型调用路径，如图8-34。

   图 8-34 获取已部署模型的调用路径

   - 获取预置服务的调用路径。在“预置服务”页签中，选择所需调用的科学计算大模型，单击“调用路径”，在“调用路径”弹窗可获取模型调用路径，如图8-35。

   图 8-35 获取预置服务的调用路径

版权所有 © 华为云计算技术有限公司

使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见**获取调用路径**。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，API Key认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。

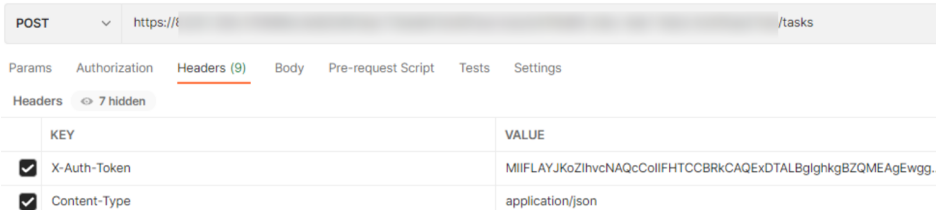
   参考**表8-32**填写请求Header参数。

   表 8-32 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
API Key认证	Content-Type	application/json
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。

   如**图8-36**，为Token认证方式的请求Header参数填写示例。

   图 8-36 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。API参数说明详见《API参考》文档。

```
{
   "name": "test-task624",
   "input": {
      "type": "obs",
      "data": [
         {
            "bucket": "pangu-weather-data",
```


版权所有 © 华为云计算技术有限公司

```
            "path": "test/"
        }
    ]
  },
  "output": {
    "obs": {
      "bucket": "pangu-weather-test",
      "path": "output/"
    }
  },
    "config": {
          "data_dirs":"{\"global-weather\": \"obs://pangu-weather-data/test\"}",
          "start_time_begin": "2022010100",
    "start_time_end": "2022010106",
    "start_time_interval_hours": 6,
    "forecast_lead_hours": 168
  }
}
```

4. 单击Postman界面“Send”，发送请求。科学计算大模型API调用成功后，会返回任务id参数**task_id**，可获取任务ID参数值。
5. 在Postman中新建一个GET请求，填入域名（将**获取调用路径**中获取的URL去除末尾的“/tasks”即为该域名），设置请求Header参数和任务ID参数。单击Postman界面的“Send”发送请求，以获取科学计算大模型的调用结果。

```
查询科学计算大模型调用详情API
GET /tasks/{task_id}
```

 版权所有 © 华为云计算技术有限公司

9 开发盘古搜索规划模型

搜索规划模型介绍
部署搜索规划大模型
调用搜索规划大模型

9.1 搜索规划模型介绍

Pangu-SearchPlan模型，用于RAG场景，提供通用意图分类/多轮查询改写/复杂查询分解/时间抽取等功能，在RAG任务中生产用于检索的query，以及根据query分类路由到后续不同的流程。

表 9-1 推理特性

推理精度	**FP16**
起推规格（推理单元）	1
QPS	10

9.2 部署搜索规划大模型

9.2.1 创建 Pangu-SearchPlan-ZH_EN 模型部署任务

平台支持部署预置的Pangu-SearchPlan-ZH_EN大模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 搜索规划模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表1 Pangu-SearchPlan-ZH_EN模型部署参数说明**完成部署参数设置。

 版权所有 © 华为云计算技术有限公司

表 9-2 Pangu-SearchPlan-ZH_EN 模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“搜索规划模型”，并选择需要进行部署的模型和版本。
	部署方式	云上部署：算法部署至平台提供的资源池中。
资源配置	计费模式	包年包月计费模式。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

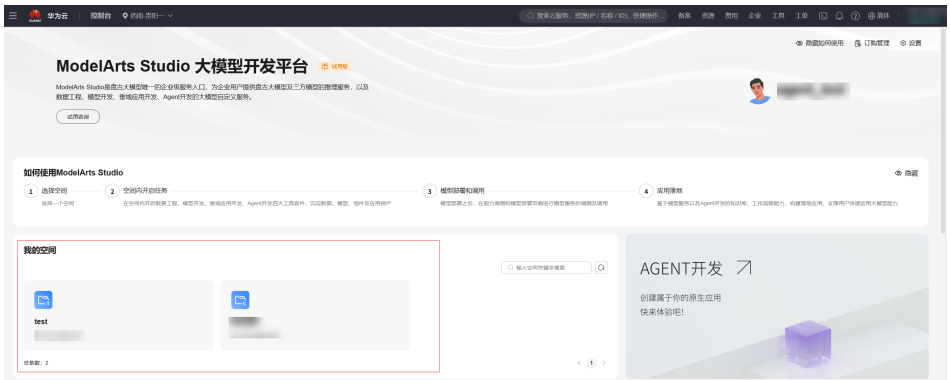
5. 参数填写完成后，单击“立即部署”。

9.2.2 查看 Pangu-SearchPlan-ZH_EN 模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-1 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建Pangu-SearchPlan-ZH_EN模型部署任务**后，可以查看模型的部署状态。

版权所有 © 华为云计算技术有限公司

当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 可单击模型名称可进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

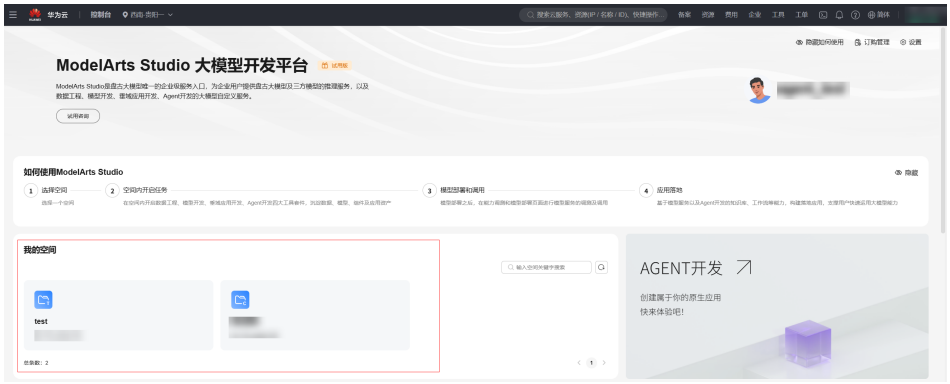
9.2.3 管理 Pangu-SearchPlan-ZH_EN 模型部署任务

模型更新

完成**创建Pangu-SearchPlan-ZH_EN模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

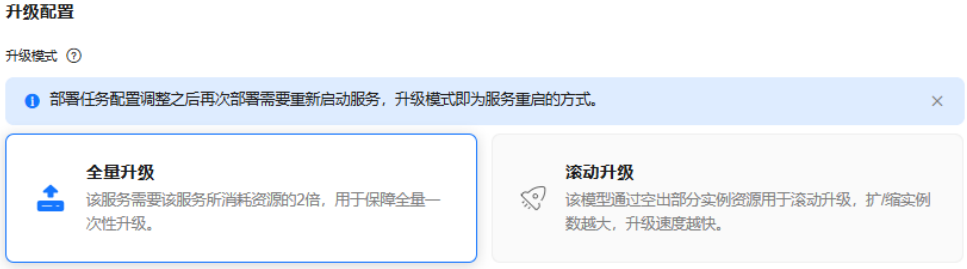
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-2 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 9-3 升级模式



 版权所有 © 华为云计算技术有限公司

说明

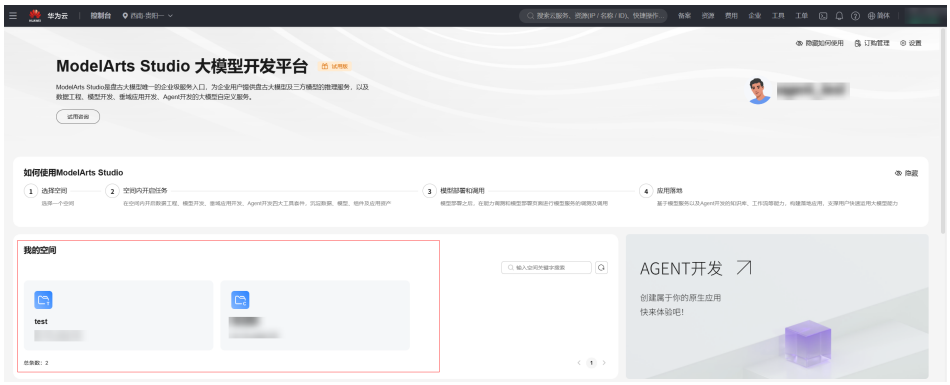
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建Pangu-SearchPlan-ZH_EN模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

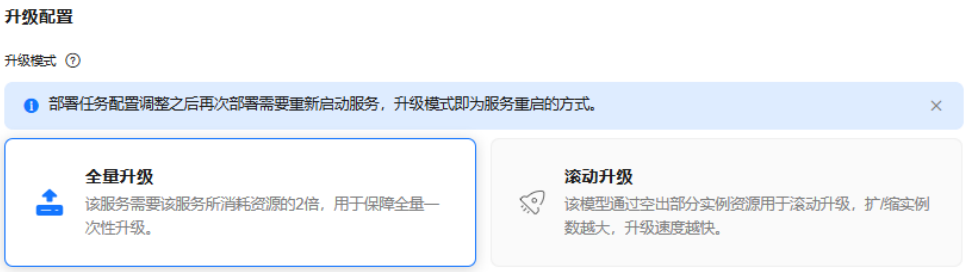
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 9-5 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

版权所有 © 华为云计算技术有限公司

模型部署实例扩缩容

完成**创建Pangu-SearchPlan-ZH_EN模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-6 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-7 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。

版权所有 © 华为云计算技术有限公司

- 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

  说明

  删除属于高危操作，删除前请确保当前任务不再需要。

完成**创建Pangu-SearchPlan-ZH_EN模型部署任务**后，可以对已创建的部署任务进行管理，包括部署任务列表查询、部署任务启动、部署任务停止、部署任务修改、部署任务删除、服务调用统计，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 9-8 我的空间

2. 在左侧导航栏中选择“模型开发 > 模型部署”，进入模型部署任务管理页面。
3. 在管理任务列表单击对应的功能，完成相关操作：
   a. 部署任务列表查询：单击部署任务列表上的搜索框，输入部署任务名称或者根据属性进行筛选，查询对应的任务；
   b. 部署任务启动：对于失败的部署任务，可以单击任务列表操作栏中“启动”按钮，重新启动该部署任务；
   c. 部署任务停止：对于部署中、运行中的部署任务，可以单击任务列表操作栏中“停止”按钮，停止该部署任务；
   d. 部署任务删除：单击任务列表操作栏中“删除”按钮，在删除二次确认弹窗中输入“DELETE”，单击“确认”，完成该部署任务删除
   e. 服务调用统计：对运行中的部署任务，可以单击任务列表操作栏中“服务调用统计”按钮，页面跳转至该部署模型的统计调用页面，可以查看该服务的调用相关数据。

9.3 调用搜索规划大模型

搜索规划属于预置模型，模型部署成功后，可以使用API实现模型调用。

表 9-3 搜索规划大模型 API 清单

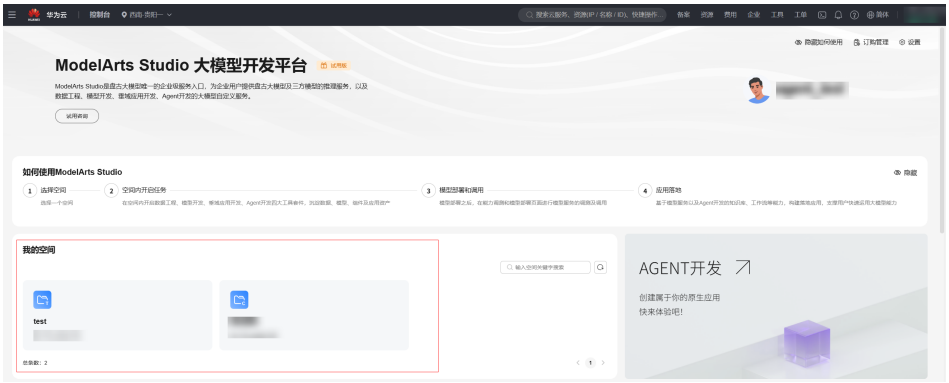
API分类	API访问后缀路径（URI）
search_plan	/app/search/v1/planning

版权所有 © 华为云计算技术有限公司

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 9-9 我的空间



2. URI获取步骤如下：
   a. 登录ModelArts Studio大模型开发平台。
   b. 进入所需工作空间。
   c. 获取请求URI。
      - 获取模型请求URI。
        若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的请求URI。
      - 与表9-3里的后缀进行拼接，获取最终的API请求路径如表9-4。

表 9-4 API 访问路径

API分类	API访问路径（URI）
search_plan	https://xxx.myhuaweicloud.com/v1/{project_id}/infer-api/proxy/service/{deployment_id}/app/search/v1/planning

使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见获取调用路径。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，API Key认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。
   参考表9-5填写请求Header参数。

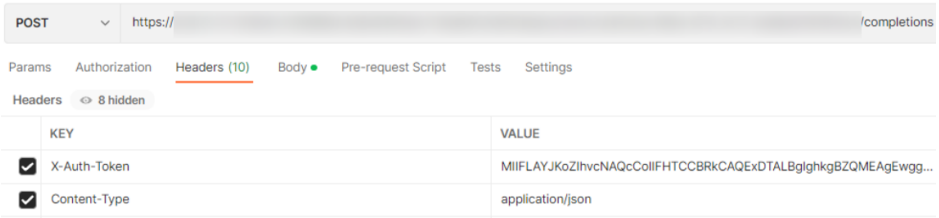
版权所有 © 华为云计算技术有限公司

表 9-5 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
V1推理接口API Key认证	Content-Type	application/json
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。

如图9-10，为Token认证方式的请求Header参数填写示例。

图 9-10 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。
   search_plan接口body样例：

```
{
   "query": "今天天气怎样？",
   "history": []
}
```

4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示搜索规划大模型API调用成功。

API Key 鉴权说明

当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，

可用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为API Key值），API服务仅校验API Key，从而实现快速响应。

版权所有 © 华为云计算技术有限公司

使用该鉴权方式前，请确保有已部署的大模型。

获取API Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。
3. 在“应用配置”中，选择已部署好的大模型，并设置应用接入名称、描述，单击“确定”。
4. 在“应用接入”列表的“API Key”列中可获取API Key值。

 版权所有 © 华为云计算技术有限公司

10 开发盘古向量&重排模型

向量&重排模型介绍
部署向量&重排大模型
调用向量&重排大模型

10.1 向量&重排模型介绍

Pangu-EmbeddingRank模型，用于大模型RAG场景。在ModelArts Studio大模型开发平台部署Pangu-EmbeddingRank模型，会生成Embedding模型服务、Rerank模型服务两个服务。这两个服务均可以在Agent开发平台中创建知识库中被使用。其中Embedding模型服务在创建知识库中，文本处理阶段，用于对文本文档进行切片，转换成向量化表示。在知识检索阶段，根据用户输入的query对切片进行召回，Rerank模型服务用于对召回的切片，按照query与切片的相关度进行精细化排序，以确保召回相关度top切片。

表 10-1 推理特性

推理精度	**FP16**
起推规格（推理单元）	1
QPS	6

10.2 部署向量&重排大模型

10.2.1 创建 Pangu-EmbeddingRank-zh 模型部署任务

平台支持部署预置的Pangu-EmbeddingRank-zh大模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 向量&重排模型 > 模型”，单击“确定”进入“创建部署”页面。

版权所有 © 华为云计算技术有限公司

4. 在“创建部署”页面，参考表10-2完成部署参数设置。

表 10-2 Pangu-EmbeddingRank-zh 模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“向量&重排模型”，并选择需要进行部署的模型和版本。
	部署方式	云上部署：算法部署至平台提供的资源池中。
资源配置	计费模式	包年包月计费模式。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

5. 参数填写完成后，单击“立即部署”。

10.2.2 查看 Pangu-EmbeddingRank-zh 模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-1 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建Pangu-EmbeddingRank-zh模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

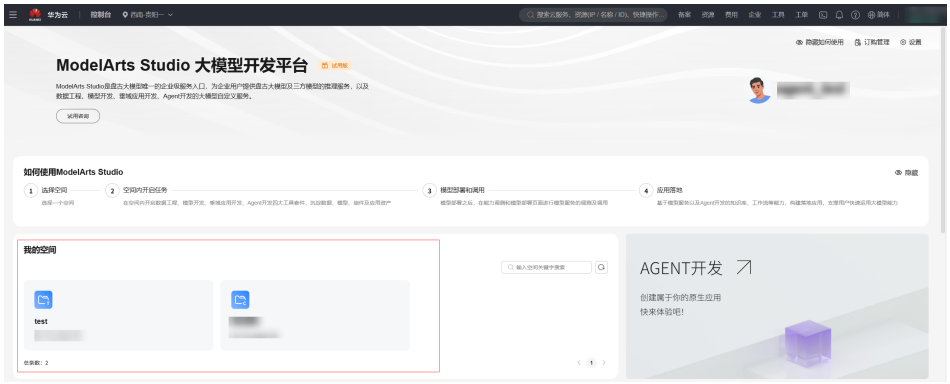
10.2.3 管理 Pangu-EmbeddingRank-zh 模型部署任务

模型更新

完成**创建Pangu-EmbeddingRank-zh模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

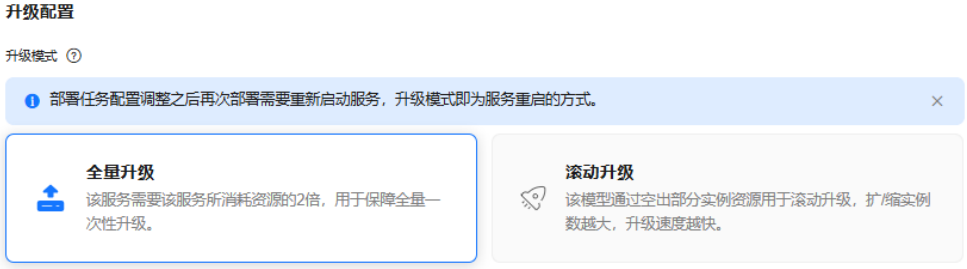
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-2 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 10-3 升级模式



 版权所有 © 华为云计算技术有限公司

说明

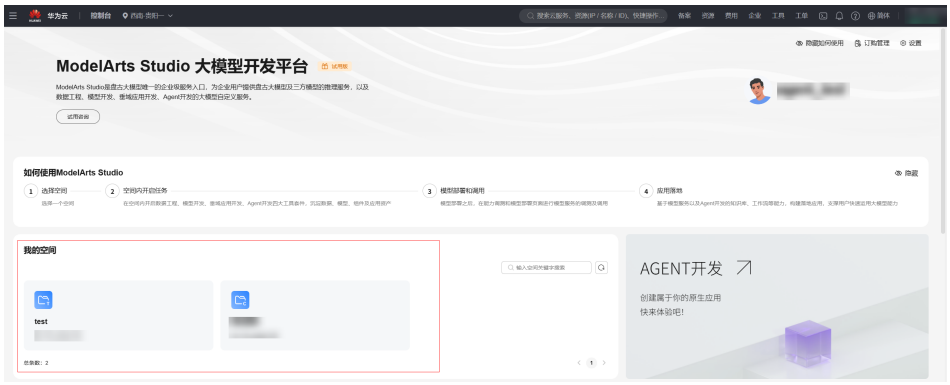
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建Pangu-EmbeddingRank-zh模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

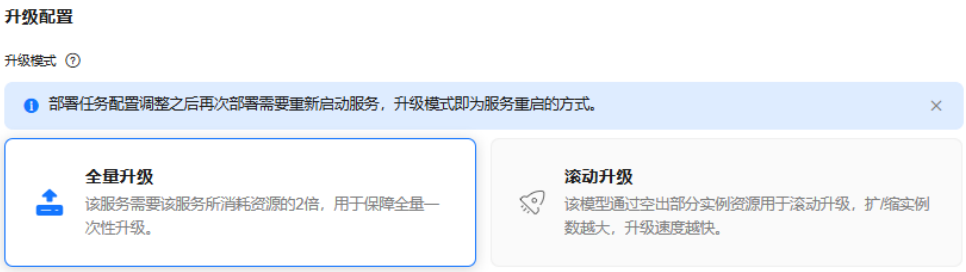
1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 10-5 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

版权所有 © 华为云计算技术有限公司

模型部署实例扩缩容

完成**创建Pangu-EmbeddingRank-zh模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-6 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-7 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。

版权所有 © 华为云计算技术有限公司

- 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

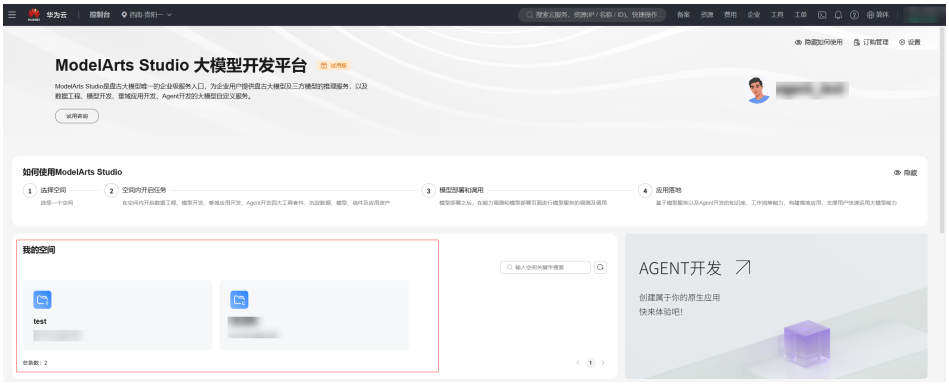
  说明

  删除属于高危操作，删除前请确保当前任务不再需要。

完成**创建Pangu-EmbeddingRank-zh模型部署任务**后，可以对已创建的部署任务进行管理，包括部署任务列表查询、部署任务启动、部署任务停止、部署任务修改、部署任务删除、服务调用统计，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 10-8 我的空间

2. 在左侧导航栏中选择“模型开发 > 模型部署”，进入模型部署任务管理页面。
3. 在管理任务列表单击对应的功能，完成相关操作：
   a. 部署任务列表查询：单击部署任务列表上的搜索框，输入部署任务名称或者根据属性进行筛选，查询对应的任务；
   b. 部署任务启动：对于失败的部署任务，可以单击任务列表操作栏中“启动”按钮，重新启动该部署任务；
   c. 部署任务停止：对于部署中、运行中的部署任务，可以单击任务列表操作栏中“停止”按钮，停止该部署任务；
   d. 部署任务删除：单击任务列表操作栏中“删除”按钮，在删除二次确认弹窗中输入“DELETE”，单击“确认”，完成该部署任务删除
   e. 服务调用统计：对运行中的部署任务，可以单击任务列表操作栏中“服务调用统计”按钮，页面跳转至该部署模型的统计调用页面，可以查看该服务的调用相关数据。

10.3 调用向量&重排大模型

向量&重排属于预置模型，模型部署成功后，可以使用API实现模型调用。

表 10-3 向量&重排大模型 API 清单

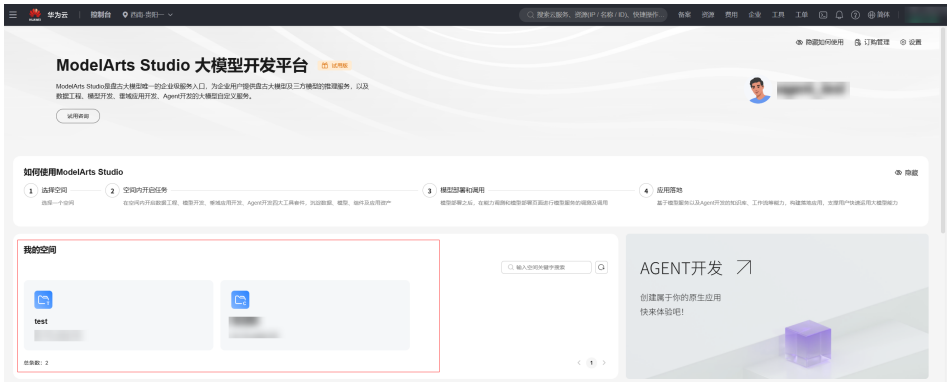
API分类	API访问后缀路径（URI）
embeddings	/app/search/v1/vector/query
rerank	/app/search/v1/rerank

版权所有 © 华为云计算技术有限公司

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 10-9 我的空间



2. URI获取步骤如下：
   a. 登录ModelArts Studio大模型开发平台。
   b. 进入所需工作空间。
   c. 获取请求URI。
      - 获取模型请求URI。
        - 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的请求URI。

图 10-10 部署后的模型调用路径前缀



      - 与表10-3里的后缀路径进行拼接，获取最终的API请求路径如表10-4。

表 10-4 拼接后的完整 URI

API分类	API访问路径（URI）
embeddings	https://xxx.myhuaweicloud.com/v1/{project_id}/infer-api/proxy/service/{deployment_id}/app/search/v1/vector/query

版权所有 © 华为云计算技术有限公司

API分类	API访问路径（URI）
rerank	https://xxx.myhuaweicloud.com/v1/{project_id}/infer-api/proxy/service/{deployment_id}/app/search/v1/rerank

使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见**获取调用路径**。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，**API Key认证**的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。

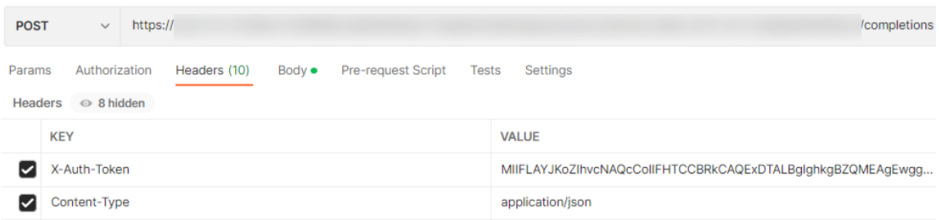
   参考**表10-5**填写请求Header参数。

表 10-5 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
API Key认证	Content-Type	application/json
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。

如**图10-11**，为Token认证方式的请求Header参数填写示例。

图 10-11 配置请求参数



版权所有 © 华为云计算技术有限公司

3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。
   - embedding接口body样例：
     ```
     {
        "query": "南京"
     }
     ```
   - rerank接口body样例：
     ```
     {
        "query": "南京",
        "ranking_order": ["title", "content"],
        "docs": [
           {
              "id": "a01",
              "title": "南京",
              "content": "简介：南京一般指南京市。 南京市，简称"
           },
           {
              "id": "a02",
              "title": "南京",
              "content": "南京市人民政府关于市政府领导同志工作分工的通知 · 南京市人民政府2021年政府信息公开工作年度报告 "
           },
           {
              "id": "a03",
              "title": "南京",
              "content": "唐代以成都为南京不到三年时间。 [1] 宋大中祥符七年（1014年），建应天府（今商丘）为南京。 [2-5] 辽代，以北京为南京。金代，以开封为南京。明永乐十九年（1421年），明成祖迁都，京师应天府（今南京）改称为南京，作为留都。"
           }
        ],
        "top_n": 10
     }
     ```
4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示向量&重排大模型API调用成功。

API Key 鉴权说明

当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，

可用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为API Key值），API服务仅校验API Key，从而实现快速响应。

使用该鉴权方式前，请确保有已部署的大模型。

获取APPI Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。
3. 在“应用配置”中，选择已部署好的大模型，并设置应用接入名称、描述，单击“确定”。
4. 在“应用接入”列表的“API Key”列中可获取API Key值。

 版权所有 © 华为云计算技术有限公司

11 开发盘古行业大模型

盘古行业NLP大模型

11.1 盘古行业 NLP 大模型

11.1.1 盘古行业 NLP 大模型介绍

盘古行业大模型在盘古NLP大模型的基础之上结合了垂直行业的海量数据，配套MA Studio平台和工具链，可化解特定行业数据构建难、模型效果调优难、场景能力评测难三大难题，助力行业客户更快、更优构建专业大模型。当前已于公有云发布金融、医疗行业大模型，满足行业通用场景开箱即用。

11.1.2 使用数据工程构建盘古行业 NLP 大模型数据集

盘古行业 NLP 大模型支持接入的数据集类型

盘古NLP大模型仅支持接入文本类数据集，数据集文件内容包括：预训练文本、单轮问答、多轮问答、带人设单轮问答、带人设多轮问答等，不同训练方式所需要使用的数据见**表11-1**，该数据集格式要求请参见**文本类数据集格式要求**。

表 11-1 训练盘古行业 NLP 大模型数据集类型要求

基模型	训练场景	数据集类型	数据集内容	文件格式
NLP	预训练	文本	预训练文本	jsonl
	微调	文本	单轮问答	jsonl、csv
		文本	多轮问答	jsonl
		文本	单轮问答（人设）	jsonl、csv
		文本	多轮问答（人设）	jsonl

版权所有 © 华为云计算技术有限公司

训练盘古行业 NLP 大模型所需数据量

使用数据工程构建盘古NLP大模型数据集进行模型训练时，所需数据量见**表11-2**。

表 11-2 构建盘古行业 NLP 大模型所需数据量

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N1	微调	-	1000条/每场景	≥ 1万条/每场景	32K
N2	微调	-	1000条/每场景	≥ 1万条/每场景	32K
N4	微调	-	1000条/每场景	≥ 1万条/每场景	4K版本：4096 32K版本：32768
	预训练	4GB(等价10亿Tokens)	-	-	4K版本：4096 32K版本：32768

评测盘古行业 NLP 大模型所需数据量

要求所有文本大小最大不超过100MB，目录下文件数量最多不超过100个。数据条数范围为：3-1000条。

构建盘古行业 NLP 大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程构建盘古NLP大模型数据集流程见**表3 盘古行业NLP大模型数据集构建流程**。

表 11-3 盘古行业 NLP 大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	**导入数据至盘古平台**
加工文本类数据集	加工文本类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	**加工文本类数据集**

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	合成文本类数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。	合成文本类数据集
	标注文本类数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。	标注文本类数据集
	配比文本类数据集	数据配比是将多个数据集按特定比例组合的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。	配比文本类数据集
发布文本类数据集	评估文本类数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。	评估文本类数据集
	发布文本类数据集	发布流程是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。 • 标准格式：平台默认的格式。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。	发布文本类数据集

11.1.3 训练盘古行业 NLP 大模型

盘古行业 NLP 大模型训练流程介绍

盘古行业NLP大模型的训练分为两个关键阶段：预训练和微调。

- **预训练阶段：**在这一阶段，模型通过学习大规模通用数据集来掌握语言的基本模式和语义。这一过程为模型提供了处理各种语言任务的基础，如阅读理解、文本生成和情感分析，但它还未能针对特定任务进行优化。

  针对**预训练阶段**，还可以继续进行训练，这一过程称为**增量预训练**。增量预训练是在已经完成的预训练的基础上继续训练模型。增量预训练旨在使模型能够适应新的领域或数据需求，保持其长期的有效性和准确性。

 版权所有 © 华为云计算技术有限公司

- **微调阶段：**基于预训练的成果，微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。这一阶段使模型能够精确执行如文案生成、代码生成和专业问答等特定场景中的任务。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。

  针对**微调阶段**，微调后的模型可以直接部署。

盘古行业 NLP 大模型选择建议

选择合适的NLP大模型类型有助于提升训练任务的准确程度。您可以根据模型**可处理最大Token长度**，选择合适的模型，从而提高模型的整体效果，详见**表11-4**。

此外，不同类型的NLP大模型在训练过程中，读取中文、英文内容时，字符长度转换为Token长度的转换比有所不同，详见**表11-5**。

表 11-4 盘古行业 NLP 大模型参数规格

模型名称	可处理最大上下文长度	可处理最大输出长度	说明
Pangu-Medical-NLP-N2-32K-3.1.35	32K	4K	2025年3月发布版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。此模型版本差异化支持INT8量化特性。
Pangu-Medical-NLP-N2-Reasoner-32K-3.0.0.1	32K	4K	医学大模型首个reasoner版本，较盘古基础大模型reasoner版本在医学行业领域效果更优。
Pangu-Finance-NLP-Reasoner-N2-128K-3.0.1.1	128K	4K	金融大模型首个reasoner版本，较盘古基础大模型reasoner版本在金融行业领域效果更优。

表 11-5 Token 转换比

模型规格	Token比（Token/英文单词）	Token比（Token/汉字）
N2	0.88	1.24

版权所有 © 华为云计算技术有限公司

说明

针对Token转换比，平台提供了**Token计算器**功能，可以根据您输入的文本计算Token数量，您可以通过以下方式使用该功能：

- 在左侧导航栏选择“体验中心”，单击“选择其他服务”，选择预置大模型或自己的大模型后，单击右上角“Token计算器”使用该功能。
- 使用API调用Token计算器，详见《API参考》“API > Token计算器”。

盘古行业 NLP 大模型训练类型选择建议

平台针对盘古行业NLP大模型提供了两种训练类型，包括预训练、微调，两者区别详见表11-6。

表 11-6 预训练、微调训练类型区别

训练方式	训练目的	训练数据	模型效果	应用场景举例
预训练	**关注通用性**：预训练旨在让模型学习广泛的通用知识，建立词汇、句法和语义的基础理解。通过大规模的通用数据训练，模型可以掌握丰富的语言模式，如语言结构、词义关系和常见的句型。	**使用大规模通用数据**：通常使用海量的无监督数据（如文本语料库、百科文章），这些数据覆盖广泛的领域和语言表达方式，帮助模型掌握广泛的知识。	**适合广泛应用**：经过预训练后，模型可以理解自然语言并具备通用任务的基础能力，但还没有针对特定的业务场景进行优化。预训练后的模型主要用于多个任务的底层支持。	通过使用海量的互联网文本语料对模型进行预训练，使模型理解人类语言的基本结构。
微调	**关注专业性**：微调是对预训练模型的参数进行调整，使其在特定任务中达到更高的精度和效果。微调的核心在于利用少量的特定任务数据，使模型的表现从通用性向具体任务需求过渡。	**使用小规模的特定任务数据**：微调通常需要小规模但高质量的标注数据，直接与目标任务相关。通过这些数据，模型可以学习到任务特定的特征和模式。	**在特定任务上具有更高的准确性**：微调后的模型在具体任务中表现更优。相较于预训练阶段的通用能力，微调能使模型更好地解决细分任务的需求。	在一个客户服务问答系统中，可以用特定领域（如电商、保险）的对话数据对预训练模型进行微调，使其更好地理解和回答与该领域相关的问题。

此外，针对微调训练任务，平台提供了两种微调方式：

- **全量微调**：适合有充足数据并关注特定任务性能的场景。在全量微调中，模型的所有参数都会调整，以适应特定任务的需求。这种方式适合样本量较大、对推理效果要求较高的任务。例如，在特定领域（如金融、医疗）中，若拥有大量标注数据，且需要更高的特定任务推理精度，则全量微调是优先选择。
- **LoRA微调**：适用于数据量较小、侧重通用任务的情境。LoRA（Low-Rank Adaptation）微调方法通过调整模型的少量参数，以低资源实现较优结果，适合

 版权所有 © 华为云计算技术有限公司

聚焦于领域通用任务或小样本数据情景。例如，在针对通用客服问答的场景中，样本量少且任务场景广泛，选择LoRA微调既能节省资源，又能获得较好的效果。

微调方式选择建议：

- 若项目中数据量有限或任务场景较为广泛，可以选择**LoRA微调**以快速部署并保持较高适用性。
- 若拥有充足数据且关注特定任务效果，选择**全量微调**有助于大幅提升在特定任务上的模型精度。

11.1.3.1 创建盘古行业 NLP 大模型训练任务

创建盘古行业 NLP 大模型预训练任务

创建NLP大模型预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-1 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表11-7**完成训练参数设置。

表 11-7 盘古行业 NLP 大模型预训练参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	数据批量大小	数据集进行分批读取训练，设定每个批次数据的大小。 通常情况下，较大的数据批量可以使梯度更加稳定，从而有利于模型的收敛。然而，较大的数据批量也会占用更多的显存资源，这可能导致显存不足，并且会延长每次训练的时长。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
	模型保存策略	save_checkpoint_steps/save_checkpoint_epoch，训练过程中是按迭代步数，还是训练轮数保存Checkpoint文件。
	Checkpoint保存间隔	save_checkpoint_steps，训练过程中每隔多少个训练步长保存一次模型Checkpoint文件。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	Checkpoint保存轮数	save_checkpoint_epoch，训练过程中每个多少训练轮数保存一次模型Checkpoint文件。
	数据预处理并发个数	定义了在预处理数据时，能够同时处理文件的并行进程数量。设定这个参数的主要目的是通过并发处理来加速数据预处理，从而提升训练效率。
训练数据配置	训练集	选择训练模型所需的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
故障自动重启	自动重启	该功能开启后，当训练作业发生故障终止本次作业时，自动从本次中断的作业继续训练。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建盘古行业 NLP 大模型增量预训练任务

在模型完成**创建盘古行业NLP大模型预训练任务**预训练后，可以对训练后的模型继续训练，该过程称为“增量预训练”。

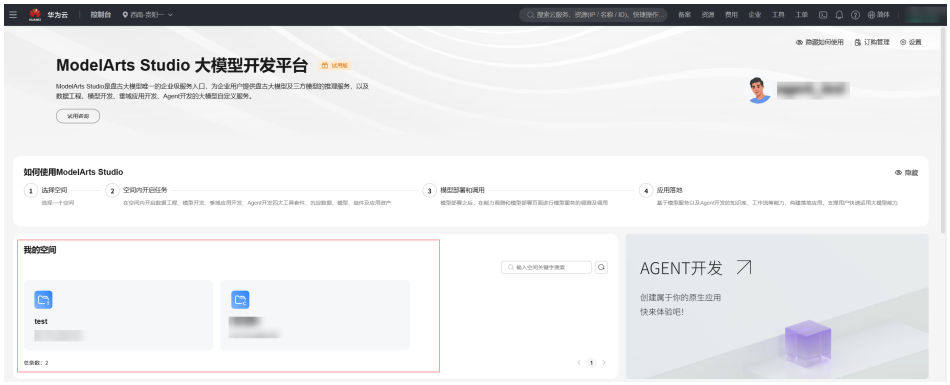
创建盘古行业NLP大模型增量预训练任务前，请确保有已完成预训练的盘古行业NLP大模型。

创建盘古行业NLP大模型增量预训练任务的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 11-2 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“我的资产 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 其余参数配置等步骤同**创建NLP大模型预训练任务**。

创建盘古行业 NLP 大模型全量微调任务

创建盘古行业NLP大模型全量微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-3 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表11-8**完成训练参数设置。

版权所有 © 华为云计算技术有限公司

表 11-8 盘古行业 NLP 大模型全量微调参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 ● 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	单步迭代时处理的数据批量大小	指定每次迭代时处理的数据批量大小。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。

 版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。
	模型保存策略	save_checkpoint_steps/save_checkpoint_epoch，训练过程中是按迭代步数，还是训练轮数保存Checkpoint文件。
	Checkpoint保存间隔	save_checkpoint_steps，训练过程中每隔多少个训练步长保存一次模型Checkpoint文件。
	Checkpoint保存轮数	save_checkpoint_epoch，训练过程中每个多少训练轮数保存一次模型Checkpoint文件。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
训练数据配置	训练集	选择训练模型所需的数据集。
	验证集	● 若选择“分割训练集”，则需进一步配置数据拆分比例。 ● 若选择“选择数据集”，则需选择导入的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
故障自动重启	自动重启	该功能开启后，当训练作业发生故障终止本次作业时，自动从本次中断的作业继续训练。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同模型训练参数默认值存在一定差异，请以前端页面展示的默认值为准。

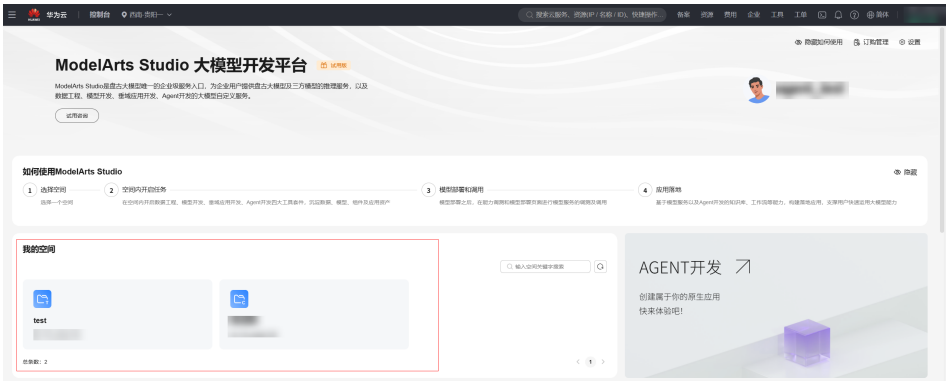
5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建行业 NLP 大模型 LoRA 微调任务

创建NLP大模型LoRA微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考表11-9完成训练参数设置。

 版权所有 © 华为云计算技术有限公司

表 11-9 NLP 大模型 LoRA 微调参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“微调”。
	训练目标	选择“LoRA微调”。 ● LoRA微调：在模型微调过程中，只对特定的层或模块的参数进行更新，而其余参数保持冻结状态。这种方法可以显著减少计算资源和时间消耗，同时在很多情况下，依然能够保持较好的模型性能。
训练参数	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	LoRA矩阵中的秩	lora_rank，在Lora矩阵中，Rank的值用于衡量矩阵的复杂度和信息量。数值较大，增强模型的表示能力，但会增加训练时长；数值越小可以减少参数数量，降低过拟合风险。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
训练数据配置	训练集	选择训练模型所需的数据集。
	验证集	● 若选择“分割训练集”，则需进一步配置数据拆分比例。 ● 若选择“选择数据集”，则需选择导入的数据集。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同模型训练参数默认值存在一定差异，请以前端页面展示的默认值为准。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

11.1.3.2 查看盘古行业 NLP 大模型训练任务状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见**表 训练状态说明**。

版权所有 © 华为云计算技术有限公司

表 11-10 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见**表2 训练指标说明**。

图 11-5 查看训练指标

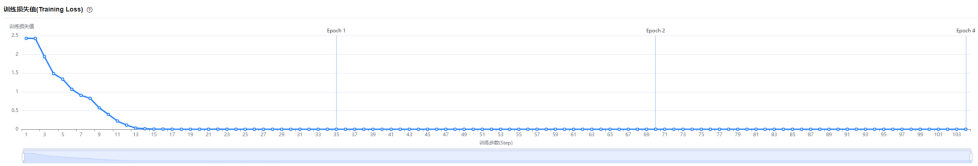


表 11-11 训练指标说明

模型	训练指标	指标说明
NLP大模型	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。

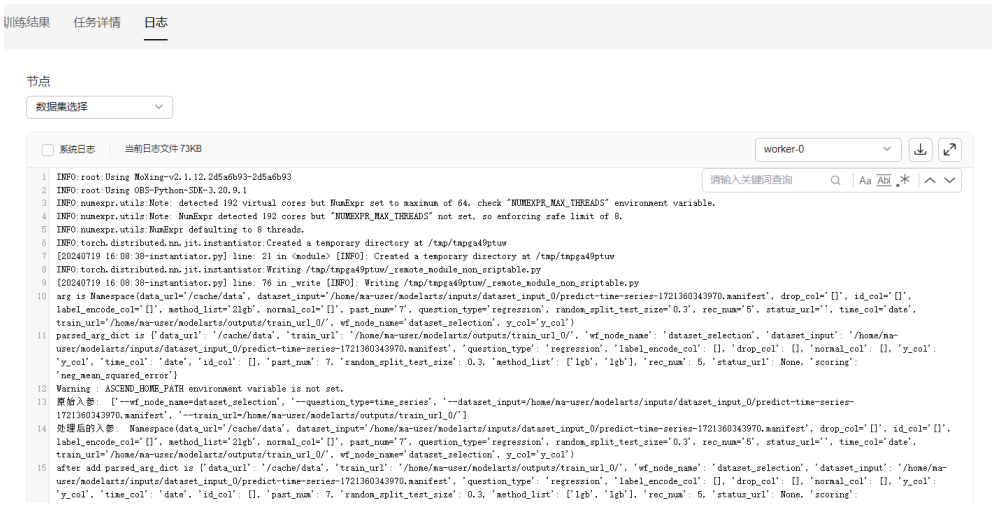
获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

 版权所有 © 华为云计算技术有限公司

图 11-6 获取训练日志



训练异常定位方法

模型训练失败时，可以按照如下定位思路对训练任务做初步定位。

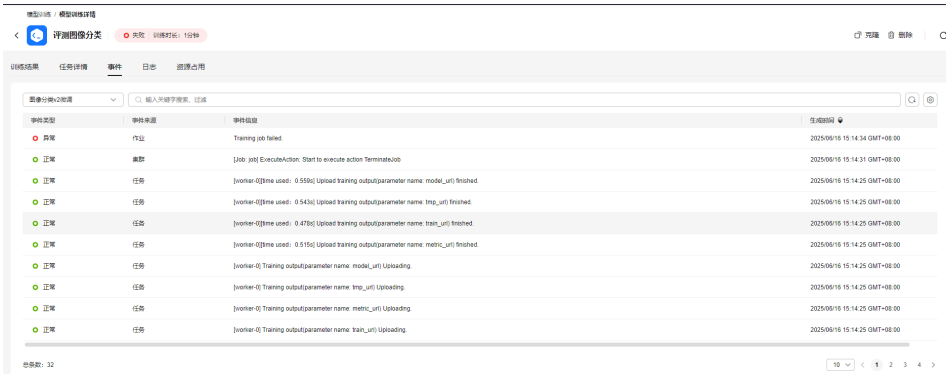
1. 在训练任务列表查看任务失败原因。

   模型启动训练后，可以在模型训练列表中查看训练任务的状态，当任务状态为“失败”时，可将鼠标放置在“失败”上，可以查看到任务失败的概要信息。



2. 单击任务名，进入任务详情页查看训练任务运行关键事件信息。

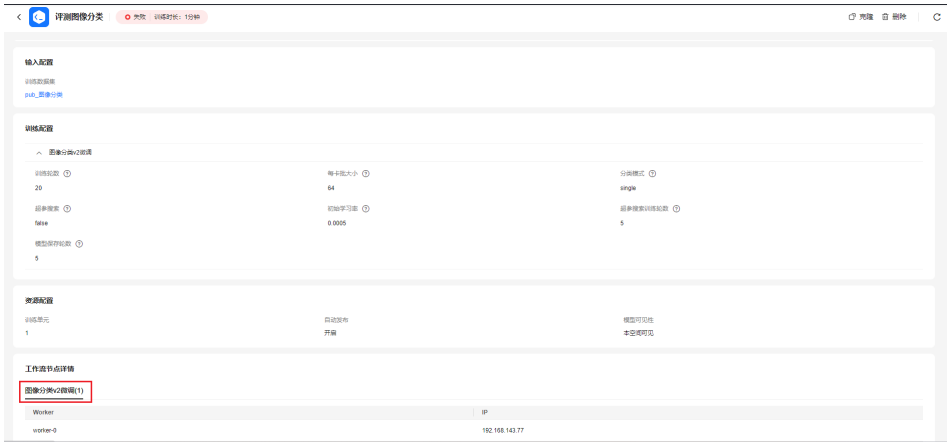
   单击“事件”页签，可查看训练任务子任务运行事件，通过事件可以观察到当前任务的运行情况。事件可以按照“正常、异常、告警”三类事件类型过滤。



3. 获取当前任务运行的worker、workerIP以及任务异常退出时的节点和退出码信息。

   在ModelArts Studio训练任务页面单击需要查看的任务名，进入“任务详情”页，在“工作流节点详情”下可以查看到当前任务所运行的节点及IP信息。

版权所有 © 华为云计算技术有限公司



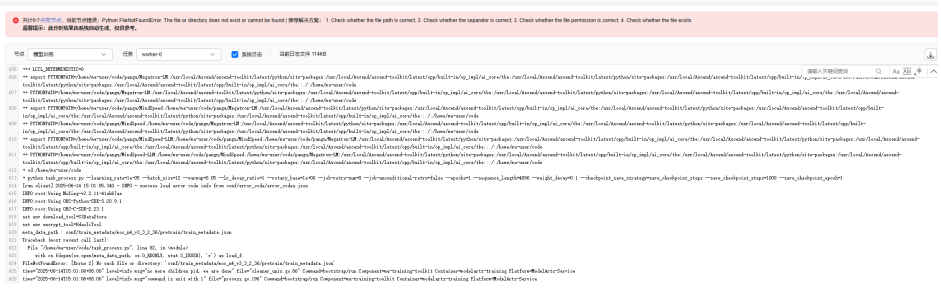
4. 当训练任务运行失败时，可查看“模型训练详情 -> 基本信息 -> 状态”，单击“查看失败节点清单”获取训练任务故障节点及退出码信息。



5. 训练作业遇到问题时，可查看“模型训练详情 -> 日志”，多数场景下的问题可以通过日志报错信息直接定位。

   如果训练作业运行失败，ModelArts Studio会自动识别导致作业失败的原因，在训练日志界面上给出提示并对部分常见训练错误给出分析建议。包含失败的可能原因、推荐的解决方案。

图 11-7 训练错误分析建议



说明

针对分布式作业，只会显示当前节点的一个分析结果，作业的失败需要综合各个节点的失败原因做一个综合判断。

11.1.3.3 发布训练后的盘古行业 NLP 大模型

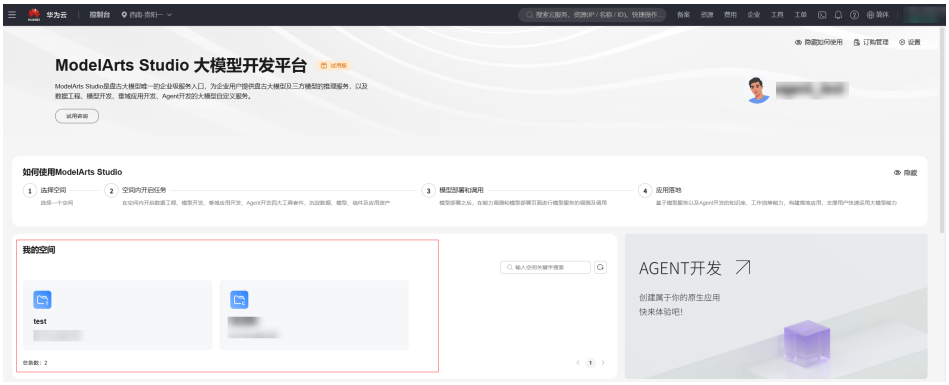
发布训练后的盘古行业NLP大模型

盘古行业NLP大模型训练完成后，需要执行发布操作，操作步骤如下：

 版权所有 © 华为云计算技术有限公司

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-8 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 11-9 训练结果页面



4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

   说明

   1、发布后的模型，支持继续进行训练&模型部署

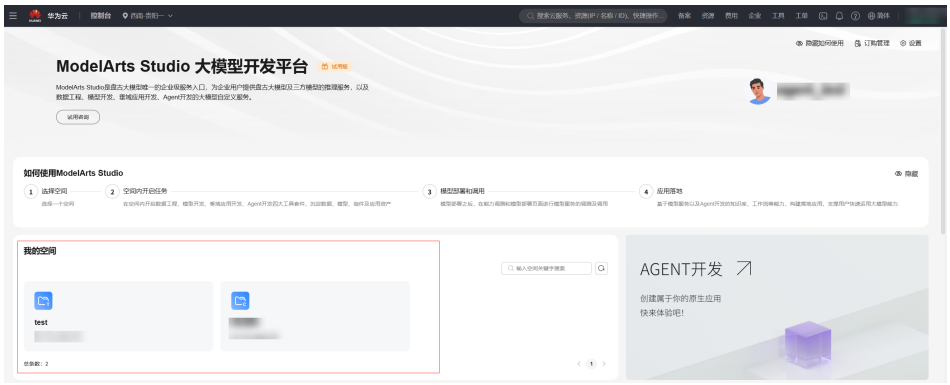
11.1.3.4 管理盘古行业 NLP 大模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 11-10 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 编辑。单击操作列的“编辑”，可以修改模型的checkpoints、训练参数、训练数据以及基本信息等。
   - 克隆。单击操作列的“克隆”，可以复制当前训练任务。
   - 停止。单击操作列的“停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“删除”，可以删除当前不需要的训练任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

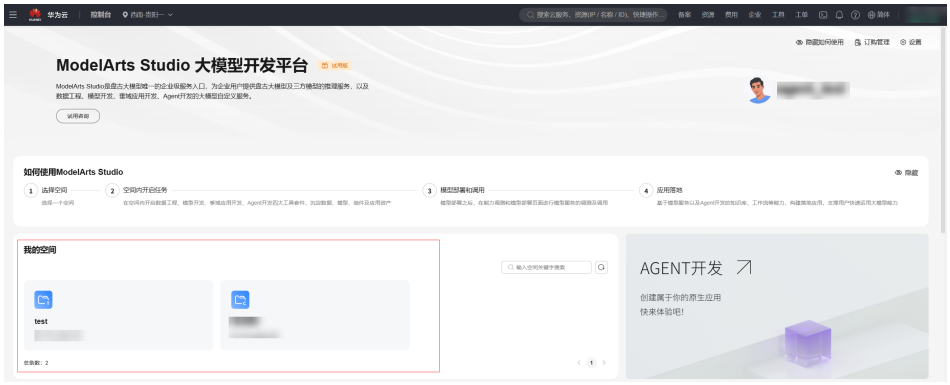
11.1.4 压缩盘古行业 NLP 大模型

模型在部署前，通过模型压缩可以降低推理显存占用，节省推理资源提高推理性能。

平台当前仅可对NLP大模型进行压缩，**支持压缩的模型清单请详见**《产品介绍》> “模型能力与规格 > 盘古NLP大模型能力与规格”。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-11 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型压缩”，单击界面右上角“创建压缩任务”。

版权所有 © 华为云计算技术有限公司

3. 在“创建压缩任务”页面，选择需要压缩的基础模型，支持选择已发布模型或未发布模型。
4. 选择压缩策略。除INT8压缩策略外，部分模型支持INT4压缩策略，可在选择模型后，根据页面展示的策略进行选择。
   - INT8：该压缩策略将模型参数压缩至8位字节，可以有效降低推理显存占用。
   - INT4：该压缩策略与INT8相比，可以进一步减少模型的存储空间和计算复杂度。
5. 配置资源。选择计费模式并设置训练单元。
6. 可选择开启订阅提醒。开启后，系统将在本次压缩任务状态变更时，向用户发送短信/邮件提醒。
7. 选择开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
8. 填写基本信息，包括任务名称、压缩后模型名称与描述，单击“立即创建”。

   当压缩任务状态为“已完成”时，表示模型已完成压缩操作。

11.1.5 部署盘古行业 NLP 大模型

云上部署

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 11-12 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表1 行业NLP大模型部署参数说明**完成部署参数设置。

版权所有 © 华为云计算技术有限公司

表 11-12 盘古行业 NLP 大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择需要进行部署的模型和版本。
	最大TOKEN长度	模型可最大请求的上下文TOKEN数。
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。 – 实例数：设置部署模型时所需的实例数。
	自定义名称	此名称是通过V2版本推理接口调用该推理服务时的唯一标识。创建后不支持修改。
安全护栏	安全护栏开关	安全护栏保障模型调用安全。
	版本选择	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置（选择云上部署时）	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

5. 参数填写完成后，单击“立即部署”。

 版权所有 © 华为云计算技术有限公司

11.1.6 调用盘古行业 NLP 大模型

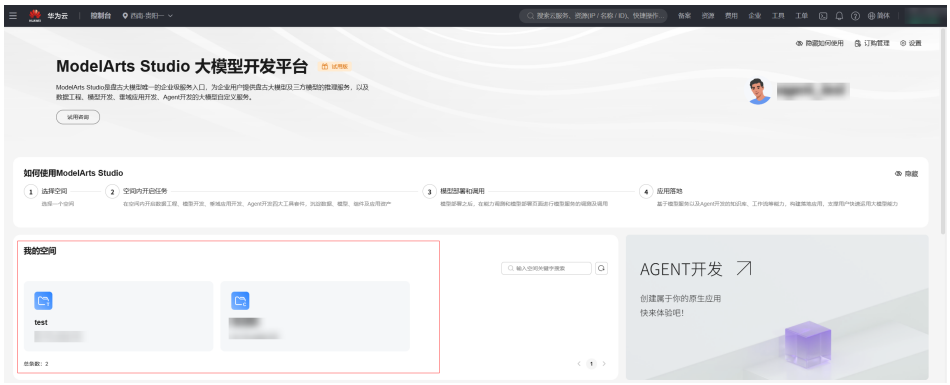
11.1.6.1 使用“体验中心”调用盘古行业 NLP 大模型

体验中心功能支持用户调用预置或训练后的NLP大模型。使用该功能前，请完成模型的部署操作，步骤详见**云上部署**。

使用“体验中心”调用盘古行业NLP大模型可实现文本对话能力，即在输入框中输入问题，模型将基于问题输出相应的回答，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-13 我的空间



2. 在左侧导航栏中选择“体验中心”，单击“文本对话”页签。
3. 选择需要调用的服务。可从“预置服务”或“我的服务”中选择。
4. 填写系统人设。如“你是一个AI助手”，若不填写，将使用系统默认人设。
5. 在页面右侧配置参数，具体参数说明见**表11-13**。

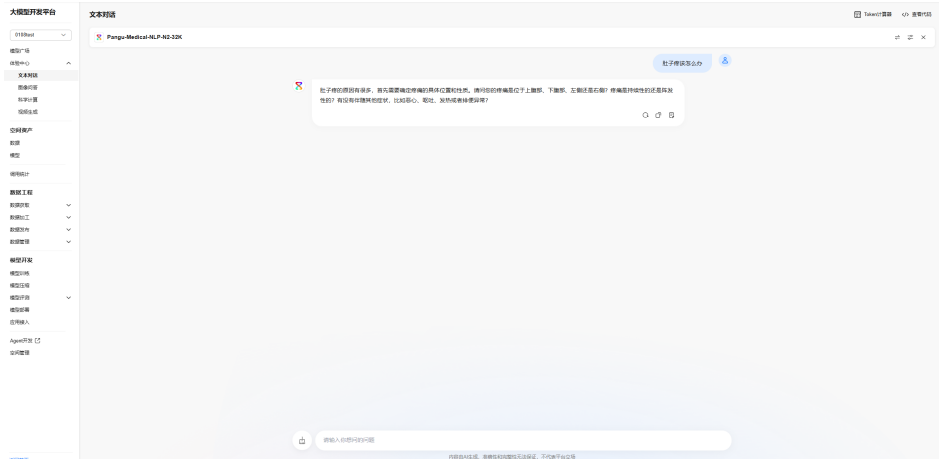
表 11-13 盘古行业 NLP 大模型体验中心参数说明

参数	说明
温度	用于控制生成文本的多样性和创造力。调高温度会使得模型的输出更多样性和创新性。
核采样	控制生成文本多样性和质量。调高核采样可以使输出结果更加多样化。
话题重复度控制	用于控制生成文本中的重复程度。调高参数模型会更频繁地切换话题，从而避免生成重复内容。
词汇重复度控制	用于调整模型对频繁出现的词汇的处理方式。调高参数会使模型减少相同词汇的重复使用，促使模型使用更多样化的词汇进行表达。
输出最大TOKEN长度	用于控制模型生成回复的长度，一般来说，较大的max_tokens值可以生成较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较简洁的回复，但也可能导致生成不完整或不连贯的内容，请根据场景选择合适的值。

版权所有 © 华为云计算技术有限公司

6. 如图2 调测行业NLP大模型，输入对话，单击“生成”，模型将输出相应的回答。

图 11-14 调测行业 NLP 大模型



11.1.6.2 使用 API 调用盘古行业 NLP 大模型

预置模型或训练后的模型部署成功后，可以使用API实现模型调用。

表 11-14 NLP 大模型 API 清单

API分类	API访问路径（URI）
文本对话	/v1/{project_id}/deployments/{deployment_id}/chat/completions

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 11-15 我的空间

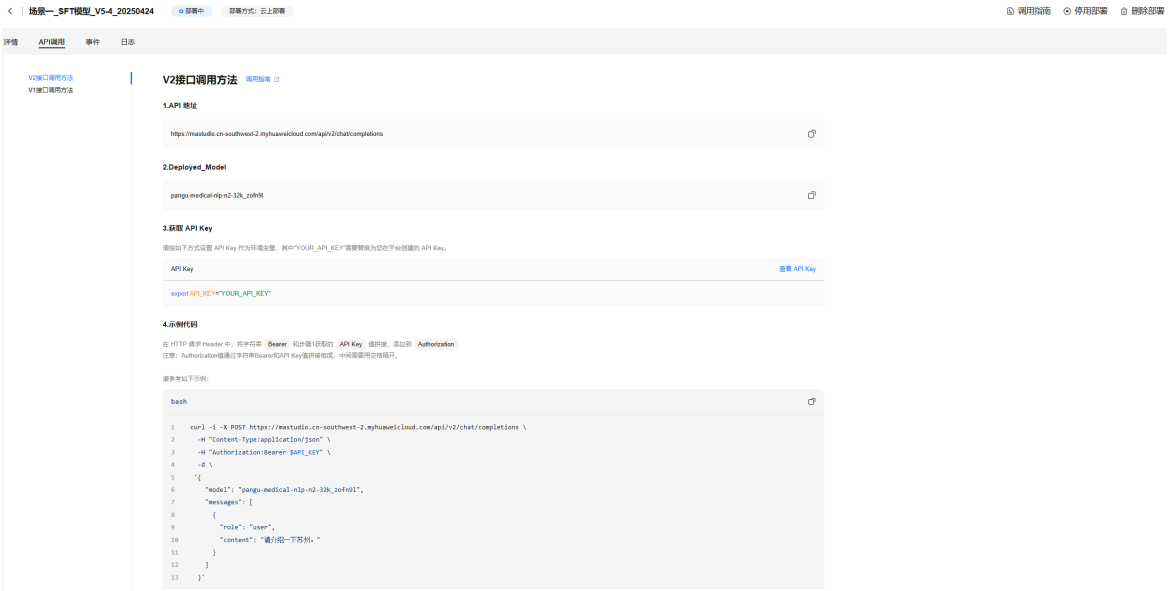


2. 获取调用路径。
   在左侧导航栏中选择“模型开发 > 模型部署”。

版权所有 © 华为云计算技术有限公司

- 获取已部署模型的调用路径。在“API调用”页签，可获取模型调用路径，如图11-16。

图 11-16 获取已部署模型的调用路径



使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见**获取调用路径**。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，API Key认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。

   参考**表11-15**填写请求Header参数。

表 11-15 请求 Header 参数填写说明

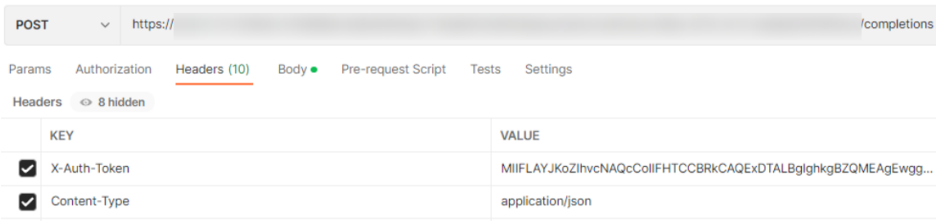
认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
API Key认证	Content-Type	application/json

版权所有 © 华为云计算技术有限公司

认证方式	参数名	参数值
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。

如图11-17，为Token认证方式的请求Header参数填写示例。

图 11-17 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "messages": [
    {
        "content": "介绍下长江，以及长江中典型的鱼类"
    }
    ],
   "temperature": 0.9,
   "max_tokens": 600
}
```

4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示NLP大模型API调用成功。

API Key 鉴权说明

当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，

可用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为API Key值），API服务仅校验API Key，从而实现快速响应。

使用该鉴权方式前，请确保有已部署的大模型。

获取APPI Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。

版权所有 © 华为云计算技术有限公司

3. 在“应用配置”中，选择已部署好的大模型，并设置应用接入名称、描述，单击“确定”。
4. 在“应用接入”列表的“API Key”列中可获取API Key值。

 版权所有 © 华为云计算技术有限公司

12 开发 Deepseek 大模型

Deepseek大模型介绍
使用数据工程构建Deepseek大模型数据集
部署Deepseek大模型
评测Deepseek大模型
调用Deepseek大模型

12.1 Deepseek 大模型介绍

平台预置DeepSeek-R1-32K及DeepSeek-V3-32K模型。

DeepSeek-R1

DeepSeek-R1满血版671B参数模型，在数学、代码、自然语言处理能力比肩OpenAI o1正式版。DeepSeek-R1-32K-0.0.1版本支持推理部署。

平台已预置推理服务，用户可通过体验中心或者API调用该推理服务。

直接部署模型时，已预置最优模型，无需进行压缩后再部署。

表 12-1 推理特性

推理精度	FP16	INT8
序列长度	32K	32K
起推规格（推理单元）	32	16
并发	256	256

DeepSeek-V3

DeepSeek-V3为MoE模型，671B参数，激活37B，在长文本、代码、数学、百科、中文能力上表现优秀。DeepSeek-V3-32K-0.0.1版本支持推理部署。

版权所有 © 华为云计算技术有限公司

直接部署模型时，已预置最优模型，无需进行压缩后再部署。

表 12-2 推理特性

推理精度	FP16	INT8
序列长度	32K	32K
起推规格（推理单元）	32	16
并发	256	256

12.2 使用数据工程构建 Deepseek 大模型数据集

构建 Deepseek 大模型数据集流程

在ModelArts Studio平台中，使用数据工程构建三方大模型数据集流程见**表12-3**。

表 12-3 三方大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务或本地的数据导入至平台统一管理，用于后续加工或发布操作。 **说明** 执行导入操作时，数据集类型需要选择**“单轮问答数据”**。	**导入数据至盘古平台**
加工其他类数据集	加工其他类数据集	通过自定义加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。	**加工其他类数据集**
发布其他类数据集	发布其他类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。	**发布其他类数据集**

12.3 部署 Deepseek 大模型

模型训练完成后，可以启动模型的部署操作。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表12-4**完成部署参数设置，启动模型部署。

版权所有 © 华为云计算技术有限公司

表 12-4 三方大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择需要进行部署的模型和版本。
	部署方式	云上部署：算法部署至平台提供的资源池中。
	Deployed_model	通过V2版本推理接口调用该推理服务时的唯一标识。
安全护栏	开启并同意授权	安全护栏保障模型调用安全。
	版本选择	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置	计费模式	限时免费。
	实例数	设置部署模型是所需的实例数，单次部署服务时，部署实例个数建议不大于10，否则可能触发限流导致部署失败。
安全护栏	开启/关闭	试用版场景默认开启安全护栏，模型推理时有效拦截有害内容。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

5. 参数填写完成后，单击“立即部署”。

12.4 评测 Deepseek 大模型

12.4.1 创建 Deepseek 大模型评测数据集

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同情景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

Deepseek大模型支持人工评测、自动评测两种评测模式。

- 人工评测：通过人工创建的评测数据集和评测指标项对模型生成的回答进行评测，评测时需要人工基于创建好的评测项对模型回答进行打分，评测完成后会基于打分结果生成评测报告。
- 自动评测：包含“基于规则”与“基于大模型”两种评测规则。

 版权所有 © 华为云计算技术有限公司

- 基于规则（相似度/准确率）自动对模型生成的回答进行评测。用户可使用评测模板中预置的专业数据集进行评测，或者自定义评测数据集进行评测。支持的预置数据集如表12-5。
- 基于大模型，使用大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景，包含评分模式与对比模式。

表 12-5 预置评测集

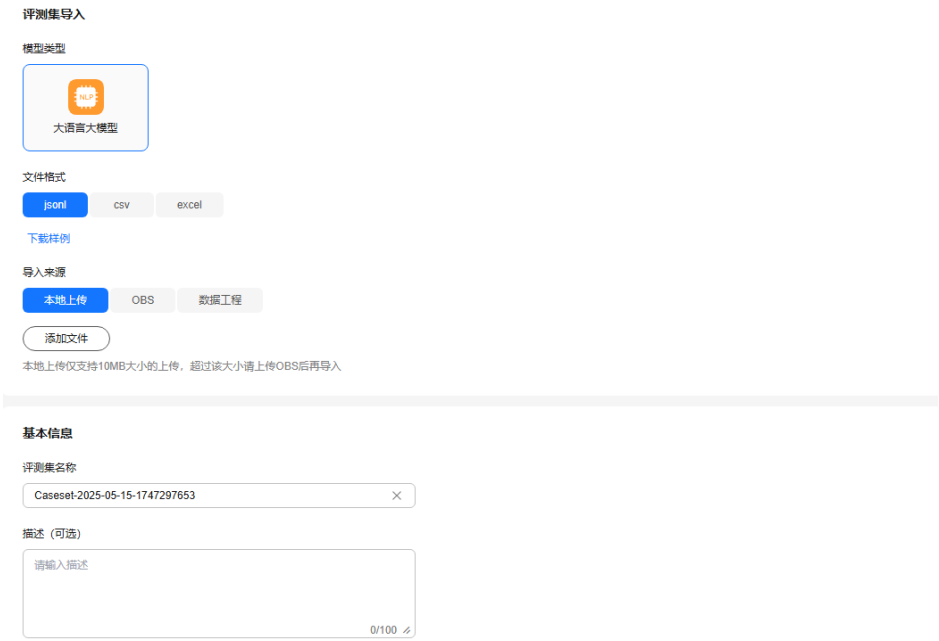
数据集类别	数据集	说明
通用知识与技能	常识知识	评测模型对于日常生活基本知识和信息的掌握程度，包括历史、地理、文化等多个方面的基础认知。
	数学能力	评测模型解决数学问题的能力，包括算数运算、代数方程求解、几何图形分析能力等。
	逻辑推理	评测模型根据已知信息进行合理推断和分析的能力，涉及演绎，归纳等多种逻辑思维过程。
	中文能力	评测大模型在中文语言和文化背景下的高级知识和推理能力。
领域知识	金融领域	评测大模型在金融领域的能力，包括快速理解和解释复杂的金融概念，风险预测和数据分析，投资建议以及金融决策支持等。

评测集支持从本地上传、OBS 文件和数据工程数据集导入

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“评测中心 > 评测集管理”，单击界面右上角“创建评测集”。
3. 在“创建评测集”页面选择所需要的“文件内容”、“文件格式”、“导入来源”，并单击“选择路径”上传数据文件。

版权所有 © 华为云计算技术有限公司

图 12-1 评测集导入



模型类型	评测数据集格式	大小
Deepseek大模型	json/csv/xlsx	本地上传最大10M，OBS最大100M

数据工程数据集创建

评测数据集的创建步骤与训练数据集一致，本章节仅做简单介绍，详细步骤请参见**使用数据工程构建Deepseek大模型数据集**。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面选择所需要的“文件内容”、“文件格式”、“导入来源”，并单击“选择路径”上传数据文件。

   Deepseek大模型评测数据集支持的格式见**表12-6**。

   表 12-6 评测数据集格式

模型类型	评测数据集格式
Deepseek大模型	文本-单轮问答-jsonl格式

4. 上传数据文件后，填写“数据集名称”与“描述”，单击“立即创建”。
5. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
6. 在“创建发布任务”页面选择数据集模态并选择数据集文件。
7. 单击“下一步”，选择发布格式，填写名称，选择数据集可见性，单击“确定”。

版权所有 © 华为云计算技术有限公司

说明

如果评测盘古大模型， 需要在发布数据集时，将数据集格式发布为“盘古格式”。

12.4.2 创建 API 服务

API管理模块可以对评测服务的API进行管理，在创建评测任务的时候选择评测服务即可。可以对API进行创建、编辑、删除。

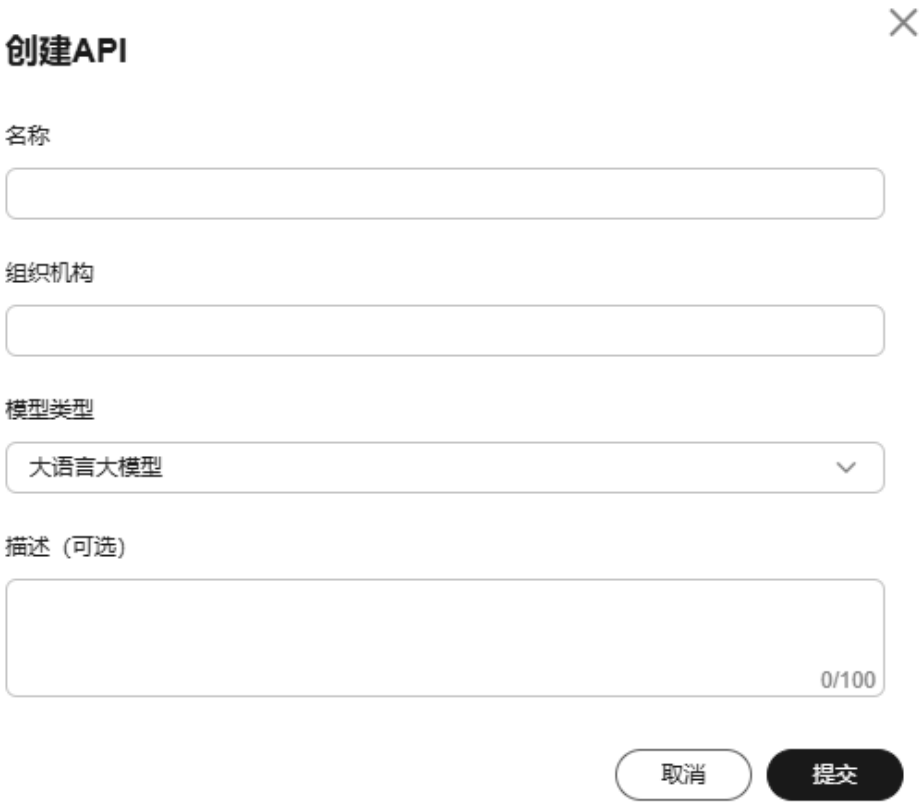
1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 12-2 我的空间



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。
3. 定义API名称、组织机构，并选择模型类型，单击“提交”。

图 12-3 创建 API



版权所有 © 华为云计算技术有限公司

4. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。
5. 单击“新增版本”，并配置模型版本、并发数、QPM、URL、Param、Body、Response。
   a. 模型版本：自定义。
   b. 并发数：可以根据模型的性能进行配置，范围1~32。
   c. QPM：可以根据模型的性能进行配置，范围1~5000。
   d. URL：模型推理地址。
   e. Param：用户可以根据接口需要配置，可选，可以定义header和请求参数信息。
   f. Body(自定义请求体)：推理请求体，可以自定义也可以选择预置的openai或者tgi结构。
   g. Response(响应体)：通过jsonpath的形式定义怎么获取响应体里面的内容。jsonpath语法的作用是从响应体的json字段中提取出所需的数据。

 版权所有 © 华为云计算技术有限公司

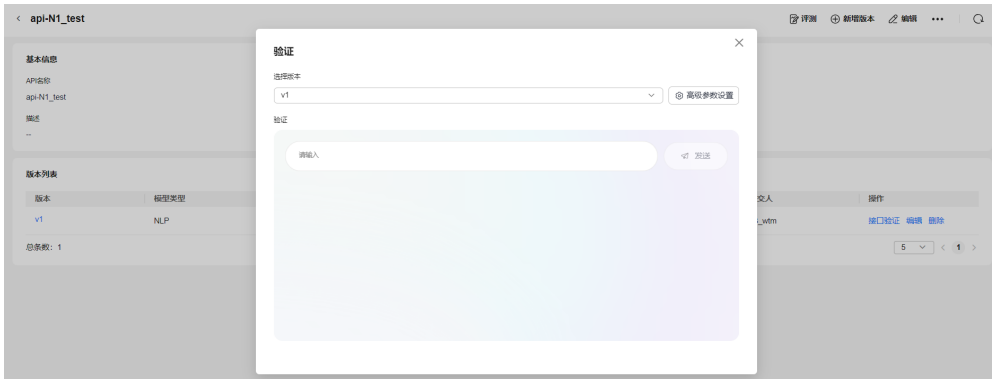
图 12-4 新增版本



 版权所有 © 华为云计算技术有限公司

6. 创建API版本后，需要验证API是否可用，只有验证通过的API才可以进行评测。在“API管理”页面单击创建好的API任务名称，单击“接口验证”输入问题后查看接口是否有响应内容输出。

图 12-5 接口验证

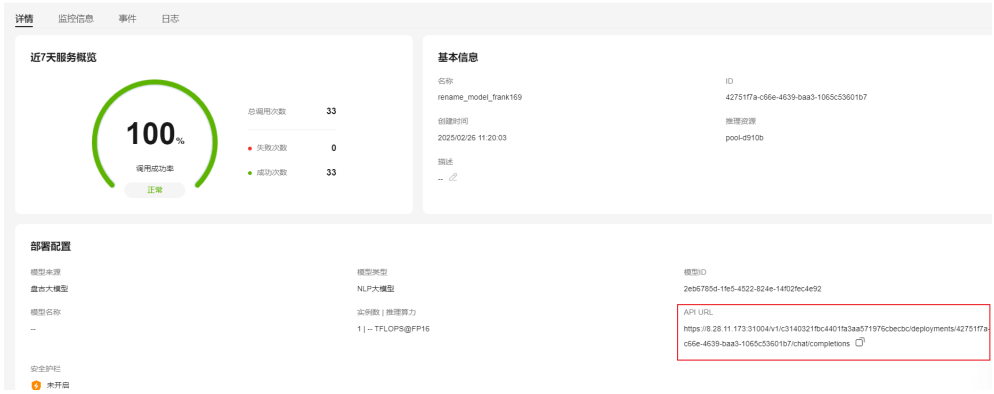


创建 API 服务示例

本示例中使用已经部署，且正常运行的盘古NLP模型创建API服务。

1. 在“模型开发 > 模型部署”页面，找到已部署的NLP模型，单击服务名称，在详情页复制API URL。

图 12-6 获取 API URL



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。定义API名称、组织结构，模型类型选择NLP大模型。
3. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。单击“新增版本”，参考表12-7完成配置。

表 12-7 配置 API 服务

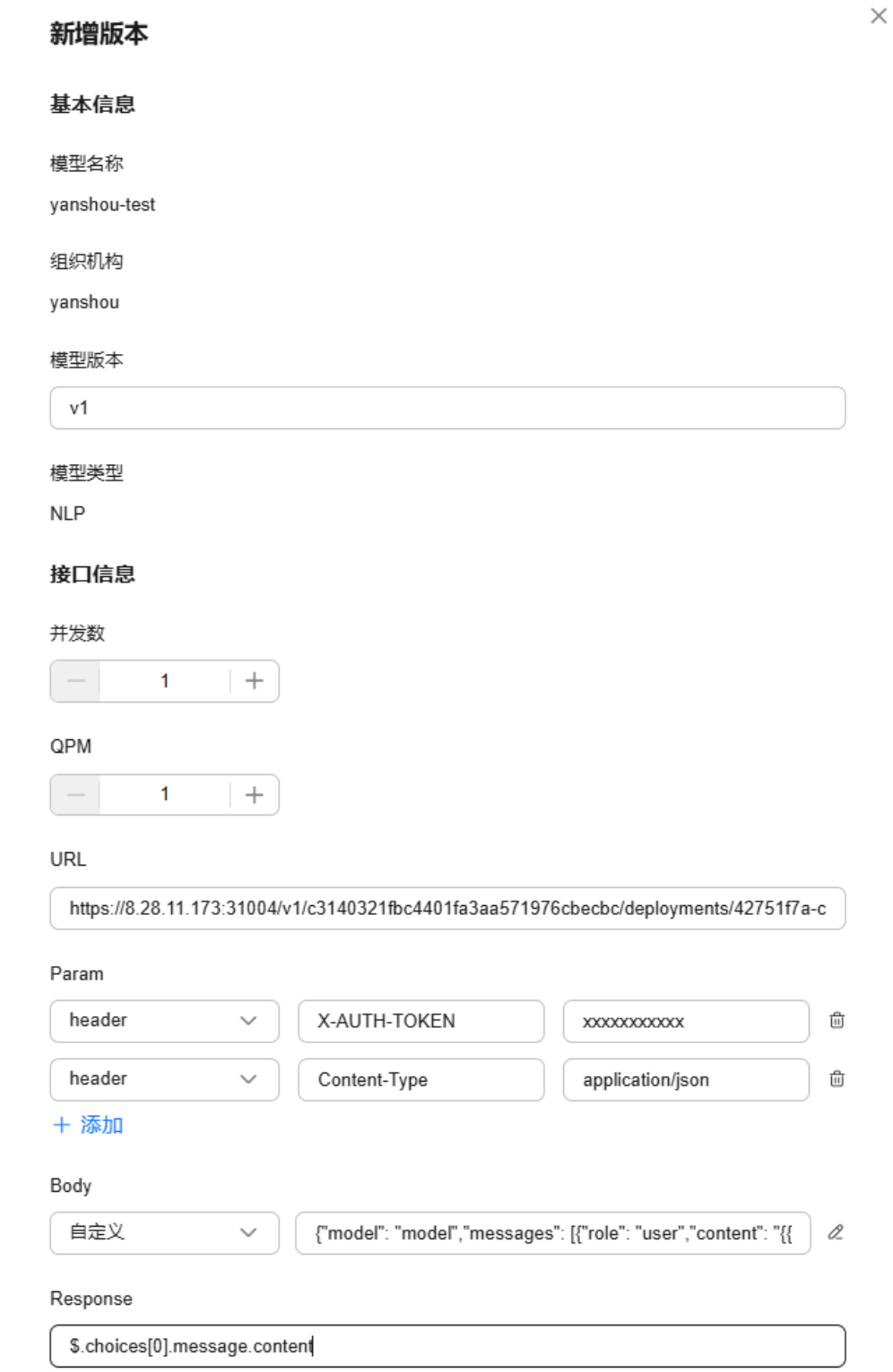
参数	填写说明
模型版本	可自定义。
并发数	设置为1。
QPM	设置为1。

版权所有 © 华为云计算技术有限公司

参数	填写说明
URL	填写获取到的API URL。
Param	共填写两个参数，均设置为header类型。 参数1：参数名X-AUTH-TOKEN，参数值为Token值，参考《 API参考 》文档“如何调用REST API > 认证鉴权”章节获取Token。 参数2：参数名Content-Type，参数值application/json。
Body	选择自定义，填写内容： {"model": "model","messages": [{"role": "user","content": "{{message}}"}],"max_tokens": 1024, "temperature": 0.7,"top_p": 0.95,"presence_penalty": 0.5,"frequency_penalty": 0.5,"stream": false}
Response	填写： $.choices[0].message.content

版权所有 © 华为云计算技术有限公司

图 12-7 配置 API 服务



12.4.3 创建 Deepseek 大模型评测任务

创建DeepSeek大模型评测任务前，请确保已完成**创建Deepseek大模型评测数据集**操作。

版权所有 © 华为云计算技术有限公司

创建 Deepseek 大模型自动评测任务（基于规则）

创建DeepSeek大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 12-8 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考表12-8完成部署参数设置。

表 12-8 Deepseek 大模型自动评测任务参数说明（基于规则）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测规则	选择“基于规则”：基于规则自动打分，即基于相似度/准确率进行打分，对比模型预测结果与标注数据的差异，适合标准选择题或简单问答场景。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。 • 单个评测集：由用户指定评测指标（F1分数、准确率、BLEU、Rouge）并上传评测数据集进行评测。选择“单个评测集”时需要上传待评测数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

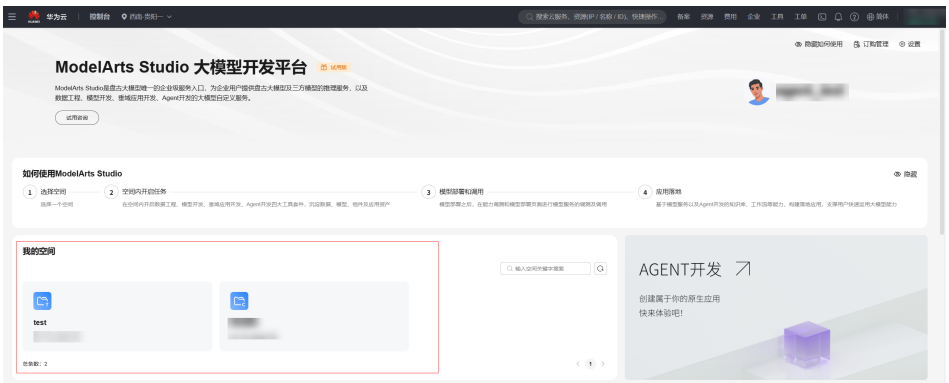
4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”查看模型评测结果，包括模型详细的得分以及评测明细。

创建 Deepseek 大模型自动评测任务（基于大模型）

创建Deepseek大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 12-9 我的空间



2. 在左侧导航栏中选择“评测中心> 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考表12-9完成部署参数设置。

版权所有 © 华为云计算技术有限公司

表 12-9 Deepseek 大模型自动评测任务参数说明（基于大模型）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测规则	选择“基于大模型”：使用能力更强的大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景。
	选择模式	● 评分模式：裁判模型将根据设置的评分标准对模型推理结果自动进行打分 ● 对比模式：裁判模型将对比两个模型在每个评测题目上的表现，选择win、lose、tie展示对比结果 ，对比模式下服务来源必须选择2个服务，默认所选择的第一个服务作为基准模型。
	评测数据集	选择模型评测所需的数据集。
	评测结果存储位置	模型评测结果的存储位置。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
裁判员配置	裁判模型	• 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 • 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
	打分规则	打分prompt模板可以选择预置，也可以选择自定义。创建自定义prompt模板需要单击“添加自定义规则->新建”，根据页面输入名称，人设，任务描述，是否包含问题，是否包含参考答案，打分策略，评测指标以及指标，最终单击“保存模板”。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”查看模型评测结果，包括模型详细的得分以及评测明细。

创建 DeepSeek 大模型人工评测任务

创建DeepSeek大模型人工评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 12-10 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“评测中心> 评测任务”，单击界面右上角“创建人工评测任务”。
3. 在“创建人工评测任务”页面，参考**表12-10**完成部署参数设置。

表 12-10 DeepSeek 大模型人工评测任务参数说明

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言大模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测指标	由用户自定义评测指标并填写评测标准。
	评测数据集	待评测的数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 人工评测”页面。
5. 当状态为“待评测”时，可以单击操作列“在线评测”进入评测页面。
6. 依据页面提示对评估效果区域进行评测打分，全部数据评测完成后单击“提交”。
   – 评测详情页面，单击“盲测”，会隐藏模型名称，进行盲评。
   – 单击“存疑”或者“作废”进行用例的存疑或作废，若取消存疑或作废，单击“取消存疑”或“取消作废”进行处理。

版权所有 © 华为云计算技术有限公司

- 单击备注下方“单击以添加备注”，可以进行新增备注。
- 评测页面，长按鼠标左键选中需要标记的文本内容，单击“标记”可以标记成重点内容。

图 12-11 人工评测



7. 返回“评测中心 > 评测任务 > 人工评测”页面，单击操作列“评测报告”查看模型评测结果。

   评测完成之后，进入人工评测列表页面，单击“人工复核”，进行复核评测，复核完成之后，单击“提交”，提交评估结果。

12.4.4 查看 Deepseek 大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 12-12 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型评测 > 任务管理”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**Deepseek大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 评测明细”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

版权所有 © 华为云计算技术有限公司

Deepseek 大模型评测指标说明

Deepseek大模型支持自动评测与人工评测，各指标说明如表12-11、表12-12、表12-13。

表 12-11 Deepseek 大模型自动评测指标说明-不使用预置评测集

评测指标（自动评测-不使用评测模板）	指标说明
F1_SCORE	精准率和召回率的调和平均数，数值越高，表明模型性能越好。
BLEU-1	模型生成句子与实际句子在单字层面的匹配度，数值越高，表明模型性能越好。
BLEU-2	模型生成句子与实际句子在词组层面的匹配度，数值越高，表明模型性能越好。
BLEU-4	模型生成结果和实际句子的加权平均精确率，数值越高，表明模型性能越好。
ROUGE-1	模型生成句子与实际句子在单个词的相似度，数值越高，表明模型性能越好。
ROUGE-2	模型生成句子与实际句子在两个词的相似度，数值越高，表明模型性能越好。
ROUGE-L	模型生成句子与实际句子在最长公共子序列的相似度，数值越高，表明模型性能越好。
PRECISION	问答匹配的精确度，模型生成句子与实际句子相比的精确程度，数值越高，表明模型性能越好。

表 12-12 Deepseek 大模型自动评测指标说明-使用预置评测集

评测指标（自动评测-使用评测模板）	指标说明
评测得分	每个数据集上的得分为模型在当前数据集上的通过率；评测能力项中若有多个数据集则按照数据量的大小计算通过率的加权平均数。
综合能力	综合能力是计算所有数据集通过率的加权平均数。

表 12-13 Deepseek 大模型人工评测指标说明

评测指标（人工评测）	指标说明
准确性	模型生成答案正确且无事实性错误。

版权所有 © 华为云计算技术有限公司

评测指标（人工评测）	指标说明
average	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
goodcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
badcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分以下的占比。
用户自定义的指标	由用户定义的指标，如有用性、逻辑性、安全性等。

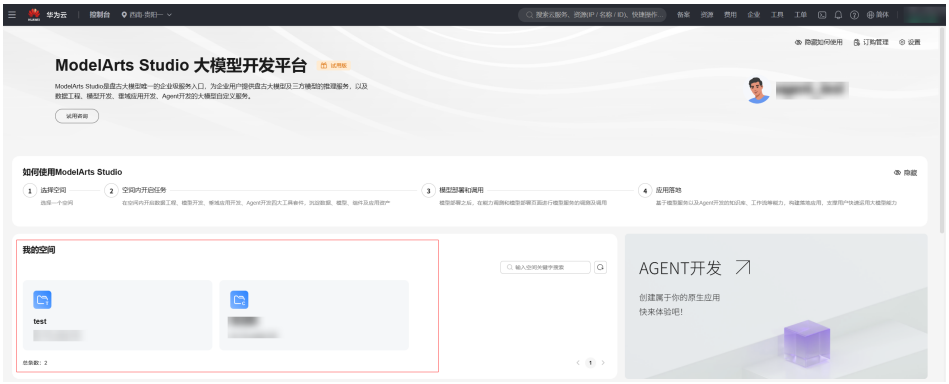
12.4.5 管理 Deepseek 大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 12-13 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

     说明

     删除属于高危操作，删除前请确保当前任务不再需要。

版权所有 © 华为云计算技术有限公司

12.5 调用 Deepseek 大模型

获取三方模型 API 请求地址

API请求地址由API接口域名+API访问路径（URI）拼接形成。

模型部署成功后，可以通过如下接口调用三方模型。

表 12-14 Deepseek 模型接口（以文本补全为例）

API分类	API访问路径（URI）
Deepseek-xxx	{api-url}/generate

API接口域名可以在“模型开发 > 模型部署”页面获取，获取方式如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称在“详情”页面获取API接口域名。

   只有处于“运行中”状态的模型才可以正常调用。

   图 12-14 获取 Deepseek 模型 API 接口域名



3. 获取API接口域名后，结合API访问路径拼接成完整的API接口。

获取 Token

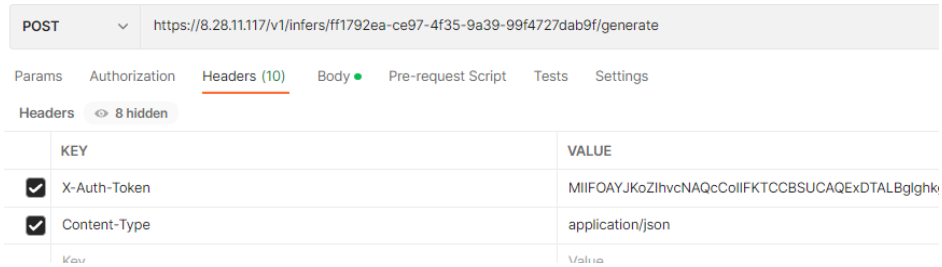
参考《API参考》文档“如何调用REST > API认证鉴权”章节获取Token。

使用 API 调用 Deepseek 大模型

1. 在Postman中新建POST请求，并填入API请求地址。
2. 参考图12-15填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**获取Token**中获取的Token值。

版权所有 © 华为云计算技术有限公司

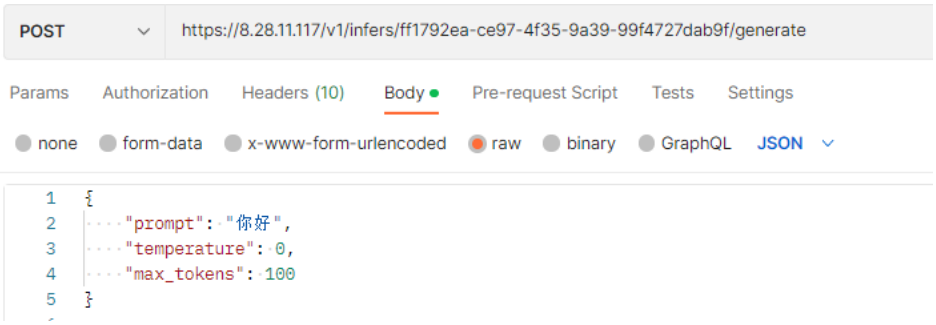
图 12-15 填写三方模型 API



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "prompt": "你好",
   "temperature": 0,
   "max_tokens": 100
}
```

图 12-16 填写三方模型请求 Body



4. 单击Postman界面“Send”按钮，发送请求。

版权所有 © 华为云计算技术有限公司

13 开发三方大模型

使用数据工程构建三方大模型数据集
部署三方大模型
评测三方大模型
调用三方大模型
三方大模型训推高可用性

13.1 使用数据工程构建三方大模型数据集

构建三方大模型数据集流程

在ModelArts Studio平台中，使用数据工程构建三方大模型数据集流程见**表13-1**。

表 13-1 三方大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务或本地的数据导入至平台统一管理，用于后续加工或发布操作。 **说明：**执行导入操作时，数据集类型选择“其他”。	**导入数据至盘古平台**
加工其他类数据集	加工其他类数据集	通过自定义加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。	**加工其他类数据集**
发布其他类数据集	发布其他类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。	**发布其他类数据集**

版权所有 © 华为云计算技术有限公司

构建三方大模型数据集步骤

在ModelArts Studio大模型开发平台中，使用数据工程构建三方大模型数据集流程可参考**表13-1**。

创建导入任务

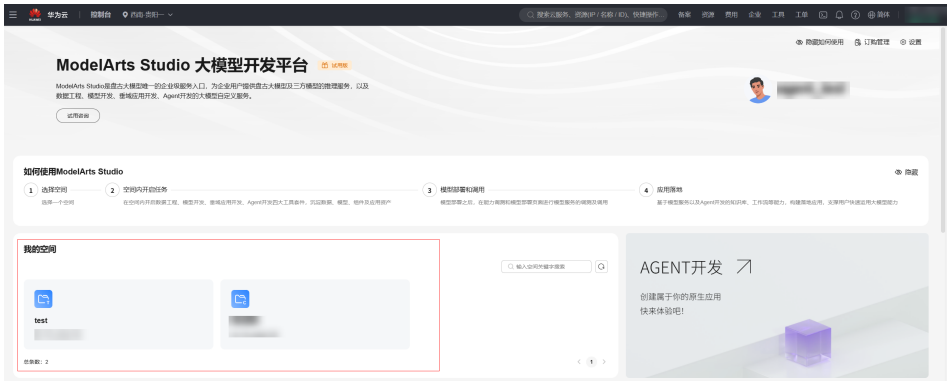
创建导入任务前，请先按照上文要求提前准备数据。

平台支持使用OBS服务导入数据，请详见**通过控制台快速使用OBS**。

创建导入任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击空间名称进入操作空间。

   图 13-1 我的空间

2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 选择“其他”类型的数据集，并选择训练数据所在的OBS存储位置。
4. 单击“立即创建”创建数据集。
5. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
6. 选择“其他”类型的数据集，并勾选创建的数据集。单击“下一步”。设置数据用途、数据集可见性，填写数据集名称、描述，设置扩展信息后（可选），单击“确定”进行数据集发布操作。

13.2 部署三方大模型

13.2.1 模型部署形态和部署说明

模型名称	云上部署	边缘部署
DeepSeek-R1-distill-Qwen-32B	支持，2推理单元起部署	支持，2推理单元起部署
DeepSeek-R1-distill-LLama-70B	支持，4推理单元起部署	支持，4推理单元起部署

版权所有 © 华为云计算技术有限公司

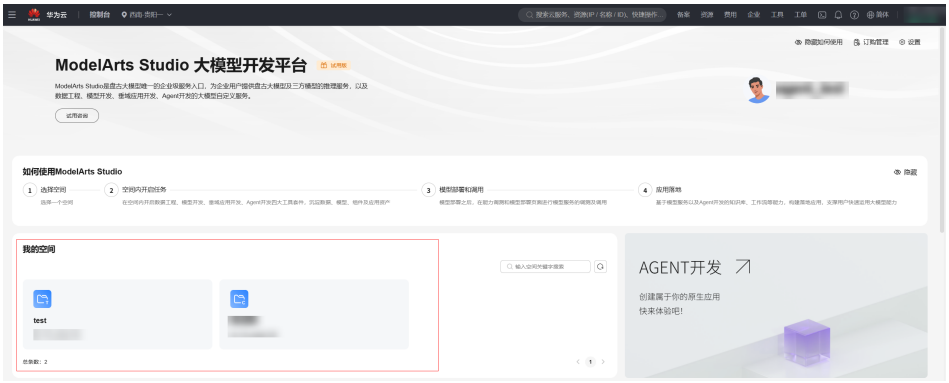
模型名称	云上部署	边缘部署
DeepSeek-R1-distill-LLama-8B	支持，1推理单元起部署	支持，1推理单元起部署
Qwen3-235B-A22B	支持，16推理单元起部署	支持，16推理单元起部署
Qwen3-32B	支持，2推理单元起部署	支持，2推理单元起部署
Qwen3-30B-A3B	支持，2推理单元起部署	支持，2推理单元起部署
Qwen3-14B	支持，1推理单元起部署	支持，1推理单元起部署
Qwen3-8B	支持，1推理单元起部署	支持，1推理单元起部署
Qwen2.5-72B	支持，4推理单元起部署	支持，4推理单元起部署
QWQ-32B	支持，4推理单元起部署	支持，4推理单元起部署

13.2.2 创建三方大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 13-2 我的空间

2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“模型广场 > 大语言模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表13-2**完成部署参数设置。

版权所有 © 华为云计算技术有限公司

表 13-2 三方大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“大语言模型”，并选择需要进行部署的模型和版本。
	部署方式	支持“云上部署”。云上部署指算法部署至平台提供的资源池中。
	自定义名称	此名称是通过V2版本推理接口调用该推理服务时的唯一标识。创建后不支持修改。
安全护栏	安全护栏开关	安全护栏保障模型调用安全。
	版本选择	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

5. 参数填写完成后，单击“立即部署”。

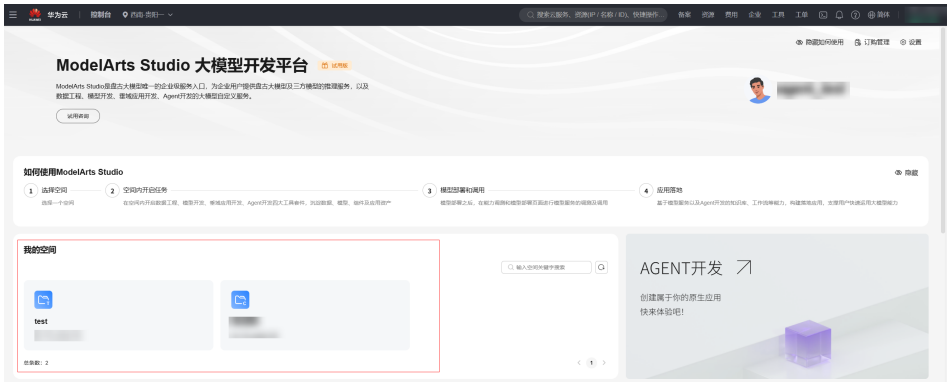
13.2.3 查看三方大模型部署任务详情

部署任务创建成功后，可以查看大模型部署任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 13-3 我的空间

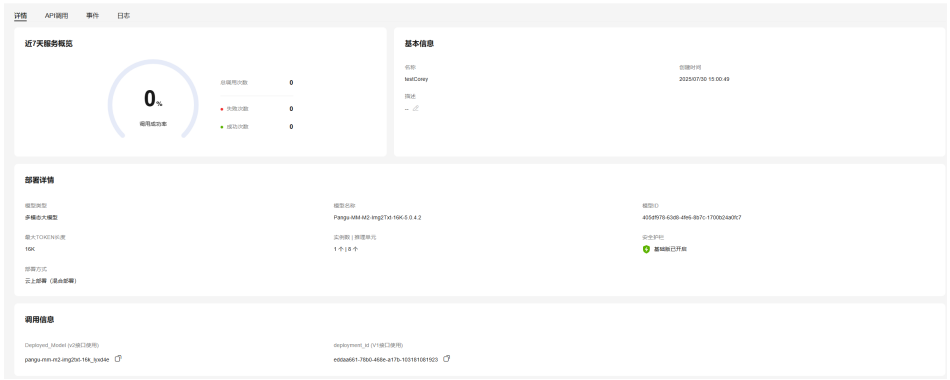


2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建三方大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

图 13-4 部署详情



13.2.4 管理三方大模型部署任务

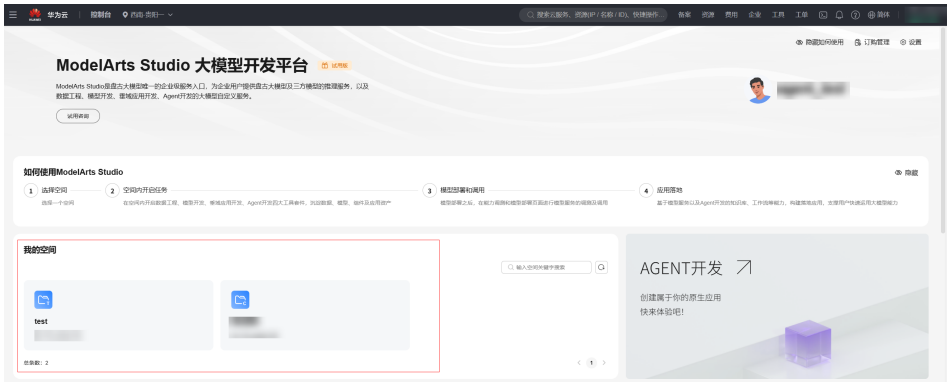
模型更新

完成**创建三方大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

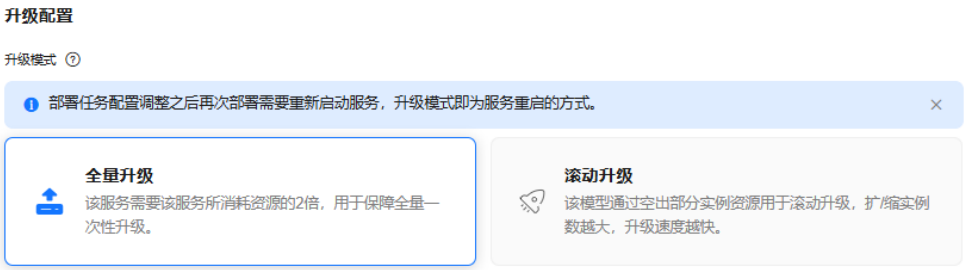
版权所有 © 华为云计算技术有限公司

图 13-5 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 选择模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 13-6 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

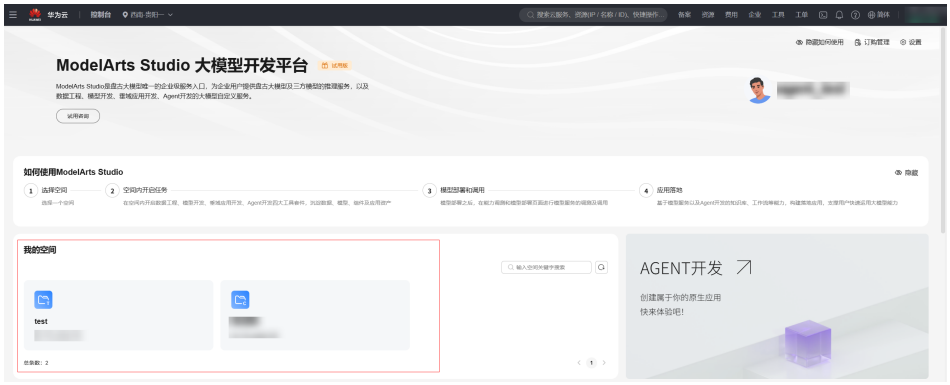
修改部署配置

完成**创建三方大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

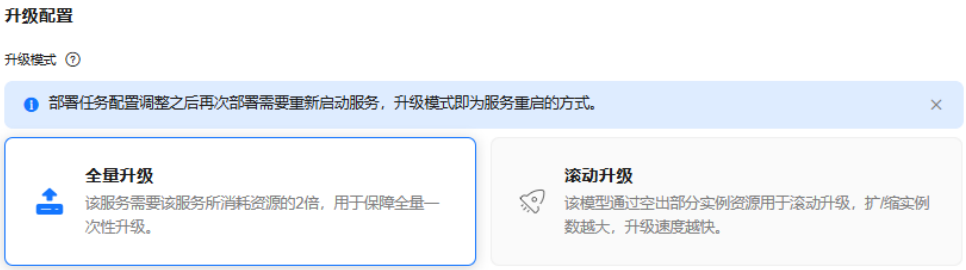
版权所有 © 华为云计算技术有限公司

图 13-7 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 13-8 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

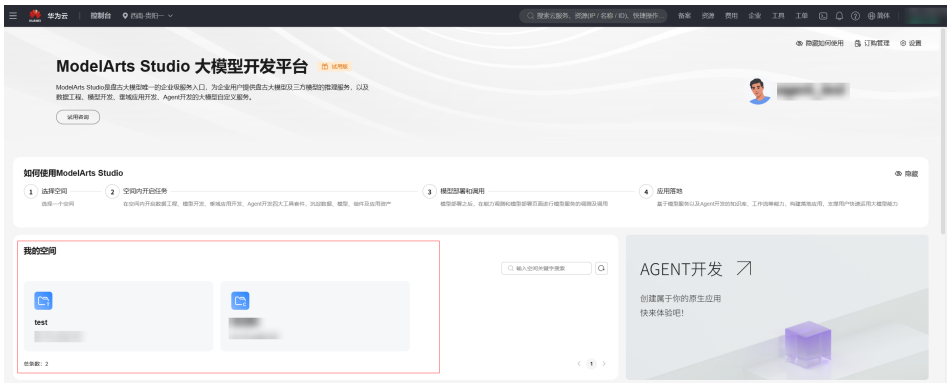
模型部署实例扩缩容

完成**创建三方大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 13-9 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-10 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

版权所有 © 华为云计算技术有限公司

13.3 评测三方大模型

13.3.1 创建三方大模型评测数据集

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同情景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

三方大模型支持人工评测、自动评测两种评测模式。

- 人工评测：通过人工创建的评测数据集和评测指标项对模型生成的回答进行评测，评测时需要人工基于创建好的评测项对模型回答进行打分，评测完成后会基于打分结果生成评测报告。
- 自动评测：包含“基于规则”与“基于大模型”两种评测规则。
  - 基于规则（相似度/准确率）自动对模型生成的回答进行评测。用户可使用评测模板中预置的专业数据集进行评测，或者自定义评测数据集进行评测。支持的预置数据集如**表13-3**。
  - 基于大模型，使用大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景，包含评分模式与对比模式。

表 13-3 预置评测集

数据集类别	数据集	说明
通用知识与技能	常识知识	评测模型对于日常生活基本知识和信息的掌握程度，包括历史、地理、文化等多个方面的基础认知。
	数学能力	评测模型解决数学问题的能力，包括算数运算、代数方程求解、几何图形分析能力等。
	逻辑推理	评测模型根据已知信息进行合理推断和分析的能力，涉及演绎，归纳等多种逻辑思维过程。
	中文能力	评测大模型在中文语言和文化背景下的高级知识和推理能力。
领域知识	金融领域	评测大模型在金融领域的能力，包括快速理解和解释复杂的金融概念，风险预测和数据分析，投资建议以及金融决策支持等。

评测集支持从本地上传、OBS 文件和数据工程数据集导入

1. 登录ModelArts Studio平台，进入所需空间。

版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“评测中心 > 评测集管理”，单击界面右上角“创建评测集”。
3. 在“创建评测集”页面选择所需要的“模型类型”、“文件格式”、“导入来源”，并单击“选择路径”上传数据文件。参见表13-4

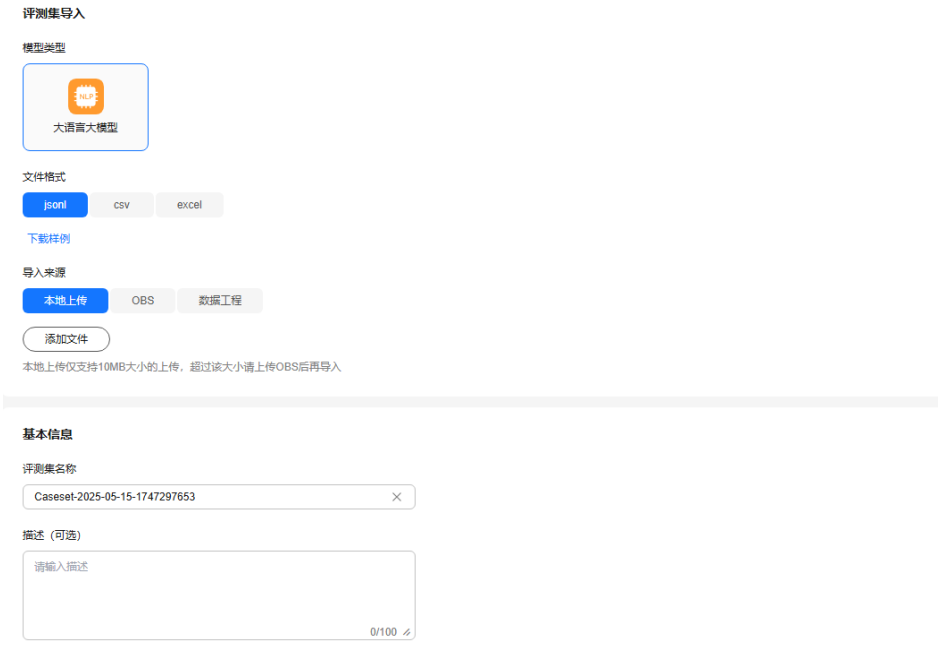


表 13-4 评测数据集参数说明

模型类型	评测数据集格式	大小
NLP大模型	json/csv/xlsx	本地上传最大10M，OBS最大100M

数据工程数据集创建

评测数据集的创建步骤与训练数据集一致，本章节仅做简单介绍，详细步骤请参见使用数据工程构建三方大模型数据集。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面选择所需要的“数据集类型”、“文件格式”、“导入来源”，并单击“存储位置”上传数据文件。

   NLP大模型评测数据集支持的格式见表13-5。

表 13-5 评测数据集格式

模型类型	评测数据集格式
三方大模型	文本-单轮问答-jsonl格式

4. 上传数据文件后，填写“数据集名称”与“描述”，单击“立即创建”。

版权所有 © 华为云计算技术有限公司

5. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
6. 在“创建发布任务”页面选择数据集模态并选择数据集文件。
7. 单击“下一步”，选择发布格式，填写名称，选择数据集可见性，单击“确定”。

说明

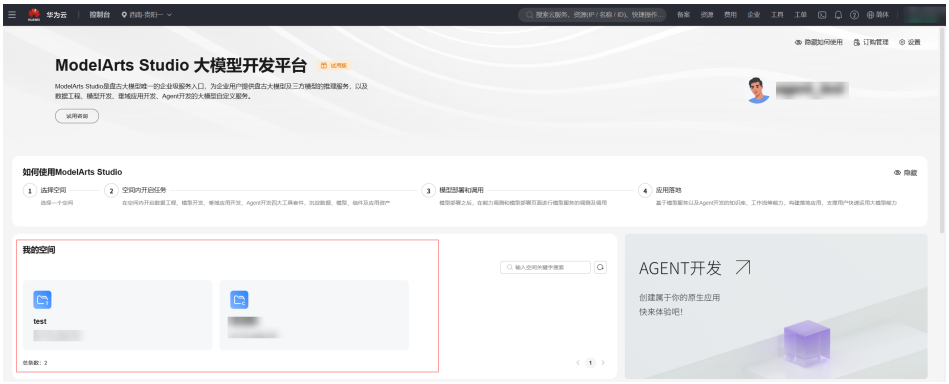
如果评测盘古大模型， 需要在发布数据集时，将数据集格式发布为“盘古格式”。

13.3.2 创建 API 服务

API管理模块可以对评测服务的API进行管理，在创建评测任务的时候选择评测服务即可。可以对API进行创建、编辑、删除。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 13-11 我的空间



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。
3. 定义API名称、组织机构，并选择模型类型，单击“提交”。

图 13-12 创建 API



版权所有 © 华为云计算技术有限公司

4. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。
5. 单击“新增版本”，并配置模型版本、并发数、QPM、URL、Param、Body、Response。
   a. 模型版本：自定义。
   b. 并发数：可以根据模型的性能进行配置，范围1~32。
   c. QPM：可以根据模型的性能进行配置，范围1~5000。
   d. URL：模型推理地址。
   e. Param：用户可以根据接口需要配置，可选，可以定义header和请求参数信息。
   f. Body(自定义请求体)：推理请求体，可以自定义也可以选择预置的openai或者tgi结构。
   g. Response(响应体)：通过jsonpath的形式定义怎么获取响应体里面的内容。jsonpath语法的作用是从响应体的json字段中提取出所需的数据。

 版权所有 © 华为云计算技术有限公司

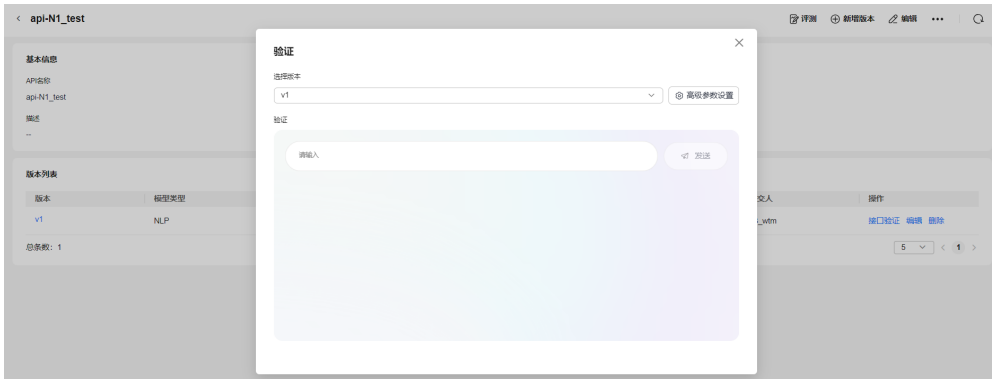
图 13-13 新增版本



 版权所有 © 华为云计算技术有限公司

6. 创建API版本后，需要验证API是否可用，只有验证通过的API才可以进行评测。在“API管理”页面单击创建好的API任务名称，单击“接口验证”输入问题后查看接口是否有响应内容输出。

图 13-14 接口验证

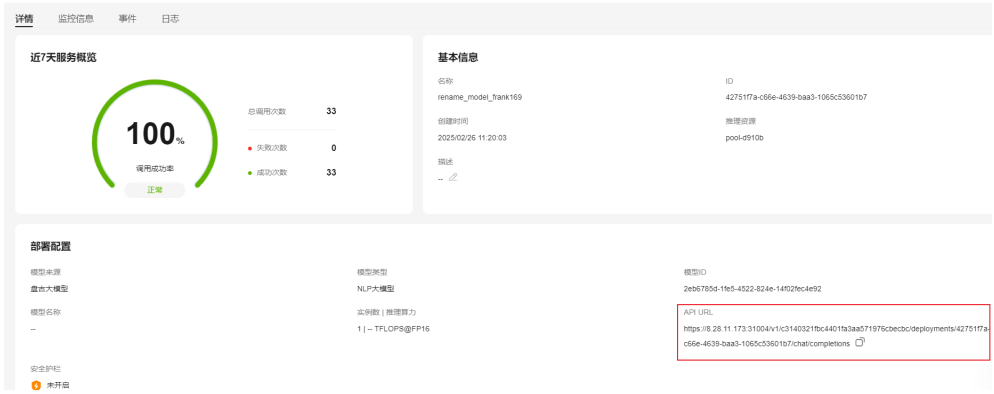


创建 API 服务示例

本示例中使用已经部署，且正常运行的三方大模型创建API服务。

1. 在“模型开发 > 模型部署”页面，找到已部署的三方大模型，单击服务名称，在详情页复制API URL。

图 13-15 获取 API URL



2. 在左侧导航栏中选择“评测中心 > API管理”，单击界面右上角“创建API”。定义API名称、组织结构，模型类型选择三方大模型。
3. API创建好以后，通过右侧新增版本来定义API的信息，可以新建多个版本。单击“新增版本”，参考表13-6完成配置。

表 13-6 配置 API 服务

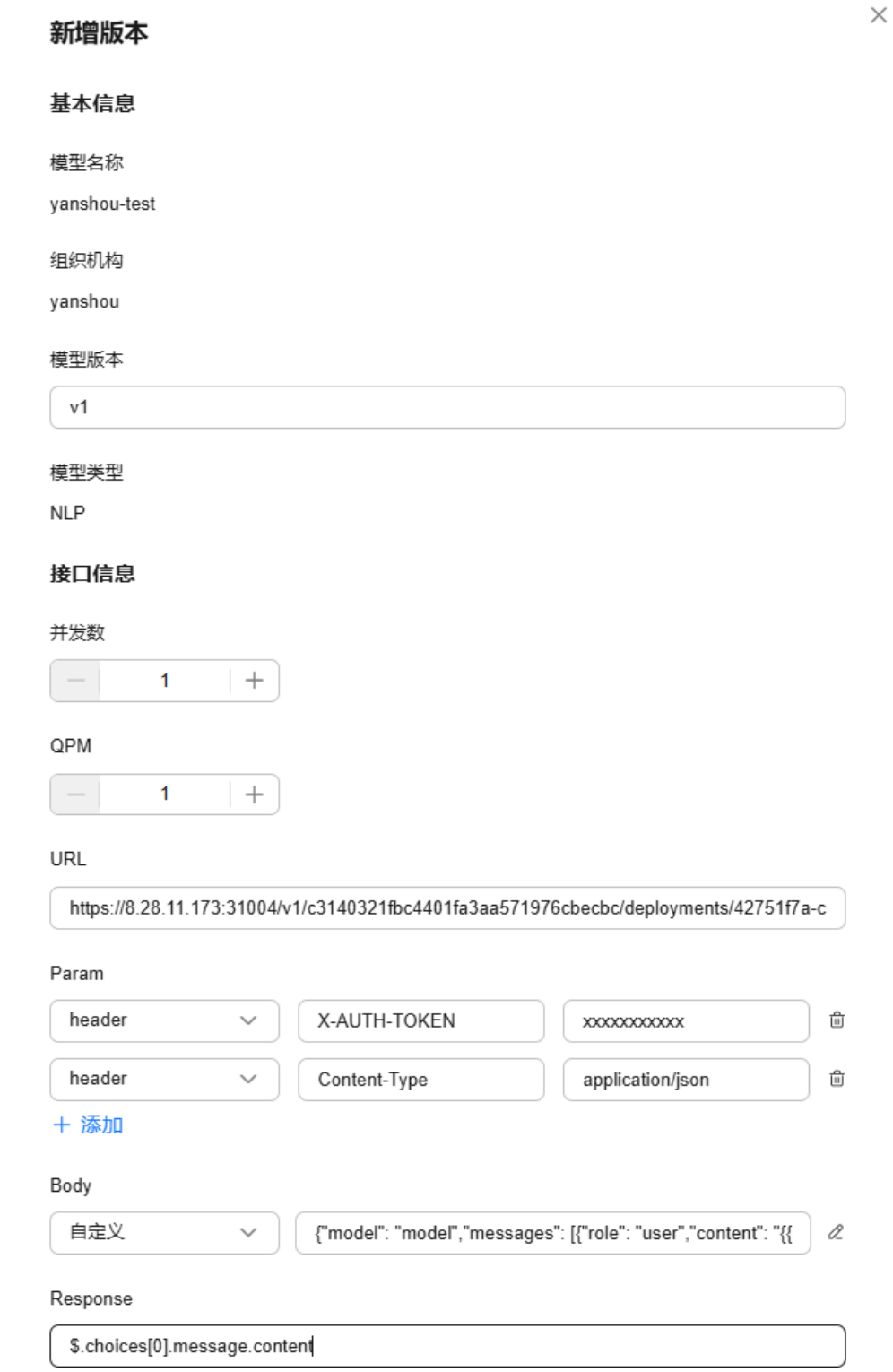
参数	填写说明
模型版本	可自定义。
并发数	设置为1。
QPM	设置为1。

版权所有 © 华为云计算技术有限公司

参数	填写说明
URL	填写获取到的API URL。
Param	共填写两个参数，均设置为header类型。 参数1：参数名X-AUTH-TOKEN，参数值为Token值，参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。 参数2：参数名Content-Type，参数值application/json。
Body	选择自定义，填写内容： {"model": "model","messages": [{"role": "user","content": "{{message}}"}],"max_tokens": 1024, "temperature": 0.7,"top_p": 0.95,"presence_penalty": 0.5,"frequency_penalty": 0.5,"stream": false}
Response	填写： $.choices[0].message.content

版权所有 © 华为云计算技术有限公司

图 13-16 配置 API 服务



13.3.3 创建三方大模型评测任务

创建三方大模型评测任务前，请确保已完成**创建三方大模型评测数据集**操作。

版权所有 © 华为云计算技术有限公司

创建三方大模型自动评测任务（基于规则）

创建三方大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-17 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考表13-7完成部署参数设置。

表 13-7 三方大模型自动评测任务参数说明（基于规则）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测规则	选择“基于规则”：基于规则自动打分，即基于相似度/准确率进行打分，对比模型预测结果与标注数据的差异，适合标准选择题或简单问答场景。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。 • 自定义评测集：由用户指定评测指标（F1分数、准确率、BLEU、Rouge）并上传评测数据集进行评测。选择“自定义评测集”时需要上传待评测数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

创建三方大模型自动评测任务（基于大模型）

创建三方大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-18 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面，参考**表13-8**完成部署参数设置。

版权所有 © 华为云计算技术有限公司

表 13-8 三方大模型自动评测任务参数说明（基于大模型）

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测规则	选择“基于大模型”：使用能力更强的大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景。
	选择模式	● 评分模式：裁判模型将根据设置的评分标准对模型推理结果自动进行打分 ● 对比模式：裁判模型将对比两个模型在每个评测题目上的表现，选择win、lose、tie展示对比结果 ，对比模式下服务来源必须选择2个服务，默认所选择的第一个服务作为基准模型。
	评测数据集	选择需要评测的数据集。NLP多轮问答场景仅支持基于大模型自动评测，可选择多轮问答评测数据集。
	评测结果存储位置	模型评测结果的存储位置。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
裁判员配置	裁判模型	• 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 • 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
	打分规则	打分prompt模板可以选择预置，也可以选择自定义。创建自定义prompt模板需要单击“添加自定义规则->新建”，根据页面输入名称，人设，任务描述，是否包含问题，是否包含参考答案，打分策略，评测指标以及指标，最终单击“保存模板”。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

创建三方大模型人工评测任务

创建三方大模型人工评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-19 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建人工评测任务”。
3. 在“创建人工评测任务”页面，参考**表13-9**完成部署参数设置。

表 13-9 三方大模型人工评测任务参数说明

参数分类	参数名称	参数说明
选择服务	评测类型	选择“大语言模型”。
	服务来源	当前仅支持通过“外部服务”调用API做评测。单次最多可评测10个模型。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openAI、tgi、自定义三种格式。 – openAI格式即是由OpenAI公司开发并标准化的一种大模型请求格式； – tgi格式即是Hugging Face团队推出的一种大模型请求格式。接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。jsonpath介绍可参考https://github.com/json-path/JsonPath。
评测配置	评测指标	由用户自定义评测指标并填写评测标准。
	评测数据集	待评测的数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“评测任务 > 人工评测”页面。
5. 当状态为“待评测”时，可以单击操作列“在线评测”进入评测页面。
6. 依据页面提示对评估效果区域进行评测打分，全部数据评测完成后单击“提交”。
   – 评测详情页面，单击“盲测”，会隐藏模型名称，进行盲评。
   – 单击“存疑”或者“作废”进行用例的存疑或作废，若取消存疑或作废，单击“取消存疑”或“取消作废”进行处理。

 版权所有 © 华为云计算技术有限公司

- 单击备注下方“单击以添加备注”，可以进行新增备注。
- 评测页面，长按鼠标左键选中需要标记的文本内容，单击“标记”可以标记成重点内容。

图 13-20 人工评测



7. 返回“评测中心 > 评测任务 > 人工评测”页面，单击操作列“评测报告”查看模型评测结果。

   评测完成之后，进入人工评测列表页面，单击“人工复核”，进行复核评测，复核完成之后，单击“提交”，提交评估结果。

13.3.4 查看三方大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-21 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**三方大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 服务结果分析”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

版权所有 © 华为云计算技术有限公司

三方大模型评测指标说明

三方大模型支持自动评测与人工评测，各指标说明如**表13-10**、**表13-11**、**表13-12**。

表 13-10 三方大模型自动评测指标说明-不使用预置评测集

评测指标（自动评测-自定义评测集）	指标说明
准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
F1分数	精确率和召回率的调和平均数，分数越高表示模型在这两个指标上表现越好，即模型在精确率和召回率之间取得了更好的平衡。
BLEU-1	模型生成句子与实际句子在单字层面的匹配度，分数越高，表示模型效果越好。
BLEU-2	模型生成句子与实际句子在中词组层面的匹配度，分数越高，表示模型效果越好。
BLEU-4	模型生成结果和实际句子的加权平均精确率，分数越高，表示模型效果越好。
ROUGE-1	将模型生成结果和标注结果按1-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。
ROUGE-2	将模型生成结果和标注结果按2-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。
ROUGE-L	将模型生成结果和标注结果按最长公共子序列（longest-gram）拆分后，计算出的召回率，分数越高，表示模型效果越好。

表 13-11 三方大模型自动评测指标说明-使用预置评测集

评测指标（自动评测-使用评测模板）	指标说明
评测得分	每个数据集上的得分为模型在当前数据集上的通过率；评测能力项中若有多个数据集则按照数据量的大小计算通过率的加权平均数。
综合能力	综合能力是计算所有数据集通过率的加权平均数。

版权所有 © 华为云计算技术有限公司

表 13-12 三方大模型人工评测指标说明

评测指标（人工评测）	指标说明
准确性	模型生成答案正确且无事实性错误。
average	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
goodcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
badcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分以下的占比。
用户自定义的指标	由用户定义的指标，如有用性、逻辑性、安全性等。

13.3.5 管理三方大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 13-22 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

     说明

     删除属于高危操作，删除前请确保当前任务不再需要。

版权所有 © 华为云计算技术有限公司

13.4 调用三方大模型

13.4.1 使用 API 调用三方大模型

预置模型或训练后的模型部署成功后，可以使用“文本对话”API实现模型调用。三方大模型推理服务支持使用盘古推理接口（V1推理接口）调用，也支持使用业界通用的OpenAi格式接口（V2推理接口）调用。V1接口、V2接口的鉴权方式不同，请求体和返回体略有差异。

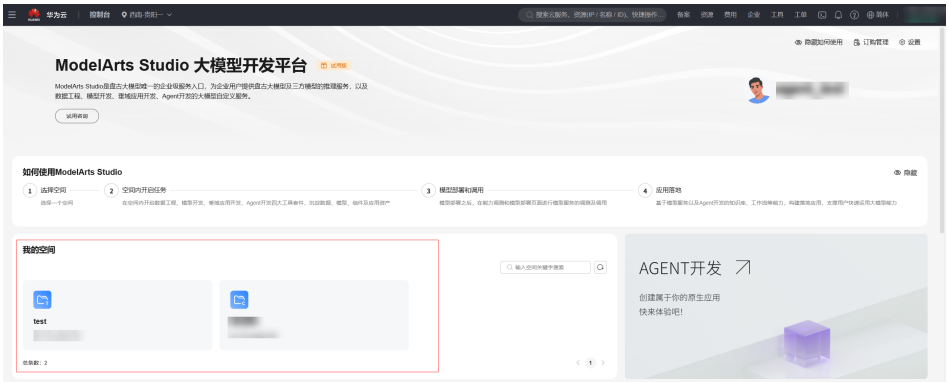
表 13-13 NLP 大模型推理 API 清单

API分类	API访问路径（URI）
V1推理接口	/v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	/api/v2/chat/completions

获取调用路径

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

   图 13-23 我的空间

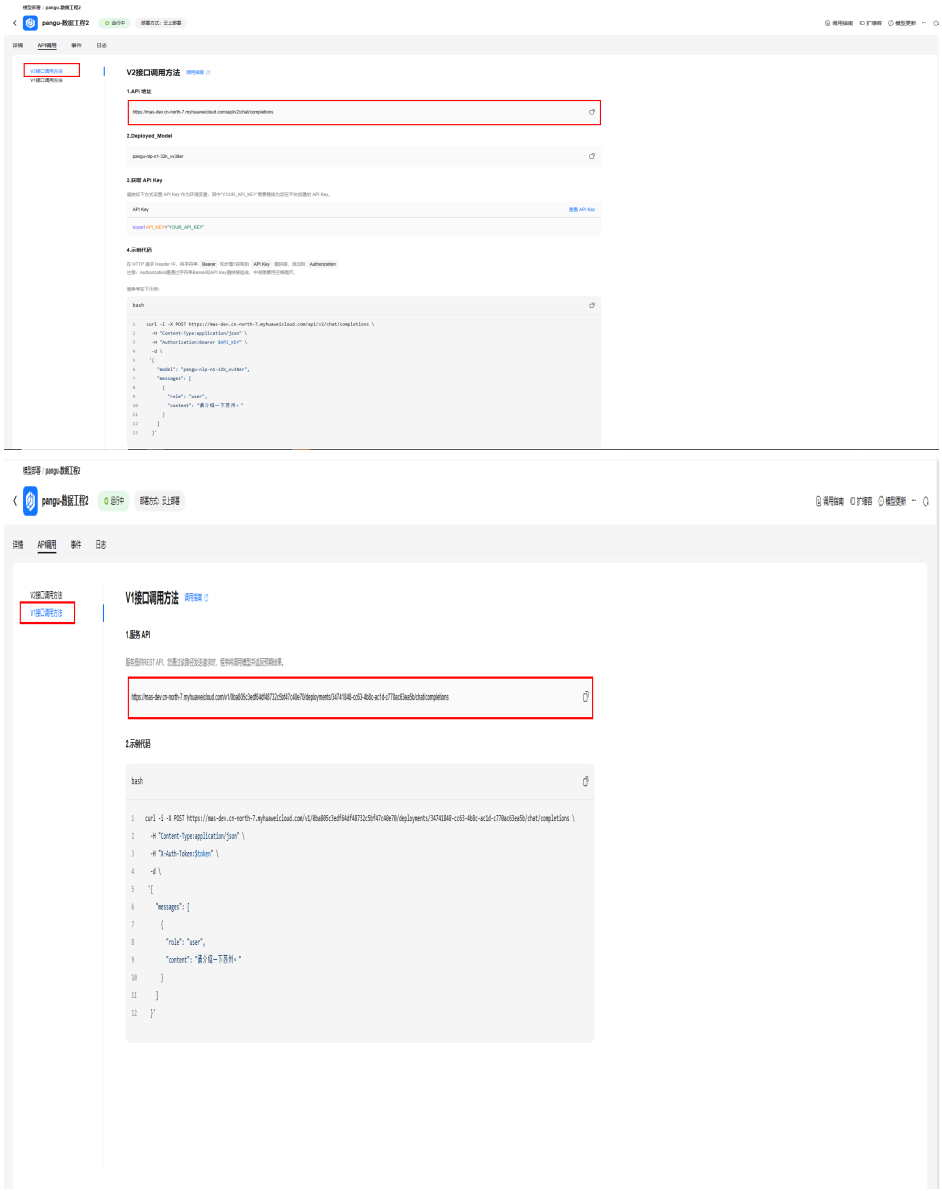
2. 获取调用路径。

   在左侧导航栏中选择“模型开发 > 模型部署”。

   - 获取已部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“API调用”页签，可获取模型调用路径，并根据页签的调用方法进行调用，如图13-24。

版权所有 © 华为云计算技术有限公司

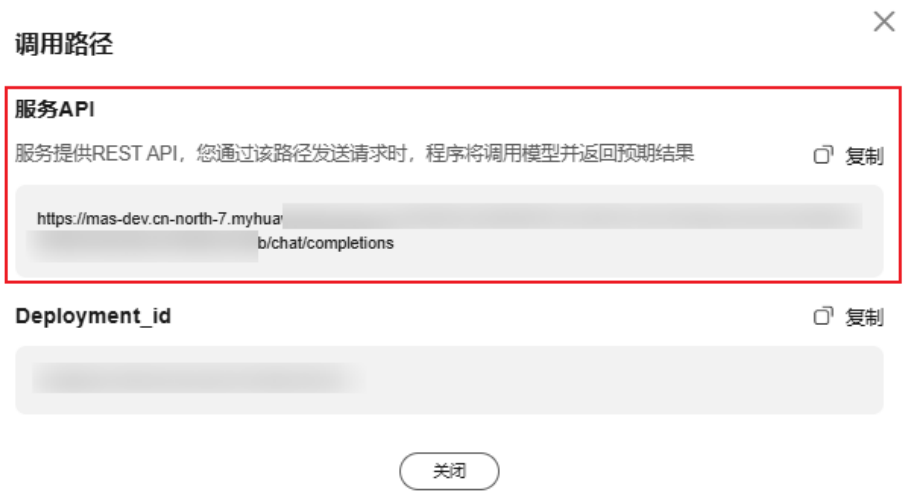
图 13-24 获取已部署模型的调用路径



- 获取预置服务的调用路径。在“预置服务”页签中，选择所需调用的NLP大模型，单击“调用路径”，在“调用路径”弹窗可获取模型调用路径，如图13-25。

版权所有 © 华为云计算技术有限公司

图 13-25 获取预置服务的调用路径



获取边缘部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“详情”页签，可获取模型调用路径。

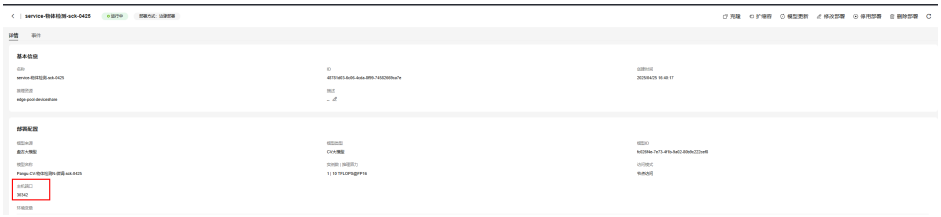
负载均衡方式：

模型路径为http://{ELB IP}:{ELB负载端口}/{API URL}/{推理的调用URL}，ELP IP需获取对应的公网IP，各个部分的获取如图所示：



节点方式：

模型路径为http://{节点所在的IP}:{主机端口}/{推理的调用URL}，节点所在IP为边缘池的工作节点的IP，各个部分的获取如图所示：



使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见**获取调用路径**。
2. 调用API有两种认证方式，包括Token认证和API Key认证。其中，API Key认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过API Key认证调用请求。

   参考**表13-14**填写请求Header参数。

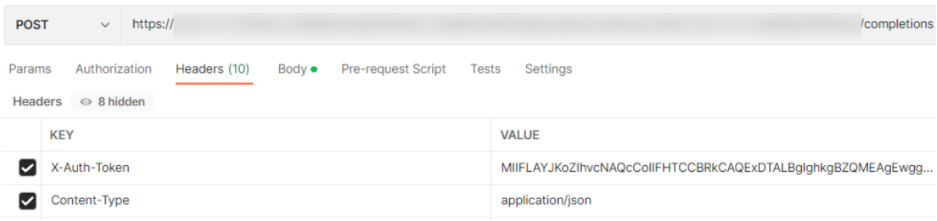
版权所有 © 华为云计算技术有限公司

表 13-14 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
V1推理接口API Key认证	Content-Type	application/json
	X-Apig-AppCode	API Key值，获取API Key步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“API Key”操作列中可获取API Key值。
V2推理接口API Key认证	Content-Type	application/json
	Authorization	用户创建应用接入获取的API Key，拼接“Bearer ”后的字符串。示例：Bearer d59******9C3

如图13-26，为Token认证方式的请求Header参数填写示例。

图 13-26 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "messages": [
     {
        "content": "介绍下长江，以及长江中典型的鱼类"
     }
    ],
   "temperature": 0.9,
   "max_tokens": 600
}
```

4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示NLP大模型API调用成功。

版权所有 © 华为云计算技术有限公司

API Key 鉴权说明

当用户部署的推理服务需要开放给其他用户调用时，原有Token认证无法支持，可用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为API Key值），API服务仅校验API Key，从而实现快速响应。

获取API Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“系统管理与统计 > 应用接入”，单击界面右上角“创建应用接入”。
3. 在“关联服务”中，选择全部服务或者指定已部署好的推理服务，并设置应用接入名称、描述，单击“确定”。
4. 在“应用接入”列表的“API Key”列中可获取API Key值。

13.5 三方大模型训推高可用性

13.5.1 训练日志失败分析

三方大模型训练常见报错及解决方案请详见表13-15。

表 13-15 三方大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空。	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。
训练日志提示“root: XXX valid number is 0”。	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。

版权所有 © 华为云计算技术有限公司

常见报错	问题现象	原因分析	解决方案
训练日志提示“ValueError: label_map not match”。	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

 版权所有 © 华为云计算技术有限公司

14 开发盘古多模态图像问答大模型

使用数据工程构建图像问答大模型数据集
训练图像问答大模型
部署图像问答大模型
评测图像问答大模型
调用图像问答大模型

14.1 使用数据工程构建图像问答大模型数据集

多模态大模型支持接入的数据集类型

盘古多模态大模型支持接入图片类数据集，不同场景所需数据见**表14-2**，数据集格式要求请参见**图片类数据集格式要求**。

表 14-1 多模态大模型

类别	模型	简介
多模态大模型	Pangu-MM-M2-Img2Txt-16K-V5.0.4.2	盘古多模态理解百亿级大模型，可以对图像、文本多模态信息进行综合理解和分析，支持16K序列，支持预训练、微调、在线推理。

版权所有 © 华为云计算技术有限公司

表 14-2 训练多模态大模型数据集类型要求

基模型	训练场景	文件类型	文件内容	文件格式	文件样例
多模态大模型	预训练	图片	图片+Caption	lmdb（jpg）+jsonl文件	图片以tar包格式存储，图片描述文本的类型为jsonl，示例如下： {"image_name":"a.jpg","tar_name": "001.tar", "caption"="一段描述"} • image_name：图片名称，图片名称需要与tar包内的图片名一致。 • tar_name：图片压缩包名称。 • caption：图片的描述文本。
	微调	图片	图片+QA对	lmdb（jpg）+jsonl文件	图片以tar包格式存储，图片描述文本的类型为jsonl，示例如下： {"tar_name":"images_0.tar","image_name":"100110.jpg","conversations":[{"question":"问题1","answer":"回答1"},{"question":"问题2","answer":"回答2"}] } • tar_name：图片压缩包名称。 • image_name：图片名称，图片名称需要与tar包内的图片名一致。 • conversations：图片的描述文本，数组结构。内容为问答对，即基于图片内容进行提问并给出回答，问答对可以是多个。

训练多模态大模型所需数据量

使用数据工程构建盘古多模态大模型数据集进行模型训练时，所需数据量见表14-3。

表 14-3 多模态数据大小说明

训练类型	最小数据量（图文对数据）	推荐数据量（图文对数据）
预训练	1000万对	3000万对
微调	20~30万对	100万对

 版权所有 © 华为云计算技术有限公司

多模态训练数据为tar+jsonl构成训练文件，单个tar包大小不超过500MB，文件总大小不超过100GB。

评测多模态大模型所需数据量

评测多模态大模型需要使用图片+QA对格式的数据，建议评测数据量在1000条以下。

构建多模态大模型数据集流程

在ModelArts Studio平台中，使用数据工程构建盘古多模态大模型数据集流程见**使用数据工程构建图像问答大模型数据集**。

表 14-4 盘古多模态大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务或本地的数据导入至平台统一管理，用于后续加工或发布操作。	**导入数据至盘古平台**
加工图片类数据集	加工图片类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	**加工图片类数据集**
	标注图片类数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。	**标注图片类数据集**
	配比图片数据集	数据配比是将多个数据集按特定比例组合并发布为“发布数据集”的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。	**配比图片类数据集**
发布图片类数据集	评估图片类数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。	**评估图片类数据集**

版权所有 © 华为云计算技术有限公司

流程	子流程	说明	操作指导
	发布图片类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**默认格式、盘古格式**。 • 默认格式：平台默认的格式。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。	发布图片类数据集

数据 prompt 拼接样例

默认prompt为：

```
[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10]
```

输入示例1-**单轮对话**：

```
{"conversations": [{"question": "图片中是否存在垃圾溢出垃圾桶的不文明卫生问题？", "answer": "是"}]}
```

数据处理后：

```
[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：[unused18][unused19].....[unused19][unused20]图片中是否存在垃圾溢出垃圾桶的不文明卫生问题？[unused10][unused9]助手：是[unused10]
```

输入示例2-**多轮对话**：

```
{"conversations": [{"question": "图片中是否存在垃圾溢出垃圾桶的不文明卫生问题？", "answer": "是"},
{"question": "图中是否存在共享单车按规定停放的情况？", "answer": "否"}]}
```

数据处理后：

```
[unused9]系统：你生成综合质量（有用性，事实性，无害性）极好的回复。[unused10][unused9]系统：你是一个擅长于文章写作、文本分析、代码编写、数学运算、逻辑推理的AI助手，专注于针对用户的问题给出高质量解答。[unused10][unused9]用户：[unused18][unused19].....[unused19][unused20]图片中是否存在垃圾溢出垃圾桶的不文明卫生问题？[unused10][unused9]助手：是[unused10][unused9]用户：图中是否存在共享单车按规定停放的情况？[unused10][unused9]助手：否[unused10]
```

14.2 训练图像问答大模型

14.2.1 图像问答大模型训练流程与选择建议

多模态大模型训练流程介绍

多模态大模型的训练分为两个关键阶段：预训练和微调。

版权所有 © 华为云计算技术有限公司

- **预训练阶段：**在这一阶段，模型通过学习大规模通用数据集来掌握语言的基本模式和语义。这一过程为模型提供了处理各种语言任务的基础，如阅读理解、文本生成和情感分析，但它还未能针对特定任务进行优化。

  预训练阶段完成之后，还可以继续进行训练，这一过程称为**增量预训练**。增量预训练是在已经完成的预训练的基础上继续训练模型。增量预训练旨在使模型能够适应新的领域或数据需求，保持其长期的有效性和准确性。
- **微调阶段：**基于预训练的成果，微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。这一阶段使模型能够精确执行如文案生成、代码生成和专业问答等特定场景中的任务。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。

多模态大模型选择建议

选择合适的多模态大模型类型有助于提升训练任务的准确程度。您可以根据模型**可处理上下文Token长度**，选择合适的模型，从而提高模型的整体效果，详见**表14-5**。

表 14-5 不同系列多模态大模型对处理文本的长度差异

模型	场景	上下文Token长度	模型功能简介
盘古-MM-M2-图生文模型-16K	-	16K	百亿参数多模态理解大模型的图生文功能模型，对应功能模型，支持预训练和SFT微调工作流，支持16K序列，基于Snt9B卡可支持64卡起训，Snt9B卡支持8卡部署，此模型版本支持在线推理、能力调测特性。

多模态大模型训练类型选择建议

平台针对多模态大模型提供了两种训练类型，包括预训练、微调，二者区别详见**表14-6**。

 版权所有 © 华为云计算技术有限公司

表 14-6 预训练、微调训练类型区别

训练方式	训练目的	训练数据	模型效果	应用场景举例
预训练	**关注通用性**：预训练旨在让模型学习广泛的通用知识，建立词汇、句法和语义的基础理解。通过大规模的通用数据训练，模型可以掌握丰富的语言模式，如语言结构、词义关系和常见的句型。	**使用大规模通用数据**：通常使用海量的无监督数据（如文本语料库、百科文章），这些数据覆盖广泛的领域和语言表达方式，帮助模型掌握广泛的知识。	**适合广泛应用**：经过预训练后，模型可以理解自然语言并具备通用任务的基础能力，但还没有针对特定的业务场景进行优化。预训练后的模型主要用于多个任务的底层支持。	通过使用海量的互联网文本语料对模型进行预训练，使模型理解人类语言的基本结构。
微调	**关注专业性**：微调是对预训练模型的参数进行调整，使其在特定任务中达到更高的精度和效果。微调的核心在于利用少量的特定任务数据，使模型的表现从通用性向具体任务需求过渡。	**使用小规模的特定任务数据**：微调通常需要小规模但高质量的标注数据，直接与目标任务相关。通过这些数据，模型可以学习到任务特定的特征和模式。	**在特定任务上具有更高的准确性**：微调后的模型在具体任务中表现更优。相较于预训练阶段的通用能力，微调能使模型更好地解决细分任务的需求。	在一个客户服务问答系统中，可以用特定领域（如电商、保险）的对话数据对预训练模型进行微调，使其更好地理解和回答与该领域相关的问题。

14.2.2 创建图像问答大模型训练任务

创建多模态大模型预训练任务

创建多模态大模型预训练任务步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-1 我的空间



版权所有 © 华为云计算技术有限公司

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“创建训练任务”页面，模型来源选择“模型广场”，模型类型选择“多模态大模型”，训练类型选择“预训练”。模型选择完成后，参考**表14-7**完成训练参数设置。

表 14-7 多模态大模型预训练参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： ● 来源：选择“模型广场”。 ● 类型：选择“多模态大模型”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	高级设置 checkpoints	断点续训，模型训练任务过程中，用于保存模型权重和状态的机制。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。 开启断点续训后对于未训练完成的模型，可以通过单击训练任务名称，进入“训练结果”页签，通过“编辑训练”入口选中需要进行断点续训的checkpoint进行续训。 **说明** 保存过多的checkpoint会占用较大的OBS存储。
训练参数	热身比例	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练。使用热身可以使得模型收敛速度更快，效果更佳。
	学习率	学习率决定了每次训练时模型参数更新的幅度。选择合适的学习率非常重要：如果学习率太大，模型可能会无法收敛；如果学习率太小，模型的收敛速度会变得非常慢。

 版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	模型保存步数	指每训练一定数量的步骤（或批次）后，模型的状态就会被保存下来。
	单步迭代时处理的数据批量大小	全局batch_size，指定每次迭代时处理的数据批量大小。在数据并行和流水线并行开启情况下，全局batch_size等于local_batch_size乘micro_size乘data_parallelism。
	训练轮数	指完成全部训练数据集训练的次数。
	优化器	优化器参数指的是用于更新模型权重的优化算法的相关参数，可以选择adam优化器。
	学习率衰减比率	学习率衰减后的比率，用于控制训练过程中学习率的下降幅度。经过衰减后，学习率的最低值由初始学习率和衰减比率决定。其计算公式为：最低学习率 = 初始学习率 * 学习率衰减比率。也就是说，学习率在每次衰减后不会低于这个计算出来的最低值。
数据配置	训练数据	选择训练模型所需的数据集。要求数据集经过发布操作，发布数据集操作方法请参见**发布训练后的多模态大模型**。
资源配置	资源池	选择训练模型所需的资源池，以及算力。 当前展示的完成本次训练所需要的最低算力要求，单击“+”可增加训练算力。如果无法增加则表明已经到达资源池的最高算力。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建多模态大模型增量预训练任务

在模型完成**创建多模态大模型预训练任务**后，可以对训练后的模型继续训练，该过程称为“增量预训练”。

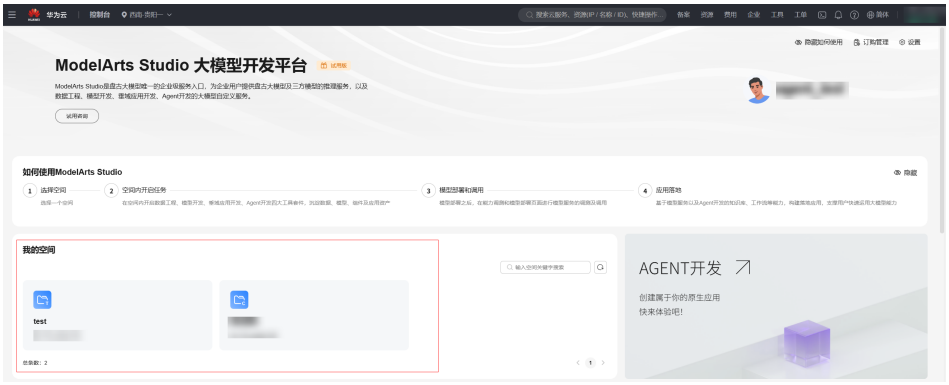
创建多模态大模型增量预训练任务前，请确保有已完成预训练并发布的多模态大模型。

创建多模态大模型增量预训练任务的步骤如下：

 版权所有 © 华为云计算技术有限公司

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 14-2 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“选择模型”页面，“来源”选择“我的资产”，“类型”选择“多模态大模型”，并从模型列表中，选择已完成预训练并发布的多模态大模型。
4. 其余参数配置等步骤同**创建多模态大模型预训练任务**章节描述一致。

创建多模态大模型微调任务

创建多模态大模型全量微调任务步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-3 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“创建训练任务”页面，模型类型选择“多模态大模型”，训练类型选择“微调”。模型选择完成后，设置plog日志，参考**表14-9**完成训练参数设置。

版权所有 © 华为云计算技术有限公司

表 14-8 设置 plog 日志

参数类别	参数名称	说明
高级设置	plog日志	开启后plog日志功能，会记录模型训练过程中的执行过程、状态、错误等信息，盘古技术支持人员可通过plog日志定位模型训练问题。训练任务开始后可以在承载租户的OBS桶中查看。plog日志分级如下： ● DEBUG：记录模型训练过程中代码的执行流、变量状态、参数的变化等。 ● INFO：记录模型训练过程中的运行信息，通常包括一些重要的状态变化、执行的关键步骤、模型训练进度等。 ● WARNING：记录模型训练过程中的告警信息，可用于识别潜在的问题或异常情况。 ● ERROR：记录模型训练过程中的执行错误信息。

表 14-9 多模态大模型微调训练参数说明

参数分类	训练参数	参数说明
训练配置	模型类型	选择“多模态大模型”
	训练类型	选择“微调”。
	基础模型	选择微调训练所用的基础模型， 可从“预置模型”或者“我的模型”中进行选择。
	高级设置 checkpoints	断点续训，模型训练任务过程中，用于保存模型权重和状态的机制。“关闭”则不开启断点续训；“自动”保存训练过程中所有的checkpoint；“自定义”保存指定数量的checkpoint。需要注意：保存过多的checkpoint会占用较大的OBS存储。 开启断点续训后对于未训练完成的模型，可以通过单击训练任务名称，进入“训练结果”页签，通过“编辑训练”入口选中需要进行断点续训的checkpoint进行续训。

 版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
训练参数	热身比例	热身阶段占整体训练的比例。刚开始训练时若选择一个较大的学习率，可能带来模型的不稳定，选择使用warmup热身的方式，可以使得开始训练的热身阶段内学习率较小，在热身阶段的小学习率下，模型可以慢慢趋于稳定，等模型相对稳定后再逐渐提升至预设的最大学习率进行训练。使用热身可以使得模型收敛速度更快，效果更佳。
	学习率	学习率决定了每次训练时模型参数更新的幅度。选择合适的学习率非常重要：如果学习率太大，模型可能会无法收敛；如果学习率太小，模型的收敛速度会变得非常慢。
	模型保存步数	指每训练一定数量的步骤（或批次）后，模型的状态就会被保存下来。
	单步迭代时处理的数据批量大小	设置在并行训练中，每个微批次包含的数据批量大小，适当的数据批量大小能够确保训练各个阶段都能充分利用计算资源，提升并行效率。
	训练轮数	指完成全部训练数据集训练的次数。
	优化器	优化器参数指的是用于更新模型权重的优化算法的相关参数，可以选择adam优化器。
	学习率衰减比率	学习率衰减后的比率，用于控制训练过程中学习率的下降幅度。经过衰减后，学习率的最低值由初始学习率和衰减比率决定。其计算公式为：最低学习率 = 初始学习率 * 学习率衰减比率。也就是说，学习率在每次衰减后不会低于这个计算出来的最低值。
数据配置	训练数据	选择训练模型所需的数据集。要求数据集经过发布操作，发布数据集操作方法请参见**发布数据集**。
资源配置	资源池	选择训练模型所需的资源池，以及算力。 当前展示的完成本次训练所需要的最低算力要求，单击“+”可增加训练算力。如果无法增加则表明已经到达资源池的最高算力。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

 版权所有 © 华为云计算技术有限公司

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

微调 CKPT 产物发布成可部署资产

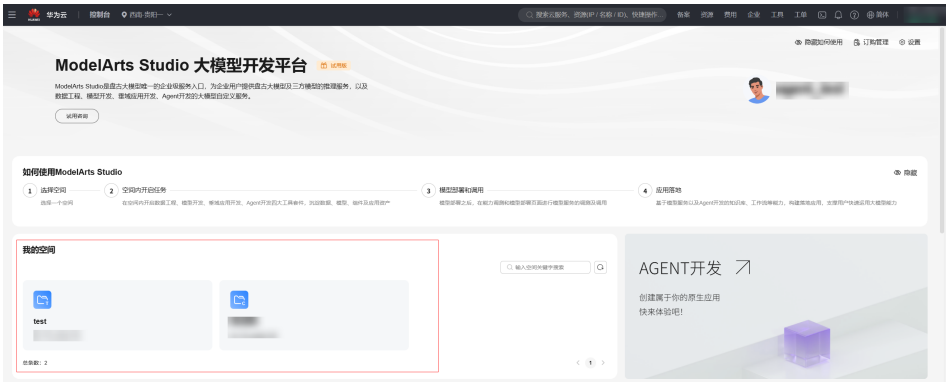
开启断点续训后的微调任务，在模型训练任务过程中，会依据用户设置的模型保存步数自动保存中间CKPT产物，用于后续的断点续训任务。

微调CKPT产物不仅支持微调的断点续训任务，也支持单独发布成推理资产，用于推理部署。原因是在一次训练任务中，可能由于训练数据量、学习率设置等因素，中间断点模型的泛化性以及推理效果更好。

微调CKPT产物发布成可部署资产，操作步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-4 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“编辑训练”，在要发布的CKPT产物下方，单击“发布”。

图 14-5 微调 CKPT 产物发布



4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

 版权所有 © 华为云计算技术有限公司

14.2.3 查看图像问答大模型训练状态与指标

模型启动训练后，模型的训练任务状态请在“模型训练”页面进行查看。单击任务名称可以进入详情页查看训练指标、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见**表14-10**。

表 14-10 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见**表14-11**。

图 14-6 查看训练指标

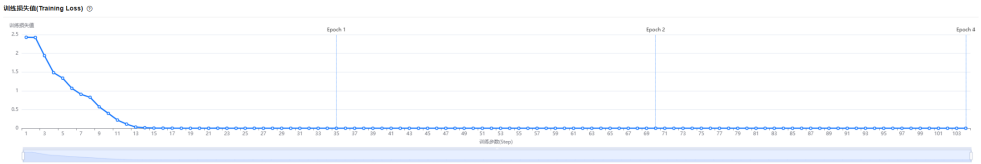


表 14-11 训练指标说明

模型	训练指标	指标说明
多模态大模型	训练损失值 (Training Loss)	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。

版权所有 © 华为云计算技术有限公司

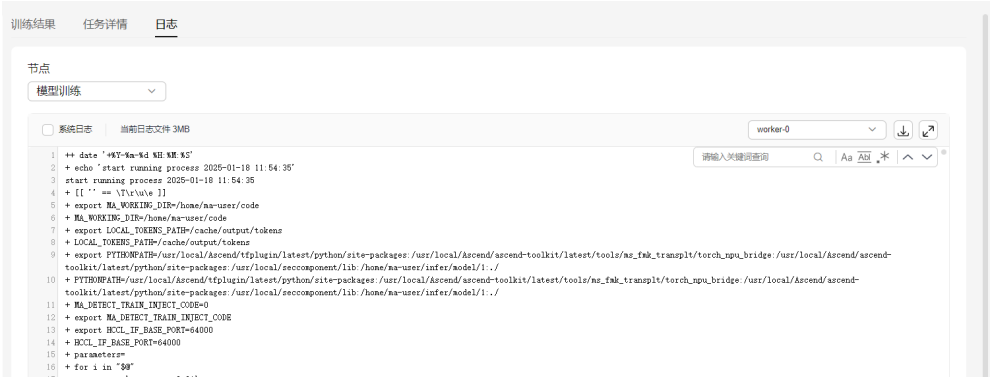
获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 14-7 获取训练日志



14.2.4 发布训练后的多模态大模型

多模态大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-8 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 14-9 发布模型



版权所有 © 华为云计算技术有限公司

4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

14.2.5 管理多模态大模型训练任务

在训练任务列表中，任务创建者可以对任务进行克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-10 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 克隆。单击操作列的“克隆”，参照**创建图像问答大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“删除”，可以删除当前不需要的训练任务。

14.2.6 多模态大模型训练常见报错与解决方案

多模态大模型训练常见报错及解决方案请详见**表14-12**。

表 14-12 多模态大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空。	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。

版权所有 © 华为云计算技术有限公司

常见报错	问题现象	原因分析	解决方案
训练日志提示“root: XXX valid number is 0”。	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。
训练日志提示“ValueError: label_map not match”。	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

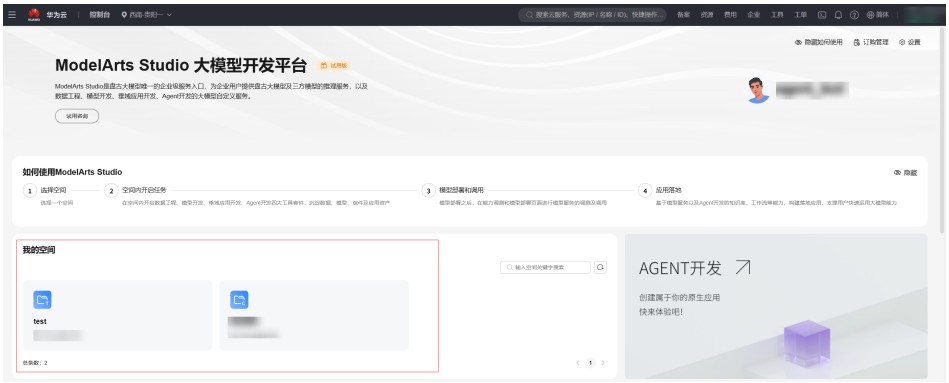
14.3 部署图像问答大模型

14.3.1 创建多模态大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-11 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，模型来源选择“模型广场”，模型类型选择“多模态大模型”，参考表14-13完成部署参数设置，启动模型部署。

 版权所有 © 华为云计算技术有限公司

表 14-13 多模态大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	模型来源	选择“模型广场”。
	模型类型	选择“多模态大模型”。
	部署模型	选择需要进行部署的模型。
	部署方式	云上部署：算法部署至平台提供的资源池中。
安全护栏	安全护栏开关	安全护栏保障模型调用安全。
	版本选择	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置（选择云上部署时）	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数，单次部署服务时，部署实例个数建议不大于10，否则可能触发限流导致部署失败。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

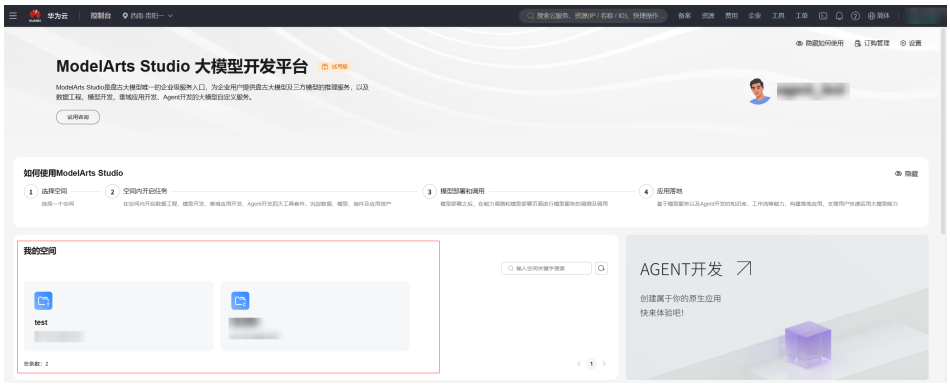
4. 参数填写完成后，单击“立即部署”。

14.3.2 查看多模态大模型部署任务详情

部署任务创建成功后，可以查看大模型部署任务详情，具体步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-12 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建多模态大模型部署任务**后，可以查看模型的部署状态。

版权所有 © 华为云计算技术有限公司

当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

图 14-13 部署详情



14.3.3 管理多模态大模型部署任务

模型更新、修改部署

部署任务创建后，如果需要修改模型部署信息，可以在详情页面单击右上角“模型更新”或者“修改部署”进行修改。更新模型时可以替换模型、修改环境变量。修改部署时，模型无法进行替换。

模型更新、修改部署后进行升级操作时，可以选择全量升级、滚动升级两种方式。

全量升级：旧版本和新版本的服务会同时运行，直到新版本完全替代旧版本。在新版本部署完成前旧版本仍然可以使用。

滚动升级：通过空出部分实例资源用于滚动升级，逐个或逐批实例停止旧版本并启动新版本。滚动升级时可以修改部署实例数。如果选择缩实例升级则先删除旧版本再进行升级操作，期间旧版本不可使用。

图 14-14 模型更新



 版权所有 © 华为云计算技术有限公司

图 14-15 修改部署

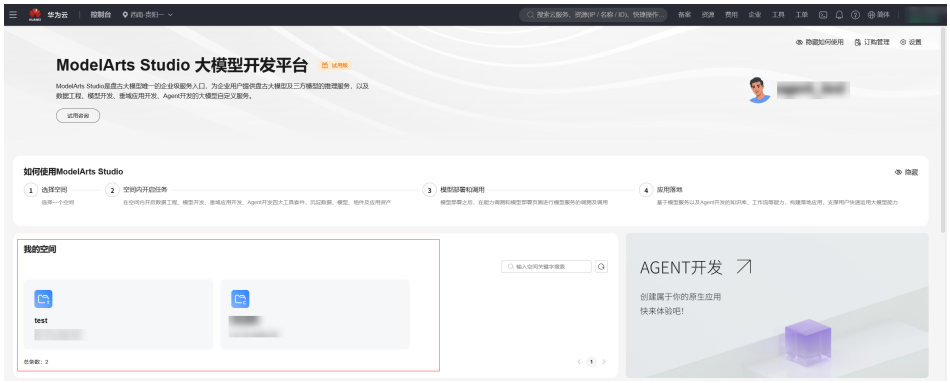


模型部署实例扩缩容

完成**创建多模态大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-16 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确认”。

14.4 评测图像问答大模型

14.4.1 创建多模态大模型评测数据集

多模态大模型支持人工评测和基于大模型的评测，人工评测方式需要通过人工创建的评测数据集和评测指标项对模型生成的回答进行评测，评测时需要人工基于创建好的评测项对模型回答进行打分，评测完成后会基于打分结果生成评测报告。

版权所有 © 华为云计算技术有限公司

创建多模态大模型的评测任务，需要用户在自己的obs桶中准备好数据。数据格式为图片存放目录+jsonl文件，目录结构如下：

图 14-17 目录结构



其中jsonl文件内容格式为图片+QA对格式：

```
{"image_name": "images/108762.jpg", "conversations": [{"answer": "红色和蓝色消防栓附近的标志独特之处在于其手写设计，看起来像是由一个七岁孩子创作的。标志上写着"Edic is CTG New West"，似乎是一个手工设计的板上有涂鸦。这种非正式和创造性的美学，结合了街头艺术元素，使该标志与更传统和正式的标志相比具有与众不同和独特的特点。", "question": "
红色和蓝色消防栓附近的标志有什么独特之处？\n<image>"}]
```

说明

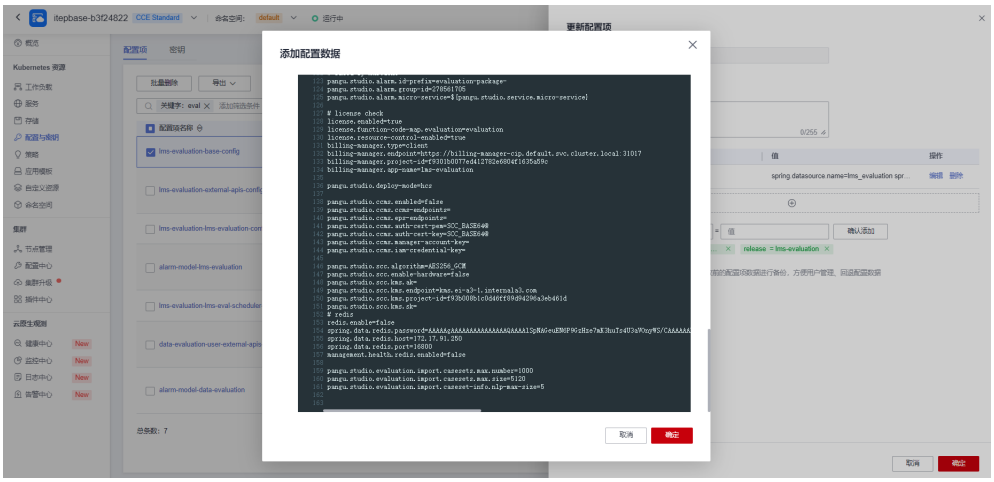
cce评测服务启动配置参数说明如下：

pangu.studio.evaluation.import.casesets.max.number: 所有评测任务数据集数量最多不超过多少，默认100条，上限与数据集所在的OBS桶容量相关。

pangu.studio.evaluation.import.casesets.max.size: 所有评测任务数据集总大小不超过多少MB，默认512MB，上限与数据集所在的OBS桶容量相关。

pangu.studio.evaluation.import.caseset-info.mllm-max-size: 多模态单次上传最大文件尺寸，默认100MB，上限与数据集所在的OBS桶容量相关。

pangu.studio.evaluation.import.caseset-info.mllm-max-data: 多模态单次上传最大数据条数，默认10000条，上限与数据集所在的OBS桶容量相关。



14.4.2 创建多模态大模型评测任务

创建多模态大模型评测任务前，请确保已完成**创建多模态大模型评测数据集**操作。

预训练的多模态大模型不支持评测。

 版权所有 © 华为云计算技术有限公司

创建多模态大模型人工评测任务

创建多模态大模型人工评测任务步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

图 14-18 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建人工评测任务”。
3. 在“创建人工评测任务”页面，参考**表14-14**完成部署参数设置。

表 14-14 多模态大模型人工评测任务参数说明

参数分类	参数名称	参数说明
评测对象	评测类型	选择“多模态大模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。
评测配置	评测指标	由用户自定义评测指标并填写评测标准。
	评测数据集	待评测的数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数设置完成后，单击“立即创建”创建评测任务。

   在“评测任务 > 人工评测”页面，单击操作列“在线评测”，进入人工评测页面。在人工评测页面对每条数据进行评估并打分，直到所有数据评估完成后，单击“提交”，提交评估结果。

   - 评测详情页面，打开“盲测”按钮，会隐藏模型名称，进行盲评。

版权所有 © 华为云计算技术有限公司

- 单击“存疑”或者“作废”进行用例的存疑或作废，若取消存疑或作废，单击“取消存疑”或“取消作废”进行处理。
- 单击备注下方“单击以添加备注”，可以进行新增备注。
- 评测页面，长按鼠标左键选中需要标记的文本内容，单击“标记”可以标记成重点内容。

图 14-19 多模态大模型人工评测



创建多模态大模型自动评测任务

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“评测中心 > 评测任务”，单击界面右上角“创建自动评测任务”。
3. 在“创建自动评测任务”页面选择“多模态大模型”，参考表14-15完成评测任务参数设置

表 14-15 自动评测任务参数说明

参数分类	参数名称	参数说明
评测对象	评测类型	选择“模型评测”、“多模态大模型”。
	服务来源	支持已部署服务。单次最多可评测10个模型。 • 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。
评测配置（基于规则）	评测规则	选择“基于规则”。
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。 • 自定义评测集：由用户指定评测指标（准确率、BLEU）并上传评测数据集进行评测。
	评测数据集	选择“单个评测集”时需要上传待评测数据集。
	评测结果存储位置	模型评测结果的存储位置。

版权所有 © 华为云计算技术有限公司

参数分类	参数名称	参数说明
评测配置（基于大模型）	评测类型	选择“多模态大模型”。
	评测规则	选择“基于大模型”。
	评测数据集	需要用户在自己的obs桶中提前准备好数据。
	评测结果存储位置	模型评测结果的存储位置。
	选择模式	• 评分模式：每个模型独立评测。 • 对比模式：多个模型对比 。
	裁判员模型	裁判员模型可以是已部署的NLP服务，也可以是外部服务。外部服务的创建可以参考**创建API服务（可选）**
	打分规则	打分prompt有评分模式的prompt和对比模式的prompt。prompt中的${metric}，${max_score}，${score_desc}和${steps}的值可以通过右边对应的编辑栏设置，最终请求给裁判模型的prompt是用设置的值替换了变量的prompt。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 评测任务参数设置完成后单击“立即创建”。
5. 返回“评测任务 > 自动评测”页面，可查看创建好的评测任务，此时任务状态显示为“初始化”，等到自动评测完成，状态显示为“已完成”时，可以单击操作列“评测报告”查看模型评测结果。

   在“评测报告”页面可以查看详细的得分以及评测数据信息。

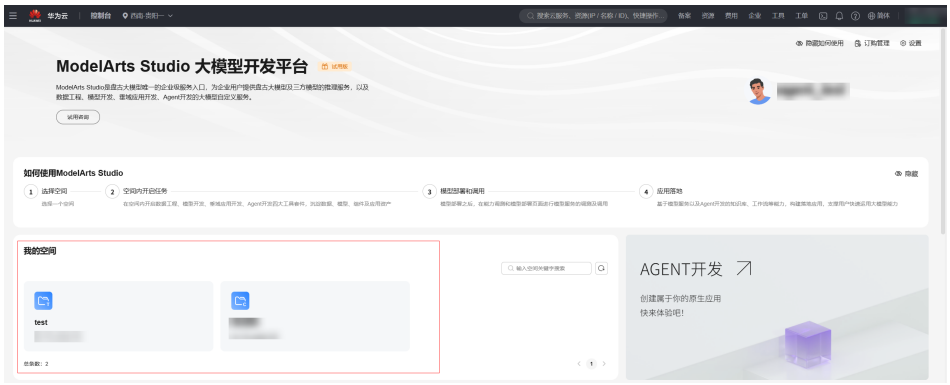
14.4.3 查看多模态大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 14-20 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**多模态大模型评测指标说明**。

多模态大模型评测指标说明

多模态大模型支持人工评测、自动评测，各指标说明如下。

表 14-16 多模态大模型评测指标说明

模型类型	评测指标（人工评测）	指标说明
多模态大模型	平均分数	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
	goodcase占比	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
	badcase占比	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分以下的占比。
	准确性及其他用户自定义的指标	由用户定义的指标。

表 14-17 多模态大模型评测指标说明

模型类型	评测指标（自动评测-基于规则-使用预置评测集）	指标说明
多模态大模型	综合能力	综合能力是计算所有数据集通过率的加权平均数。

版权所有 © 华为云计算技术有限公司

表 14-18 多模态大模型自动评测指标说明

模型类型	评测指标（自动评测-自定义评测集）	指标说明
多模态大模型	准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
	BLEU-1	模型生成句子与实际句子在单字层面的匹配度，分数越高，表示模型效果越好。
	BLEU-2	模型生成句子与实际句子在中词组层面的匹配度，分数越高，表示模型效果越好。
	BLEU-4	模型生成结果和实际句子的加权平均精确率，分数越高，表示模型效果越好。

表 14-19 多模态大模型评测指标说明

模型类型	评测指标（自动评测-基于大模型）	指标说明
多模态大模型	裁判员模型打分	数据集中每个用例，裁判员模型给的评分值。
	平均值	数据集中所有用例得分的平均值。
	中位数	数据集中所有用例得分的中位数
	标准差	数据集中所有用例得分的标准差
	分位	(win+tie)/(lose+tie)
	分位（剔除tie_bad）	剔除tie_bad所得分位，(win+tie_good)/(lose+tie_good)
	分位（剔除tie_good）	剔除tie_good所得分位，(win+tie_bad)/(lose+tie_bad)

14.4.4 管理多模态大模型评测任务

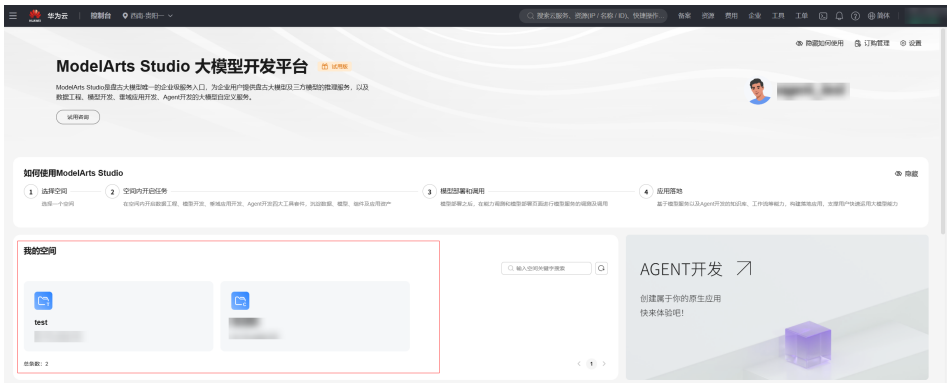
管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 14-21 我的空间



2. 在左侧导航栏中选择“评测中心 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前状态为“已完成”的不需要的评测任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

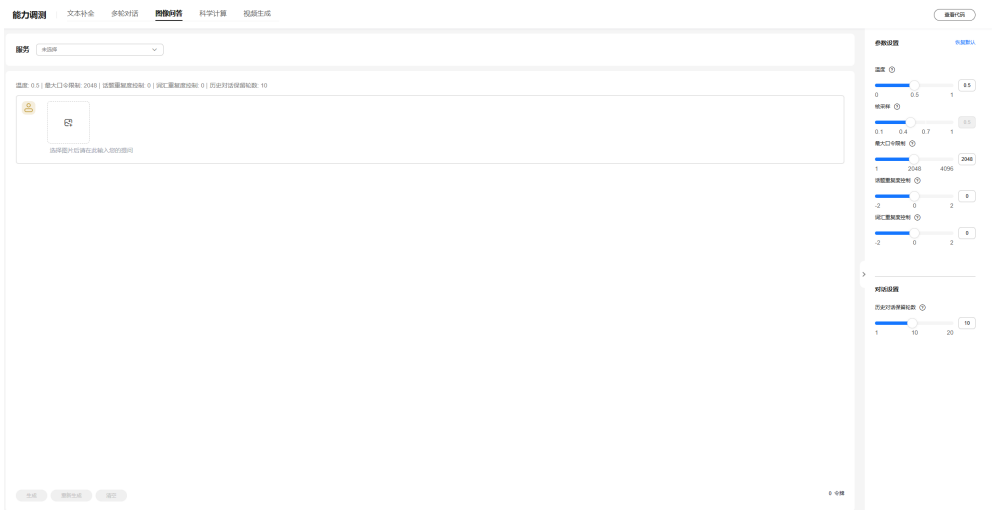
14.5 调用图像问答大模型

14.5.1 使用“体验中心”调用多模态大模型

“体验中心”功能支持用户直接调用订购的预置模型，或经过训练的模型。在体验模型功能前请先完成**创建多模态大模型部署任务**操作。

多模态大模型支持图像问答能力，在选择好订购的模型后，上传需要分析的图片，并在输入框中输入问题，模型就会基于图片内容回答问题。

图 14-22 调测多模态大模型



版权所有 © 华为云计算技术有限公司

表 14-20 多模态大模型体验中心参数说明

参数	说明
温度	用于控制生成文本的多样性和创造力。调高温度会使得模型的输出更多样性和创新性。
核采样	控制生成文本多样性和质量。调高核采样可以使输出结果更加多样化。
输出最大token长度	用于控制模型生成回复的长度，一般来说，较大的max_tokens值可以生成较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较简洁的回复，但也可能导致生成不完整或不连贯的内容，请根据场景选择合适的值。
话题重复度控制	用于控制生成文本中的重复程度。调高参数模型会更频繁地切换话题，从而避免生成重复内容。
词汇重复度控制	用于调整模型对频繁出现的Token的处理方式。调高参数会使模型减少相同词汇的重复使用，促使模型使用更多样化的词汇进行表达。
历史对话保留轮数	选择“图像问答”功能时具备此参数。表示系统能够记忆的历史对话数。

能力调测功能同时支持生成示例代码，支持生成Python、Shell代码。

14.5.2 使用 API 调用多模态大模型

获取多模态大模型 API 请求地址

API请求地址由API接口域名+API访问路径（URI）拼接形成。

模型在部署成功后，可以通过API调用多模态大模型。

表 14-21 多模态大模型 API 清单

API分类	API访问路径（URI）
对话问答	/v1/{project_id}/deployments/{deployment_id}/chat/completions

API接口域名格式为https://ip:port，获取方式参考《API参考》文档“获取大模型推理API接口域名”章节。

API访问路径（URI）中的deployment_id、project_id获取方式如下。

1. 使用最终租户登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称在“详情”页面获取API访问路径中的deployment_id参数。

 版权所有 © 华为云计算技术有限公司

只有处于“运行中”状态的模型才可以正常调用。

图 14-23 获取 deployment_id



3. 单击页面右上方用户名，并进入“个人设置”页面获取project_id。

图 14-24 获取 project_id



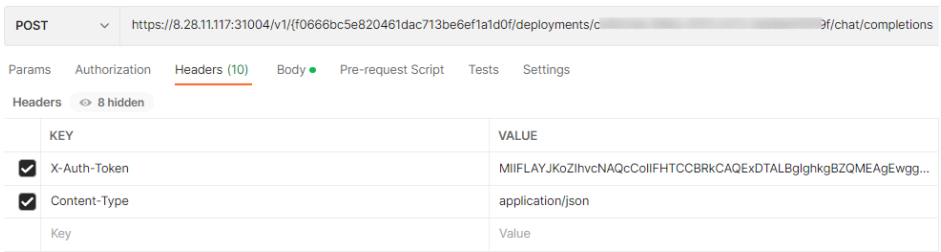
获取 Token

参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。

使用 API 调用多模态大模型

1. 在Postman中新建POST请求，并填入API请求地址。
2. 参考**图14-25**填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**获取Token**中获取的Token值。

图 14-25 填写多模态大模型 API

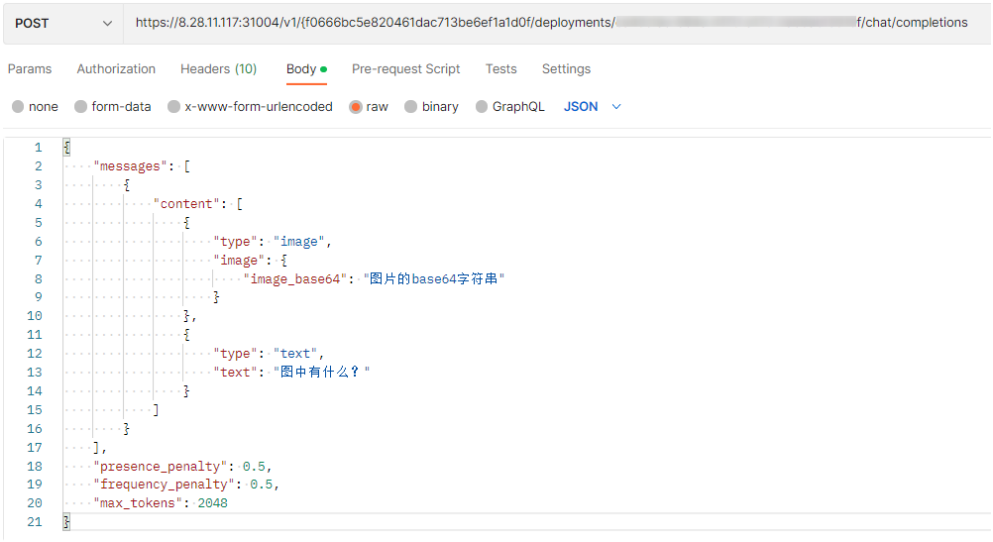


 版权所有 © 华为云计算技术有限公司

3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "messages": [
      {
         "content": [
            {
               "type": "image",
               "image": {
                  "image_base64": "图片的base64字符串"
               }
            },
            {
               "type": "text",
               "text": "图中有什么？"
            }
         ]
      }
   ],
   "presence_penalty": 0.5,
   "frequency_penalty": 0.5,
   "max_tokens": 2048
}
```

图 14-26 填写多模态大模型请求 Body



4. 单击Postman界面“Send”按钮，发送请求。

 版权所有 © 华为云计算技术有限公司

15 开发图像搜索模型

图像搜索模型介绍
使用数据工程构建图像搜索模型数据集
训练图像搜索模型
部署图像搜索模型
调用图像搜索模型

15.1 图像搜索模型介绍

基于服装图片数据标注的服装类型、款式信息、工艺类别标签进行训练。

目前支持的服装类型包括：男西服、女西服、男大衣、女大衣、男夹克、女夹克、男礼服、女礼服、男风衣、女风衣、T恤、男POLO衫、男猎装、女猎装、女马甲、男马甲、女西裙、女卫裤、男卫裤、女西裤、男西裤、女衬衣、男衬衣。

15.2 使用数据工程构建图像搜索模型数据集

图像搜索模型支持接入图片类数据集，目前使用异常检测图片数据集接口导入，数据格式为图片+txt格式，标注内容需要严格安装如下形式标注。

一个标注txt文件中每一行为一个标注信息，每行的标注信息格式如下：

- 格式一：图片名称;左上角x坐标;左上角y坐标;右下角x坐标;右下角y坐标;一级标签内容-二级标签内容-三级标签内容@

  示例一：示例.jpg;2173;157;2726;617;男衬衣-领型-八字领@
- 格式二：图片名称;左上角x坐标;左上角y坐标;右下角x坐标;右下角y坐标;一级标签内容-季节信息-场景信息-风格信息-颜色信息-花型信息-@

  示例二：示例.jpg;1476;203;3163;3037;男西服-SS-英式商务-英式商务-藏蓝-素色-@

训练数据建议大于5000，训练数据集需要划分为训练集和测试集。

 版权所有 © 华为云计算技术有限公司

15.3 训练图像搜索模型

15.3.1 创建图像搜索模型训练任务

约束

当前图像搜索模型支持预训练，适用于图像标签类训练任务。

创建图像搜索预训练步骤

创建图像搜索模型预训练步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“选择模型”页面，选择“模型广场 > 图像搜索 > 模型”，单击“确定”进入“创建训练任务”页面。
4. 在“创建训练任务”页面，参考**表15-1**完成训练参数设置，参数默认值在创建训练任务的时候会带出。

表 15-1 图像搜索模型预训练参数说明

参数分类	训练参数	参数说明
训练配置	选择模型	可以修改如下信息： • 来源：选择“模型广场”。 • 类型：选择“图像搜索 ”，并选择训练所用的基础模型和版本。
	训练类型	选择“预训练”。
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。 • 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 • 自动：自动保存训练过程中的所有checkpoints。 • 自定义：根据设置保存指定数量的checkpoints。
训练参数	epochs	训练轮次，表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	batch_size	表示每个训练步骤中使用的样本数量。较大的批量可以提供更稳定的梯度，但可能会增加计算资源的使用和训练时间。
	training_categories	需要训练的标签，每个标签以逗号隔开，不输入时默认训练数据中的标签全部训练。

版权所有 © 华为云计算技术有限公司

参数分类	训练参数	参数说明
	数据集	选择已发布的数据集进行本次训练。
资源配置	计费模式	选择训练当前任务的计费模式。
	训练单元	选择训练模型所需的训练单元。训练单元 = 单实例训练单元 * 实例数。 当前展示的完成本次训练所需要的最低训练单元要求。
	单实例训练单元数	选择单实例训练单元数。
	实例数	选择实例数。 当实例数大于1，将启动多节点分布式训练。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
发布模型	开启自动发布	开启自动发布后，模型训练完成的最终产物会自动发布为空间资产，以便对模型进行压缩、部署、评测等操作或共享给其他空间。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

5. 参数填写完成后，单击“立即创建”。
6. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

15.3.2 查看图像搜索模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见**表15-2**。

表 15-2 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。

 版权所有 © 华为云计算技术有限公司

训练状态	训练状态含义
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见图15-1和表2 训练指标说明。

图 15-1 查看训练指标

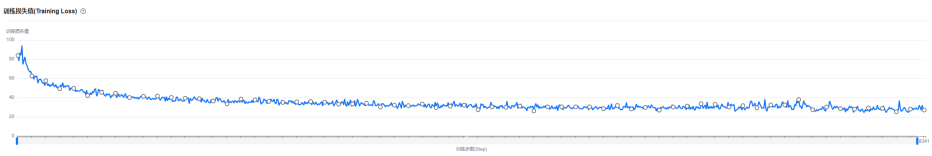


表 15-3 训练指标说明

模型	训练指标	指标说明
Pangu-AIKits-ImageSearch-Large-Tagging-1.0.0	训练损失	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	Recall	召回率是指在所有实际为正类的样本中，被模型正确预测为正类的比例。数值越高，表明模型在检测正类样本时的全面性越高。
	mAP	类别的平均精确度，单类别时显示的就是精准率，即指在模型预测为正类的样本中，真正类样本的比例。数值越高，表明模型在检测正类样本时的准确性越高。
	AP50	模型在测试集评估中预测框和目标框交并比大于50%时的平均精度。
	F1	精准率和召回率的调和平均数，用于综合评估分类模型的性能。

获取训练日志

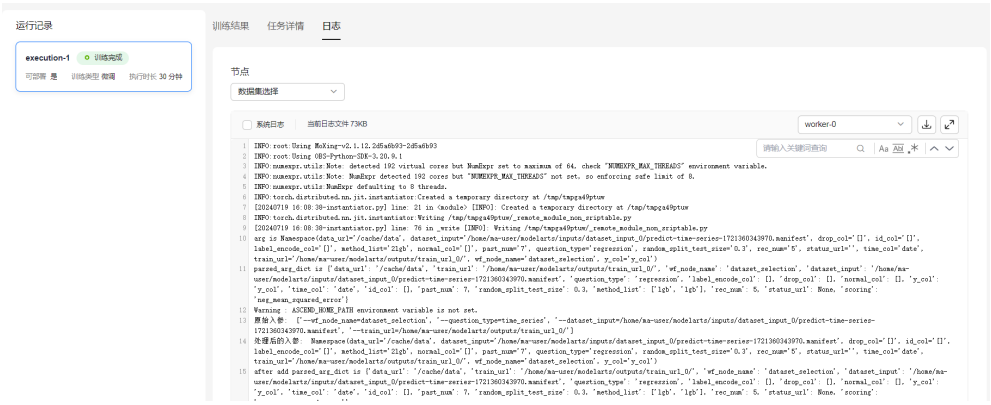
单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

版权所有 © 华为云计算技术有限公司

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 15-2 获取训练日志



15.3.3 发布训练后的图像搜索模型

模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 15-3 训练结果页面



4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
5. 发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

15.3.4 管理图像搜索模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、启动、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

版权所有 © 华为云计算技术有限公司

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 克隆。单击操作列的“更多 > 克隆”，参照填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“更多 > 停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的训练任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

15.4 部署图像搜索模型

15.4.1 创建图像搜索模型部署任务

平台支持部署训练后的模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“选择模型”页面，选择“我的资产 > 图像搜索模型 > 模型”，单击“确定”进入“创建部署”页面。
4. 在“创建部署”页面，参考**表1 创建图像搜索模型部署任务**完成部署参数设置。
5. 参数填写完成后，单击“立即部署”。

表 15-4 图像搜索模型部署参数说明

参数分类	部署参数	参数说明
部署配置	选择模型	可以修改如下信息： ● 来源：选择“我的资产”。 ● 类型：选择“图像搜索”，并选择需要进行部署的模型和版本。
	部署方式	选择“云上部署”。
资源配置	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。

版权所有 © 华为云计算技术有限公司

参数分类	部署参数	参数说明
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

15.4.2 查看图像搜索模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建图像搜索模型部署任务**后，可以查看模型的部署状态。

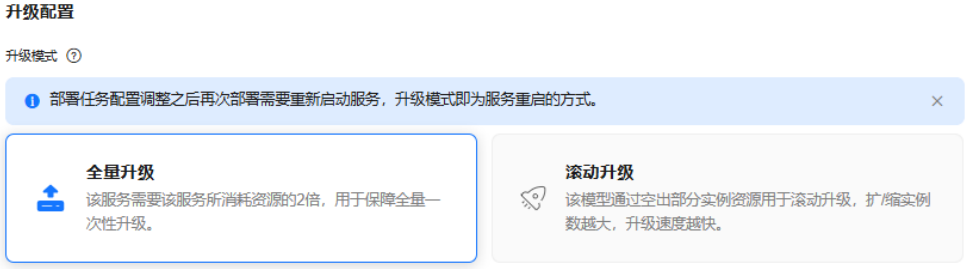
   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。
3. 单击模型名称进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

15.4.3 管理图像搜索模型部署任务

完成**创建图像搜索模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下

1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 15-4 升级模式



 版权所有 © 华为云计算技术有限公司

> **说明**
>
> 升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成创建**创建图像搜索模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

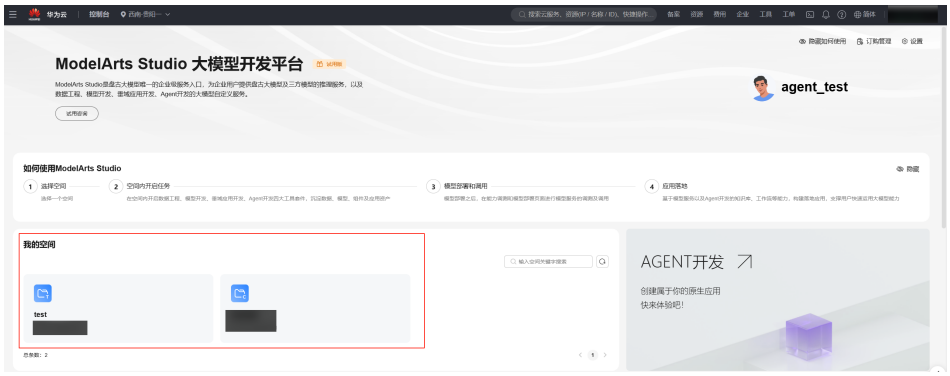
1. 登录ModelArts Studio大模型开发平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确定”。

管理模型部署任务

在左侧导航栏中选择“模型开发 > 模型部署”，在模型服务列表中，可以对服务进行启动、停止、调用统计和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 15-5 我的空间



2. 在左侧导航栏中选择“模型开发 > 模型部署”，可进行如下操作：
   - 启动。单击操作列的“启动”，可以启动被停止的模型服务，启动后可以调用该模型服务进行推理。
   - 停止。单击操作列的“停止”，可以停止处于“运行中”模型服务。
   - 调用统计。单击操作列的“更多 > 调用统计”，展示该模型的调用统计，包含调用次数和token统计。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的模型任务。

   > **说明**
   >
   > 删除属于高危操作，删除前请确保当前任务不再需要。

15.5 调用图像搜索模型

版权所有 © 华为云计算技术有限公司

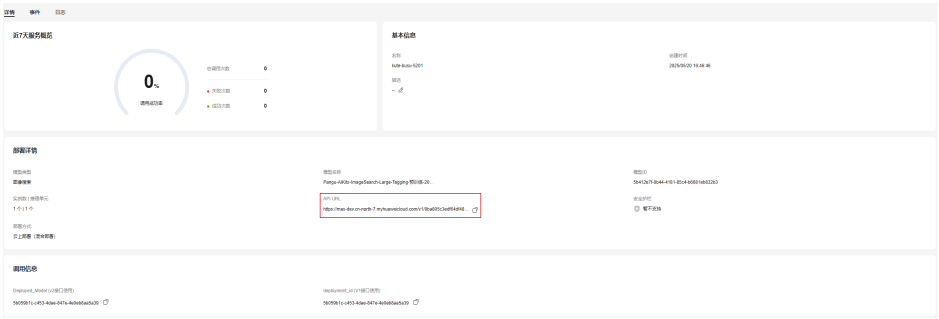
15.5.1 使用 API 调用图像搜索模型

获取图像搜索模型 API 请求地址

1. 登录ModelArts Studio平台，在“我的空间”（详见**工作空间介绍**）分页，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击“模型名称”在“详情”页面获取API请求地址。

   只有处于“运行中”状态的模型才可以正常调用。

   图 15-6 云上部署获取图像搜索模型 API 请求地址

3. 获取API接口域名后，结合API访问路径拼接成完整的API接口。

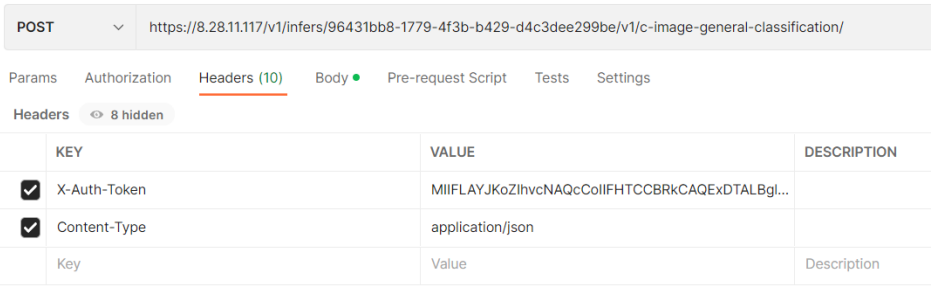
获取 Token

参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。

使用 API 调用图像搜索大模型

1. 在Postman中新建POST请求，并填入API请求地址。
2. 参考填**图15-7**写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**获取Token**中获取的Token值。

   图 15-7 填写图像搜索模型 API

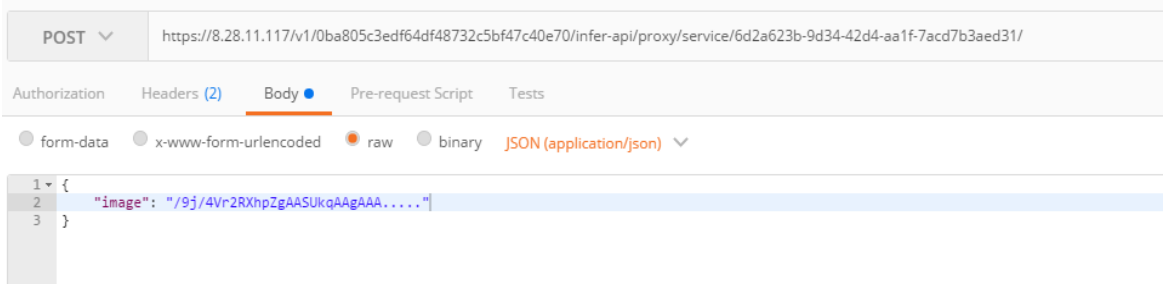
   

3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。注意需要将图片转成base64编码传入。整个请求体大小建议不超过4M，建议使用JPG、PNG、JPEG、BMP格式的图片。

```
{
  "image": "/9j/4Vr2RXhpZgAASUkqAAgAAA....."
}
```


版权所有 © 华为云计算技术有限公司

图 15-8 填写图像搜索模型请求 Body



4. 单击Postman界面“Send”按钮，发送请求。

版权所有 © 华为云计算技术有限公司

16 开发提示词工程

什么是提示词工程
获取提示词模板
撰写提示词
横向比较提示词效果
批量评估提示词效果
发布提示词

16.1 什么是提示词工程

提示词工程简介

提示词工程（Prompt Engineering）是一个较新的学科，应用于开发和优化提示词（Prompt），帮助用户有效地将大语言模型用于各种应用场景和研究领域。掌握提示词工程相关技能将有助于用户更好地了解大语言模型的能力和局限性。

提示词工程不仅是关于设计和研发提示词，它包含了与大语言模型交互和研发的各种技能和技术。提示工程在实现和大语言模型交互、对接，以及理解大语言模型能力方面都起着重要作用。用户可以通过提示词工程来提高大语言模型的安全性，还可以赋能大语言模型，如借助专业领域知识和外部工具来增强大语言模型的能力。

提示词基本要素

您可以通过简单的提示词（Prompt）获得大量结果，但结果的质量与您提供的信息数量和完善度有关。一个提示词可以包含您传递到模型的指令或问题等信息，也可以包含其他种类的信息，如上下文、输入或示例等。您可以通过这些元素来更好地指导模型，并因此获得更好的结果。提示词主要包含以下要素：

- **指令**：希望模型执行的特定任务或指令，如总结、提取、生成等。
- **上下文**：包含外部信息或额外的上下文信息，引导语言模型更好地响应。
- **输入数据**：用户输入的内容或问题。
- **输出指示**：指定输出的类型或格式。

 版权所有 © 华为云计算技术有限公司

提示词所需的格式取决于您希望语言模型完成的任务类型，并非所有以上要素都是必须的。

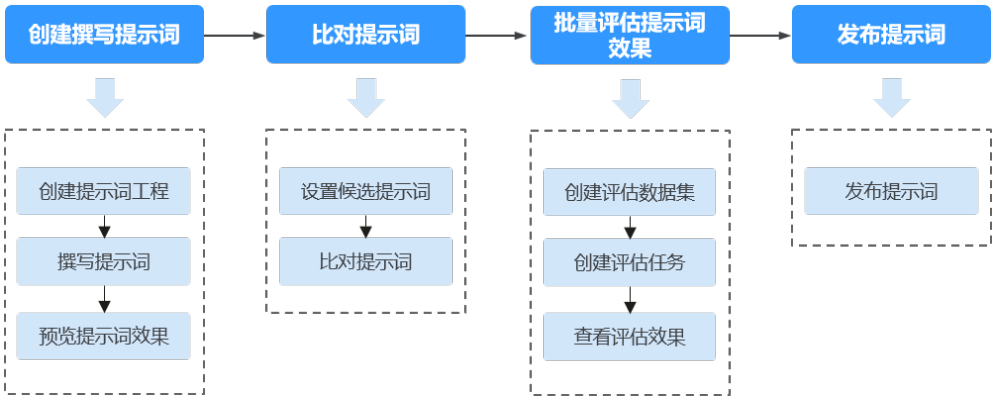
提示词工程使用流程

ModelArts Studio大模型开发平台可以辅助用户进行提示词撰写、比较和评估等操作，并对提示词进行保存和管理。

表 16-1 功能说明

功能	说明
提示词工程任务管理	提示词工程平台以提示词工程任务为管理维度，一个任务代表一个场景或一个调优需求，在提示词工程任务下可以进行提示词的调优、比较和评估。 提示词工程任务管理支持工程任务的创建、查询、修改、删除。
提示词撰写	提示词调优支持对提示词文本的编辑、提示词变量设置、提示词结果生成和调优历史记录管理。
提示词候选	提示词候选支持用户对调优后初步筛选的提示词进行候选管理，每个工程任务下可以保存上限9个候选提示词，进一步基于候选提示词进行比较和评估。
提示词比较	提示词比较支持选择两个候选提示词对其文本和参数进行比较，支持对选择的候选提示词设置相同变量值查看效果。
提示词评估	提示词评估以任务维度管理，支持评估任务的创建、查询、修改、删除。支持创建评估任务，选择候选提示词和需要使用的变量数据集，设置评估算法，执行任务自动化对候选提示词生成结果和结果评估。
提示词管理	提示词管理支持用户对满意的候选提示词进行保存管理，同时支持提示词的查询、删除。

图 16-1 提示词工程使用流程



版权所有 © 华为云计算技术有限公司

16.2 获取提示词模板

平台提供了多种任务场景的提示词模板，可以帮助用户更好地利用大模型的能力，引导模型生成更准确、更有针对性的输出，从而提高模型在特定任务上的性能。

在创建提示词工程前，可以先使用预置的提示词模板，或基于提示词模板进行改造 。如果提示词模板满足不了使用需求可再单独创建。

提示词模板可在平台“Agent 开发 > 提示词工程 > 提示词模板”中获取。

16.3 撰写提示词

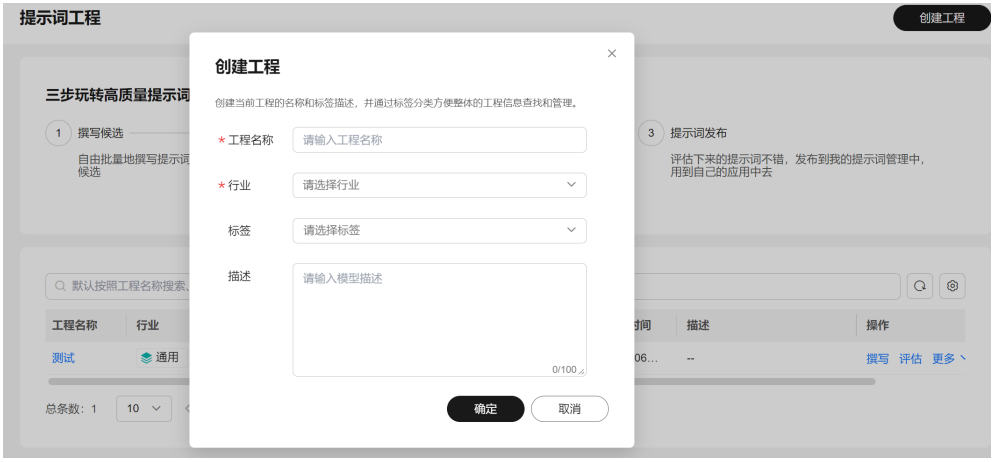
16.3.1 创建提示词工程

通过精心设计和优化提示词，可以引导大模型生成用户期望的输出。提示词工程任务的目标是通过设计和实施一系列的实验，来探索如何利用提示词来提高大模型在各种任务上的表现。

撰写提示词前需要先创建提示词工程，用于对提示词进行统一管理。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”，单击界面右上角“创建工程”。
3. 输入工程名称、描述，选择行业、标签后。单击“确定”完成工程创建。

图 16-2 创建提示词工程



16.3.2 撰写提示词

提示词是用来引导模型生成的一段文本。撰写的提示词应该包含任务或领域的关键信息，如主题、风格、格式等。

撰写提示词时，可以设置提示词变量。即在提示词中通过添加占位符{{ }}标识表示一些动态的信息，让模型根据不同的情况生成不同的文本，增加模型的灵活性和适应性。例如，将提示词设置为“你是一个旅游助手，需要给用户介绍旅行地的风土人情。请介绍下{{location}}的风土人情。”在评估提示词效果时，可以通过批量替换{{location}}的值，来获得模型回答，提升评测效率。

版权所有 © 华为云计算技术有限公司

同时，撰写提示词过程中，可以通过设置模型参数来控制模型的生成行为，如调整温度、核采样、最大Token限制等参数。模型参数的设置会影响模型的生成质量和多样性，因此需要根据不同的场景进行选择。

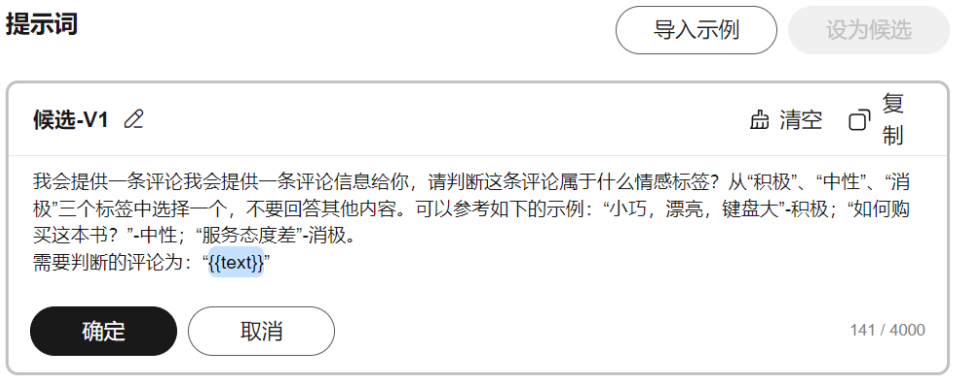
1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 16-3 提示词工程



4. 在提示词撰写区域输入提示词文本，可以插入若干个变量，变量需要使用占位符{{ }}标识。

图 16-4 撰写提示词



5. 撰写完成后，单击“确定”，平台会自动识别插入的变量。提示词中识别的变量将展示在变量定义区域。

   变量名称可以进行修改，如添加备注信息以便更好理解变量的作用。

版权所有 © 华为云计算技术有限公司

图 16-5 变量定义



说明

变量定义区域展示的是整个工程任务下定义的变量信息，候选提示词中关联的变量也会进行展示，候选提示词相关操作请参见**设置候选提示词**。

同一个提示词工程中，定义的变量不能超过20个。

6. 在“模型”区域，单击“设置”，设置提示词输入的模型和模型参数。

图 16-6 模型设置



16.3.3 预览提示词效果

提示词撰写完成后，可以通过输入具体的变量值，组成完整的提示词，查看不同提示词在模型中的使用效果。

1. 在撰写提示词页面，找到页面右侧变量输入区域，在输入框中输入具体的变量值信息。

   输入变量值后预览区域会自动组装展示提示词。也可以直接选择已创建的变量集填入变量值信息，变量集是一个excel文件，每行数据是需要输入的变量值信息，可以通过“导入”功能进行上传。

版权所有 © 华为云计算技术有限公司

图 16-7 效果预览



2. 单击“查看效果”，输出模型回复结果，用户可以基于预览的效果调整提示词文本和变量。

16.4 横向比较提示词效果

16.4.1 设置候选提示词

用户可以将效果较好的提示词设为候选提示词，并对提示词进行比对，以查看其效果。

说明

每个工程任务下候选提示词上限9个，达到上限9个时需要删除其他候选提示词才能继续添加。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

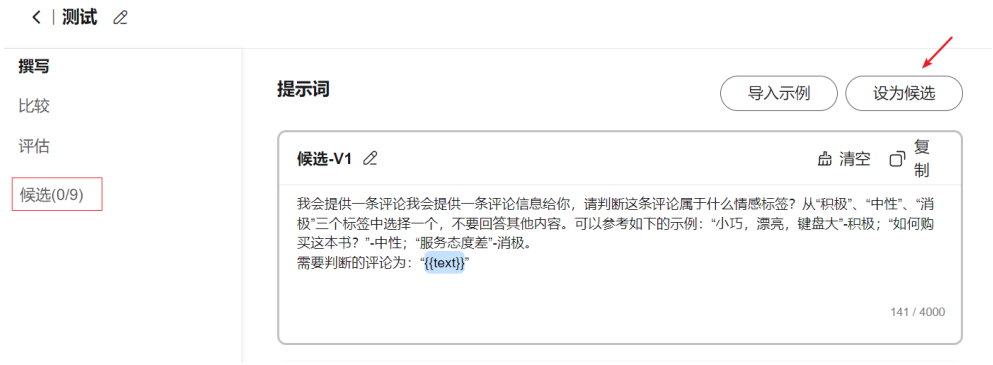
图 16-8 提示词工程



 版权所有 © 华为云计算技术有限公司

4. 在提示词撰写区域，单击“设为候选”，将当前撰写的提示词设置为候选提示词。
   候选状态的提示词将保存至左侧导航栏的“候选”中。

图 16-9 设为候选



16.4.2 横向比较提示词效果

将设置为候选的提示词横向比对，获取提示词的差异性和效果。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 16-10 提示词工程



4. 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，勾选需要进行横向比对的提示词，并单击“横向比较”。

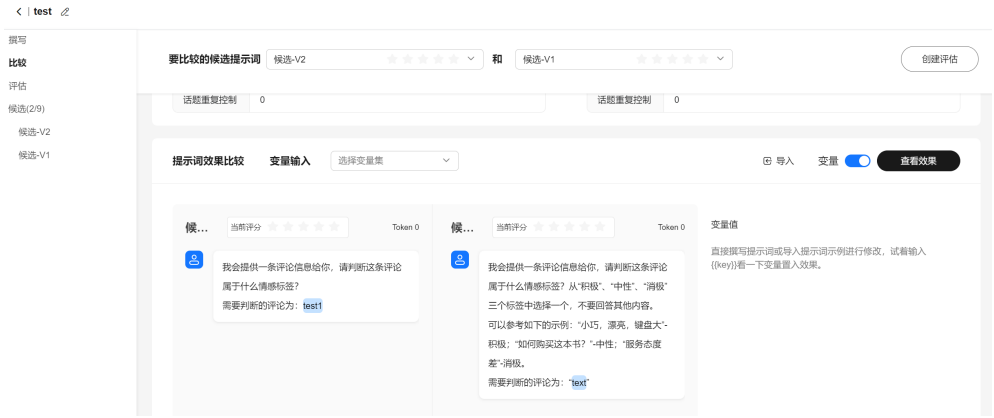
 版权所有 © 华为云计算技术有限公司

图 16-11 横向比较



5. 进入到横向比较页面，下拉页面至“提示词效果比较”模块，比较提示词的效果，输入相同的变量值，查看两个提示词生成的结果。

图 16-12 横向比对提示词效果



16.5 批量评估提示词效果

16.5.1 创建提示词评估数据集

批量评估提示词效果前，需要先上传提示词变量数据文件用于创建对应的评估数据集。

提示词变量是一种可以在文本生成中动态替换的占位符，用于根据不同的场景或用户输入生成不同的内容。其中，变量名称可以是任意的文字，用于描述变量的含义或作用。

提示词评估数据集约束限制

- 上传文件限xlsx格式。
- 数据行数不小于10行，不大于50行。
- 数据不允许相同表头，表头数量小于20个。
- 数据单条文本长度不超过1000。

说明

创建数据集时会对相关限制条件进行校验。

数据参考格式如下：

 版权所有 © 华为云计算技术有限公司

图 16-13 数据参考格式

	A	B	C	D	E	F	G
1	key1	key2	key3	result	→ 表头为提示词变量key		
2	组1-v1	组1-v2	组1-v3	组1-r			
3	组2-v1	组2-v2	组2-v3	组2-r	→ 每行为1组变量值		
4	组3-v1	组3-v2	组3-v3	组3-r			
5							

图 16-14 数据示例

comment → 变量key	result	
字打错了怎么修改？	中性	
这个效果不太好	消极	
我喜欢这个样式	积极	
请问你是有什么心事吗？	中性	
↘ 评论	↘ 预期结果	

创建提示词评估数据集

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示用例管理”，单击页面右上角“创建提示用例集”。

图 16-15 提示用例管理



3. 在“创建数据集”页面完成数据集的上传。

版权所有 © 华为云计算技术有限公司

图 16-16 创建提示词评估数据集



16.5.2 创建提示词评估任务

选择候选提示词进行批量自动化评估，步骤如下：。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。
4. 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，勾选需要进行横向比对的提示词，并单击“创建评估”。

图 16-17 创建评估



5. 选择评估使用的变量数据集和评估方法。
   - 评估用例集：根据选择的数据集，将待评估的提示词和数据集中的变量自动组装成完整的提示词，输入模型生成结果。
   - 评估方法：根据选择的评估方法，对模型生成结果和预期结果进行比较，并根据算法给出相应的得分。

版权所有 © 华为云计算技术有限公司

图 16-18 创建提示词评估任务



6. 单击“确定”，评估任务自动进入执行状态。

16.5.3 查看提示词评估结果

1. 评估任务创建完成后，会跳转至“评估”页面，在该页面可以查看评估状态。

图 16-19 查看提示词评任务状态



2. 单击“评估名称”，进入评估任务详情页，可以查看详细的评估进度，例如在图16-20中有10条评估用例，当前已评估8条，剩余2条待评估。

图 16-20 查看评估进展



3. 评估完成后，可以查看每条数据的评估结果。

   在评估结果中，“预期结果”表示变量值（问题）所预设的期望回答，“生成结果”表示模型回复的结果。通过比对“预期结果”、“生成结果”的差异可以判断提示词效果。

版权所有 © 华为云计算技术有限公司

16.6 发布提示词

通过**横向比较提示词效果**和**批量评估提示词效果**，如果找到高质量的提示词，可以将这些提示词发布至“提示词模板”中。

1. 在提示词“候选”页面，选择质量好的提示词，并单击“保存到模板库”。

图 16-21 保存提示词至模板库



2. 进入“Agent 开发 > 提示词工程 > 提示词模板”页面，查看发布的提示词。

版权所有 © 华为云计算技术有限公司

17 开发盘古大模型垂域应用

工业应用编排介绍

编排工业应用

17.1 工业应用编排介绍

工业应用编排功能是一个基于脚本组件和算法包组件的编排工具。其中，组件是用来构成应用编排的最小业务单元，即将封装好的业务功能作为应用编排中的一个节点，实现特定的业务功能。

例如，需要构建一个视频抽帧的应用，该应用需要具备读取视频数据、视频抽帧、抽帧后图片处理、输出图片数据的功能。那么可以将各个处理模块的代码脚本封装为组件，通过工业应用编排功能将各组件编排成一个完整的工作流来实现视频抽帧。除了将代码脚本封装为组件外，也可以将算法封装为组件。例如，将一个图片分类算法封装为组件，那么在视频抽帧完成后，可以叠加一个图片分类算法对图片做进一步的处理。

17.2 编排工业应用

17.2.1 创建应用组件

17.2.1.1 预置组件功能介绍

平台提供了多种预置组件方便用户使用。

表 17-1 预置组件清单

组件类别	组件名称	功能介绍
数据读写	convert_obs_file_to_base64	**convert_obs_file_to_base64组件**
	save_data_to_local_file	**save_data_to_local_file组件**

版权所有 © 华为云计算技术有限公司

组件类别	组件名称	功能介绍
	read_data_from_local_file	read_data_from_local_file组件
逻辑组件	deploy_together	deploy_together组件
自定义脚本	custom_python_script	custom_python_script组件
其他	call_stream_api	call_stream_api组件
	call_api	call_api组件

convert_obs_file_to_base64 组件

convert_obs_file_to_base64组件用于读取obs上的文件，并将其转换为base64编码，适用于类似图片的小文件base64读取转换。

表 17-2 convert_obs_file_to_base64 组件参数说明

参数分类	参数名称	参数说明
输入参数	obs_path	待读取的obs文件路径。以“obs://”开头
	ak	obs文件所在租户的访问密钥（ak）。
	sk	obs文件所在租户的密钥（sk）。
输出参数	output	文件转化为base64编码后的输出结果。

其中，ak、sk获取方法如下：

登录华为云管理控制台，鼠标移动至页面右上角用户名位置，单击“我的凭证”，在“我的凭证”页面，进入“访问密钥”页签获取ak、sk。如果没有ak、sk可以通过单击“新增访问密钥”进行创建。

图 17-1 convert_obs_file_to_base64 组件



 版权所有 © 华为云计算技术有限公司

save_data_to_local_file 组件

save_data_to_local_file组件用于将单层级的json格式数据存储到本地的csv或json文件中。

表 17-3 save_data_to_local_file 组件参数说明

参数分类	参数名称	参数说明
输入参数	file_path	本地文件的路径，支持csv和json两种文件格式。例如：/home/HwHiAiUser/test.csv
	data	要写入文件的json数据。
	encoding	文件内容编码。支持gbk和utf-8两种格式， 默认值gbk。 file_path文件为csv格式时，建议使用gbk编码。
输出参数	flag	写入json数据后文件的长度。

图 17-2 save_data_to_local_file 组件



read_data_from_local_file 组件

read_data_from_local_file组件可以用于读取本地数据，组件参数说明如下：

表 17-4 read_data_from_local_file 组件参数说明

参数分类	参数名称	参数说明
输入参数	file_path	要读取的本地文件路径，支持csv和json两种文件格式。例如：/home/HwHiAiUser/test.csv
	encoding	文件内容的编码方式。支持gbk和utf-8两种格式。
输出参数	data	从文件中读取到的json数据

版权所有 © 华为云计算技术有限公司

图 17-3 read_data_from_local_file 组件



deploy_together 组件

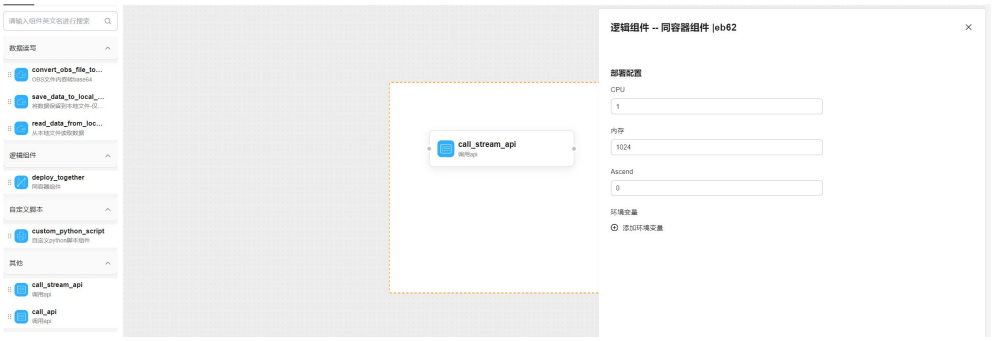
deploy_together组件中文名称为同容器组件，所有拖入同容器组件中的子组件会被整合为一个容器服务，从而简化管理和部署流程。这些子组件共享同一个容器的资源，如CPU、内存和环境变量。这些子组件在同一容器中运行，减少了跨容器通信的开销和复杂性，便于统一配置和管理。

同容器组件参数说明如下：

表 17-5 deploy_together 组件参数说明

参数分类	参数名称	参数说明
部署配置	CPU	部署需要使用的最小CPU值（物理核）。
	内存	部署需要使用的最小内存值。
	Ascend	部署使用的npu卡数量。
	环境变量	用户在组件代码中定义的外部环境变量。

图 17-4 deploy_together 组件部署配置



版权所有 © 华为云计算技术有限公司

custom_python_script 组件

custom_python_script组件与组件管理中的自定义脚本组件功能基本相同，支持用户自己编写Python代码完成特定组件功能。不同之处在于，custom_python_script组件支持在画布中进行参数的创建与删除，脚本代码的编写，调试过程中有问题无需返回组件管理页面进行处理，大大方便用户的调试过程，可以在调试完成后归档在组件管理页面中。需要注意的是，每次在画布中拖入custom_python_script组件时，均需创建其参数和补充脚本，删除时，之前创建的信息会丢失。

图 17-5 custom_python_script 组件配置



call_stream_api 组件

call_stream_api组件可以用于调用已经部署好的流式出参的模型服务。例如，nlp模型流式场景的组件参数说明如下：

表 17-6 call_stream_api 组件参数说明

参数分类	参数名称	参数说明
输入参数	body	调用服务请求所需要的请求体，为对象格式，默认为{}。
	headers	调用服务请求所需要的请求头，为对象格式，默认为{}。
	url	调用服务的请求地址，为字符串格式，画布中可通过“选择服务”按钮进行自动填充。
	request_type	调用服务的请求类型，默认为POST。
输出参数	response	调用服务的流式返回值。

版权所有 © 华为云计算技术有限公司

图 17-6 call_stream_api 组件



call_api 组件

call_api组件可以用于调用已经部署好的五大模型服务（非流式出参场景），组件参数说明如下：

表 17-7 call_api 组件参数说明

参数分类	参数名称	参数说明
输入参数	body	调用服务请求所需要的请求体，为对象格式，默认为{}。
	url	调用服务的请求地址，为字符串格式，画布中可通过“选择服务”按钮自动填充，实现调用已部署的算法包组件。
	headers	调用服务请求所需要的请求头，为对象格式，默认为{}。
	request_type	调用服务的请求类型，默认为POST。
输出参数	response	调用服务的返回值。

版权所有 © 华为云计算技术有限公司

图 17-7 call_api 组件



17.2.1.2 创建脚本组件

脚本组件使用场景

脚本组件适用于简单的代码处理任务，特别是可以通过单个代码文件解决的问题。例如，将读取本地数据功能设置为脚本组件。

脚本组件仅支持Python脚本，脚本组件可以将代码进行整合，使得代码更加模块化，同时也提高了代码的可读性和可维护性。

脚本组件使用规范

- 脚本语言：脚本组件仅支持Python脚本。
- 类定义：脚本组件必须定义一个名为Template的类。这个类名是固定的，不可修改。
- 方法定义：Template类中必须实现两个方法：__init__和process。这两个方法的名称是固定的，不可修改。
- 方法参数：__init__和process方法的参数不可添加。__init__方法没有参数，process方法有一个参数params_dict。
- 类初始化：__init__方法在脚本组件的生命周期中只会被执行一次。这个方法通常用于进行一些初始化操作。
- 数据处理：process方法是脚本组件的核心，它负责处理数据。这个方法可以从params_dict参数中获取输入数据。params_dict是一个字典，它包含了所有的输入参数，如示例注释中，a和b对应组件的两个输入参数·。
- 返回值：process方法的返回值必须是一个字典。这个字典包含了处理结果，对应组件的所有输出参数，如示例注释中，a对应该组件的一个输出参数。
- 编码声明：脚本组件的文件必须在文件头部声明编码方式，通常为# -*- coding: utf-8 -*-。
- 注释：脚本组件的代码应该包含足够的注释，以帮助其他人理解代码的功能和工作原理。

脚本组件示例：

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
```

 版权所有 © 华为云计算技术有限公司

```
class Template:
  def __init__(self):
      # todo
      pass
   def process(self, params_dict):
     # 获取入参
     # a = params_dict.get("a")
     # b = params_dict.get("b")
     # todo
     # return {"a": a}
     return None
```

创建脚本组件

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“垂域应用开发 > 应用组件 > 组件管理”，在“自定义”页签，单击右上角“创建组件”，选择“脚本组件”算法类型，单击“确定”。

图 17-8 脚本组件

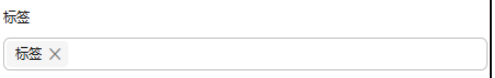


3. 进入“创建脚本组件”页面，参考表17-8设置基础信息。

表 17-8 创建脚本组件基础信息参数说明

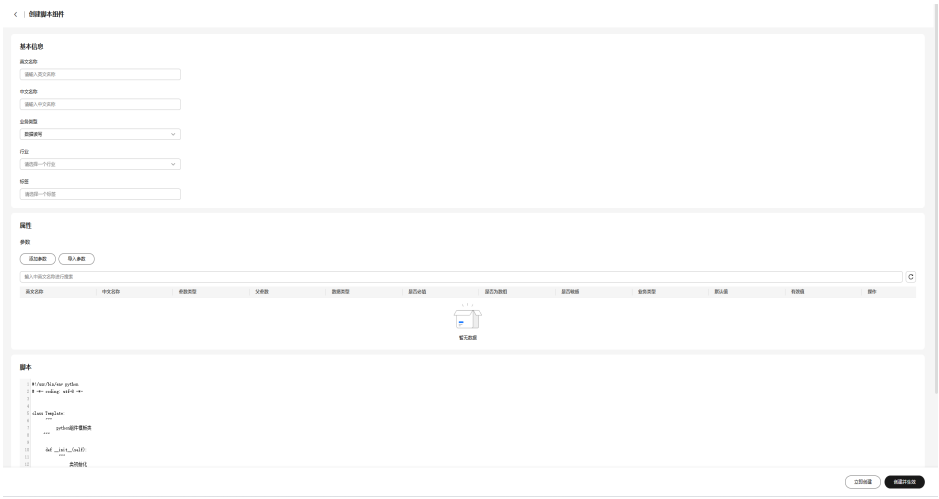
参数名称	说明
英文名称	组件的英文名称。
中文名称	组件的中文名称。
业务类型	用于区分脚本组件的业务场景。包括以下类型： 数据读写、数据预处理、数据后处理、特征加工、优化场景定义、机理模型、过程控制、图像组件、视频组件、其他。

版权所有 © 华为云计算技术有限公司

参数名称	说明
行业	用于区分脚本组件的行业类型，按行业类型进行组件筛选，非必填项，包括以下行业： 汽车、政务、交通、电力、钢铁、矿山、油气、供热。
标签	• 用于给组件绑定标签，支持绑定已有或者自定义的新标签，每个组件最多可绑定10个标签。 • 标签输入框支持创建新的标签和删除已有标签，单个空间最多可创建200个标签。 – 创建标签：输入框中输入标签名称并回车即可创建该标签。 **图 17-9** 创建标签  – 删除标签：下拉框单击指定标签的“X”号，单击“确认”，即可删除标签。 **图 17-10** 删除标签  **注意** 如果已有组件绑定该标签，删除操作会同步删除该组件与标签的绑定关系，且删除后不可找回，请谨慎操作。

 版权所有 © 华为云计算技术有限公司

图 17-11 创建脚本组件



4. 基础信息设置完成后，参考**表17-9**设置属性参数。

   属性参数具备多种功能。例如，在编排静态应用时，通常需要组合多个组件来实现某一功能，不同的组件需要定义清楚输入、输出参数的引用关系，实现数据在不同的组件间流转。

表 17-9 创建属性参数说明

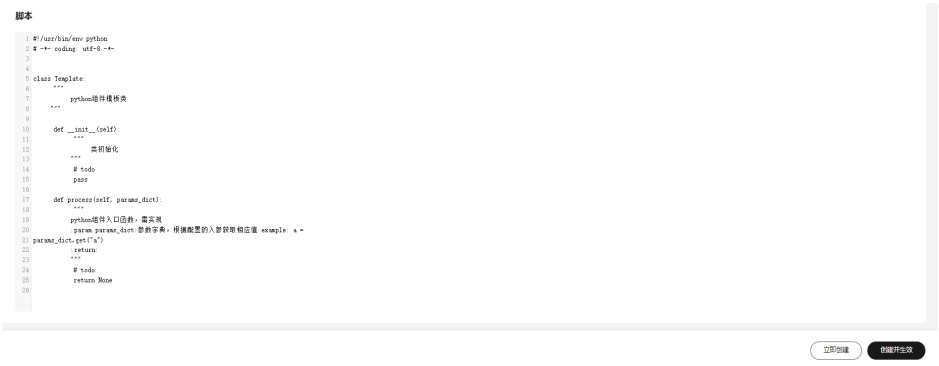
参数名称	说明
英文名称	参数的英文名称。
中文名称	参数的中文名称。
参数类型	• 入参：在编排应用时，该节点的入参，支持三种方式，包括：界面填写、其他组件节点的输出参数作为当前节点的入参、画布参数传入。 • 出参：表示执行组件后生成的结果，出参可以被其他组件节点使用。 • 流式出参：nlp等场景下支持流式出参，支持出参以流式方式展示，使用时组件应该仅包含一个流式出参，并且不包含其他出参。 • 环境变量：在组件运行时可用的全局变量，可以用环境变量存储一些共享的配置信息。例如应用需要从数据库中获取待处理的数据，可以用环境变量存储数据库的位置信息、账密信息、API密钥等。
数据类型	组件参数的数值类型。
业务类型	参数的业务类型。非必填参数，默认为空，支持模型和模型服务两种类型。 模型：用于在静态应用界通过“选择模型”，自动填充所选模型的obs地址； 模型服务：用于在静态应用界通过“选择服务”，自动填充所选模型服务的访问地址，实现组件服务调用所选模型服务。

 版权所有 © 华为云计算技术有限公司

参数名称	说明
父参数	父参数通常用于配置复杂的结构体，可以使用一个已经添加的数据参数作为这个结构的父参数，即父参数必须是一个已经定义的组件参数。 配置复杂结构体使用，父参数只能是已经添加的数据类型为对象的组件参数。
默认值	参数的默认值。
有效值	参数可以接受的合法取值范围。举例：整数：[0,10]，浮点数：[0.1,10.1]，枚举：A|B|C。
是否必填	该参数是否为必填参数。
是否敏感	设置参数是否为敏感参数，敏感参数将不允许在画布中填入，只能通过画布参数传入。
是否为数组	该参数是否为数组类型。如果一个参数是数组类型，意味着它可以包含多个值。例如，一个名为“标签”的参数可以是一个字符串数组，其中包含多个标签名称。如果参数不是数组类型，那么它只能接受单个值。

5. 填完参数无误后，填写脚本的内容，完成后单击“立即创建”，组件创建成功。

图 17-12 修改脚本



6. 组件创建完成后，在脚本列表页面，单击操作列“编辑”可以对中文名称、参数、脚本等进行修改更新。其中，组件参数有批量导入的功能，如果一个组件参数比较多，并且在其他组件中已经创建，可以复制参数代码，在需要导入的组件参数中通过“导入参数”功能填入代码，进行参数的批量导入。

 版权所有 © 华为云计算技术有限公司

图 17-13 查看组件参数

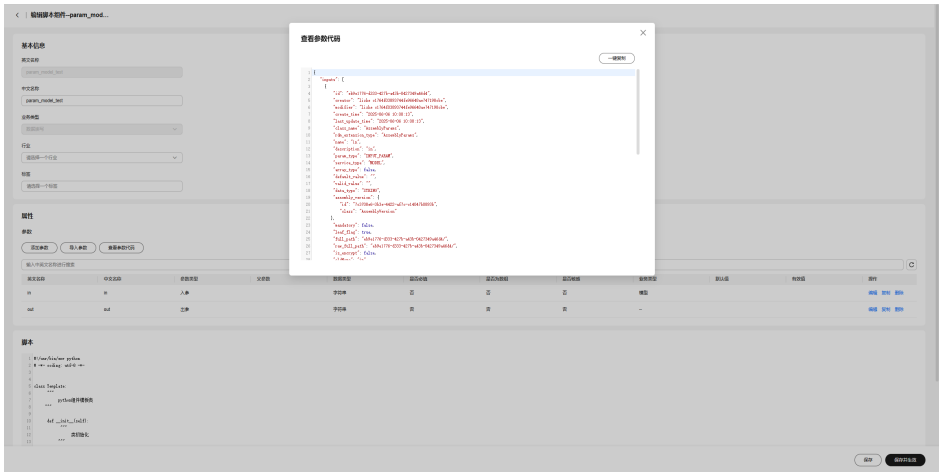
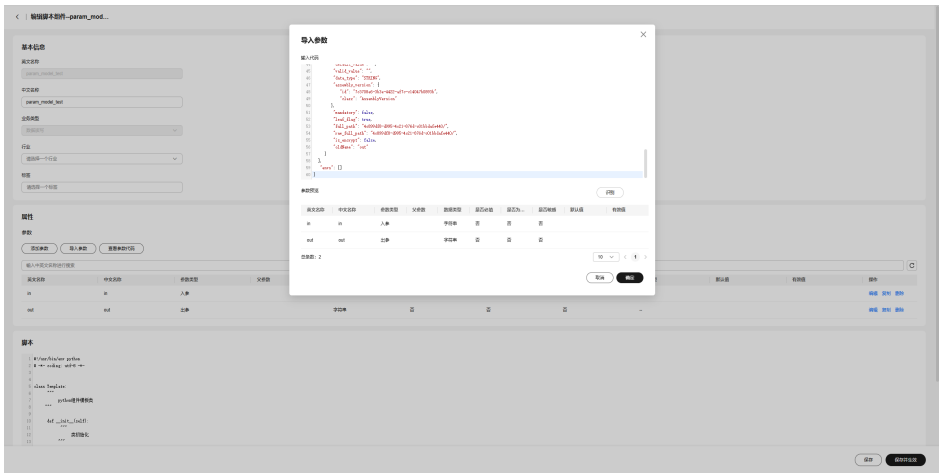
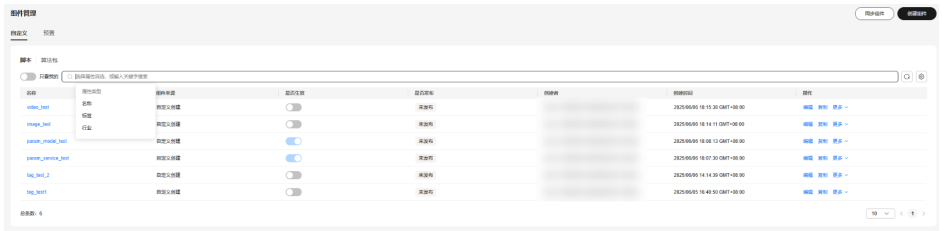


图 17-14 导入参数



7. 组件开发完成后，单击“保存并生效”后，该组件可用于**创建静态应用**中使用。
8. 在组件列表页面，可基于名称、标签、行业对组件进行筛选。

图 17-15 组件列表



17.2.1.3 创建算法包组件

算法包组件使用场景

算法包组件适合处理复杂的场景，如承载算法模型的任务。在使用时，算法包组件将被配置为一个独立的服务，以供各组件进行访问。

 版权所有 © 华为云计算技术有限公司

算法包组件使用规范

- 打包格式：算法包需要打包成tar包或tar.gz格式，不支持其他格式。解压后的目录结构不应包含第一层文件夹，应为以下结构：

  软件包目录结构：

  ```
  - main.py
  - dependency/
  - lib/                (可选：C/C++运行文件)
  - requirements.txt
  - ...
  ```

- main.py文件：main.py文件为必选项，其中需要实现一个名为AlgoProcessor的类。这个类中的__init__函数可以实现模型初始化等操作，process函数则实现调用方法。包中引用代码应遵循标准，采用绝对路径。如软件包名为A，内部各模块引入时应为from A.xxx import xxx。

  示例代码：

  ```
  #!/usr/bin/env python
  # -*- coding: utf-8 -*-
  from rest.process_base import ProcessBase
  class AlgoProcessor(ProcessBase):
      """
      Algorithm Processor
      """
      def __init__(self):
      super(AlgoProcessor, self).__init__(url="")
      def process(self, req):
      rst = {'result': "success", 'suggestion': "pass"}
      return rst, 200
  ```

- dependency文件夹：dependency文件夹为必选项，用于存放基础镜像中未包含或版本不一致的依赖包，可以是自己打的依赖包，也可以是开源依赖包。这些依赖包将配合requirements.txt一起安装。

  示例代码：

  ```
  - simon-1.0.8-py3-none-any.whl
  ```

- lib文件夹：lib文件夹为可选项，用于存放在Linux环境下编译的C/C++代码的****.so动态库，以及其他依赖包。

  示例代码：

  ```
  - Width_Spread.so
  ```

- requirements.txt文件：requirements.txt文件为必选项，用于列出dependency文件夹中包含的依赖包。

  示例代码：

  ```
  dependency/simon-1.0.8-py3-none-any.whl
  ```

- 其他代码文件：其他代码文件为可选项，可以自行组织存放并正确引用。

创建算法包组件

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“垂域应用开发 > 应用组件 > 组件管理”，在“自定义”页签，单击右上角“创建组件”，选择“算法组件包”算法类型，单击“确定”。

 版权所有 © 华为云计算技术有限公司

图 17-16 算法包组件



3. 进入“创建算法包组件”页，参考表17-10设置基础信息。

表 17-10 创建算法包组件参数说明

参数名称	说明
英文名称	组件的英文名称。
中文名称	组件的中文名称。
业务类型	用于区分算法包组件的业务场景。包括以下类型： 数据读写、数据预处理、数据后处理、特征加工、优化场景定义、机理模型、过程控制、图像组件、视频组件、其他。
行业	用于区分算法包组件的行业类型，按行业类型进行组件筛选，非必填项，包括以下行业： 汽车、政务、交通、电力、钢铁、矿山、油气、供热。

版权所有 © 华为云计算技术有限公司

参数名称	说明
标签	• 用于给组件绑定标签，支持绑定已有或者自定义的新标签，每个组件最多可绑定10个标签。 • 标签输入框支持创建新的标签和删除已有标签，单个空间最多可创建200个标签。 – 创建标签：输入框中输入标签名称并回车即可创建该标签。 **图 17-17** 创建标签  – 删除标签：下拉框单击指定标签的“X”号，单击“确认”，即可删除标签。 **图 17-18** 删除标签  **注意** 如果已有组件绑定该标签，删除操作会同步删除该组件与标签的绑定关系，且删除后不可找回，请谨慎操作。
架构类型	算法支持的硬件架构类型： • ALL：算法支持x86和arm两种架构。 • X86：算法仅支持x86架构。 • ARM：算法仅支持arm架构。
卡类型	算法包组件在进行部署时，支持的卡类型。
是否区分架构	算法包是否区分架构。
算法包	上传算法包。

 版权所有 © 华为云计算技术有限公司

图 17-19 创建算法包组件



4. 基础信息设置完成后，参考**表17-11**设置属性参数。

   属性参数具备多种功能。例如，在编排静态应用时，通常需要组合多个组件来实现某一功能，不同的组件需要定义清楚输入、输出参数的引用关系，实现数据在不同的组件间流转。

表 17-11 创建属性参数说明

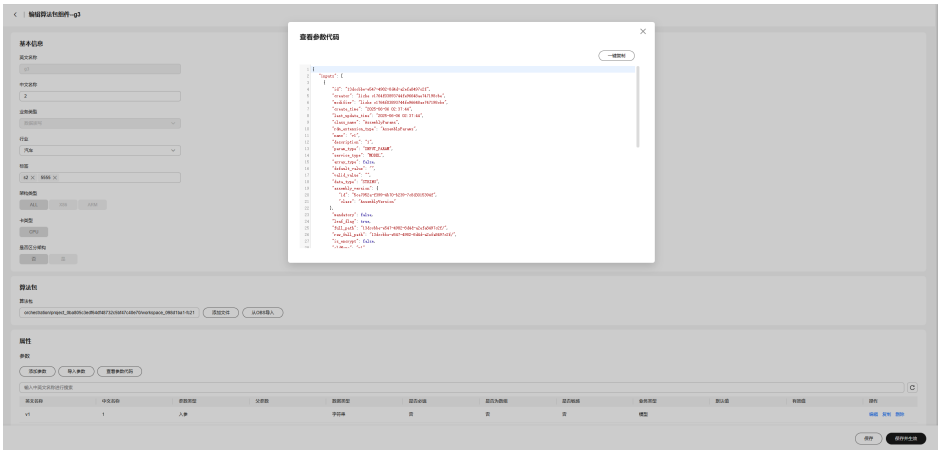
参数名称	说明
英文名称	参数的英文名称。
中文名称	参数的中文名称。
参数类型	• 入参：在编排应用时，该节点的入参，支持三种方式，包括：界面填写、其他组件节点的输出参数作为当前节点的入参、画布参数传入。 • 出参：表示执行组件后生成的结果，出参可以被其他组件节点使用。 • 流式出参：nlp等场景下支持流式出参，支持出参以流式方式展示，使用时组件应该仅包含一个流式出参，并且不包含其他出参。 • 环境变量：在组件运行时可用的全局变量，可以用环境变量存储一些共享的配置信息。例如应用需要从数据库中获取待处理的数据，可以用环境变量存储数据库的位置信息、账密信息、API密钥等。
数据类型	组件参数的数值类型。
业务类型	参数的业务类型。非必填参数，默认为空，支持模型和模型服务两种类型。 模型：用于在静态应用界通过“选择模型”，自动填充所选模型的obs地址； 模型服务：用于在静态应用界通过“选择服务”，自动填充所选模型服务的访问地址，实现组件服务调用所选模型服务。

 版权所有 © 华为云计算技术有限公司

参数名称	说明
父参数	父参数通常用于配置复杂的结构体，可以使用一个已经添加的数据参数作为这个结构的父参数，即父参数必须是一个已经定义的组件参数。 配置复杂结构体使用，父参数只能是已经添加的数据类型为对象的组件参数。
默认值	参数的默认值。
有效值	参数可以接受的合法取值范围。举例：整数：[0,10]，浮点数：[0.1,10.1]，枚举：A|B|C。
是否必填	该参数是否为必填参数。
是否敏感	设置参数是否为敏感参数，敏感参数将不允许在画布中填入，只能通过画布参数传入。
是否为数组	该参数是否为数组类型。如果一个参数是数组类型，意味着它可以包含多个值。例如，一个名为“标签”的参数可以是一个字符串数组，其中包含多个标签名称。如果参数不是数组类型，那么它只能接受单个值。

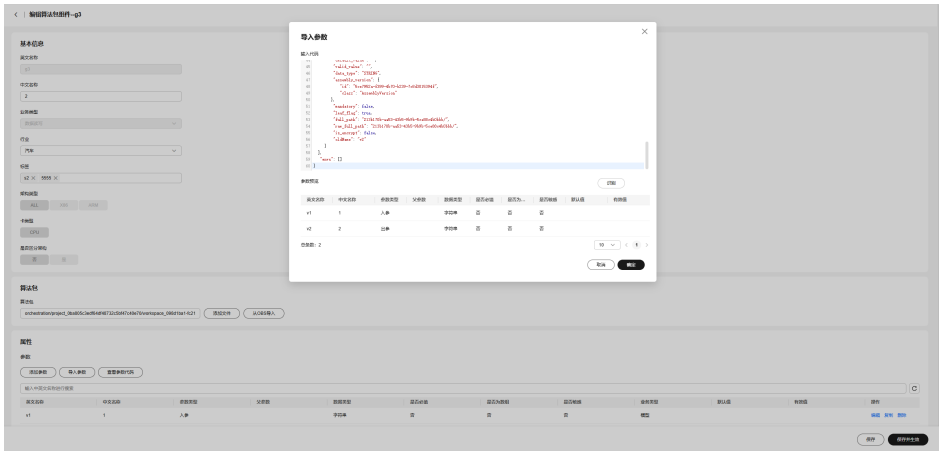
5. 参数填写完成后，单击“立即创建”。
6. 组件创建完成后，在算法包列表页面，单击操作列“编辑”可以对中文名称、参数、算法包等进行修改更新。其中，组件参数支持批量导入功能。如果一个组件参数比较多，并且这些参数在其他组件中已经创建，可以通过复制参数代码，并在需要导入参数的组件中通过“导入参数”功能填入代码，从而实现参数的批量导入。

图 17-20 查看参数代码



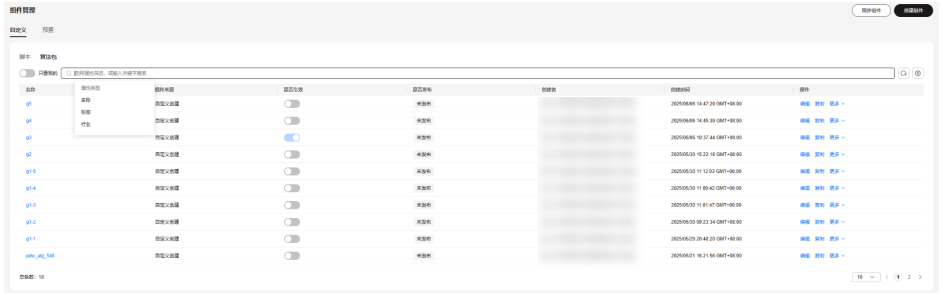
 版权所有 © 华为云计算技术有限公司

图 17-21 导入参数



7. 组件开发完成后，单击“保存并生效”后，该组件可用于**部署算法包组件**和**创建静态应用**中使用。
8. 在组件列表页面，可基于名称、标签、行业对组件进行筛选。

图 17-22 组件列表

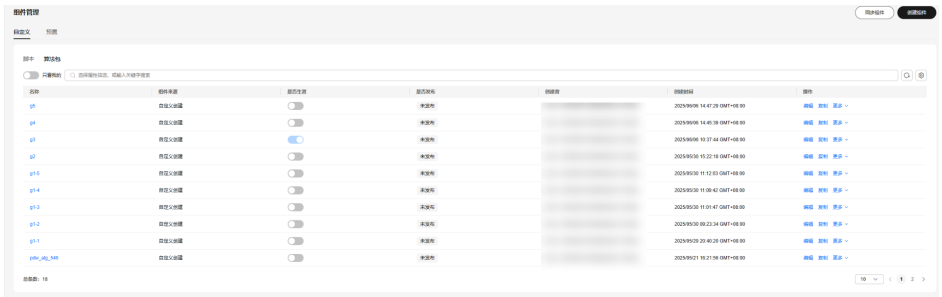


17.2.1.4 部署算法包组件

运行算法包组件依赖于CPU、NPU这类计算资源。如果应用中有使用算法包组件，需要先执行部署操作。部署后算法包就能在一个独立、隔离的环境中运行，同时也能被其他组件访问。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“垂域应用开发 > 应用组件 > 组件管理 > 自定义”，进入“算法包”页签。

图 17-23 算法包列表



版权所有 © 华为云计算技术有限公司

3. 选择“已生效”的算法包，单击操作栏的“部署”，或者进入算法包组件详情页面，单击右上角“部署”，部署算法。“未生效”的算法包组件单击列表其所在行的“是否生效”开关进行组件生效。

图 17-24 部署组件

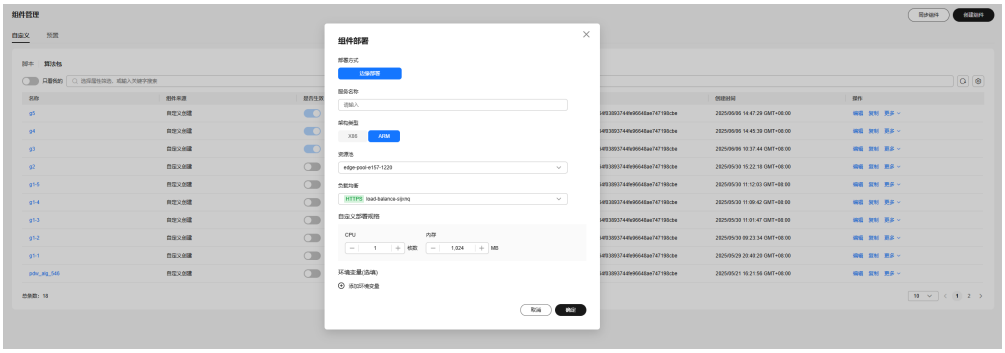


表 17-12 部署参数说明

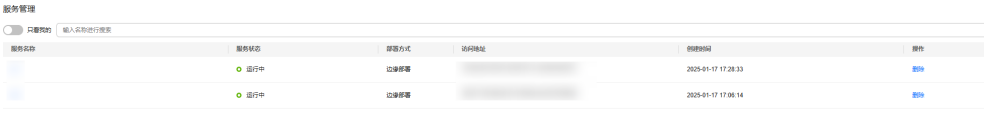
参数名称	说明
部署方式	服务的部署类型，支持边缘部署。
服务名称	部署服务的名称。
架构类型	部署的资源池架构类型，与实际算法的架构相关。
资源池	选择部署的边缘资源池。若需要使用Ascend卡，当前支持推理卡类型为Snt3p。
负载均衡	边缘部署使用的负载均衡，支持HTTPS和HTTP两种协议的负载均衡，用于后续通过负载均衡调用的部署的服务。
CPU	部署需要使用的最小CPU值（物理核），默认1Core。越复杂的模型通常需要更多的计算资源，可以考虑增加CPU的数量；如果服务需要处理大量的并发请求，建议增加CPU数量。但不能超过资源池中的可用CPU数。
内存	部署需要使用的最小内存值，默认1024MB。模型越大、处理数据越多，需要的内存也就越大。可以依据实际模型情况和待处理的数据量情况进行调整。
Ascend	部署使用的npu卡数量。Ascend设备通常用于深度学习模型的加速，当模型支持在Ascend上运行，且资源池中包含Ascend资源时，才可以进行配置，确保分配的Ascend数量在资源池范围内。当前支持推理卡类型为Snt3p，单卡算力为70T FLOPS@fp16，加载不同推理模型所消耗的算力资源不同，需要卡数的计算公式为：卡数量≈总算力/卡算力。计算得到的卡数量可以向上取整后作为参数取值。 若资源池进行过算力切分，此处的卡数量表示的是逻辑卡数。若资源池没有经过算力切分，则表示为物理卡数。例如，物理卡切分成了7份，那一个逻辑卡的算力就是10T FLOPS@fp16。
环境变量	用户在组件代码中定义的外部环境变量。

版权所有 © 华为云计算技术有限公司

如果使用视频解码、视频抽帧类算法，CPU和内容推荐设置为4CPU，8092M内存，若算法需要使用Ascend设备运行，可将数量设置为1。CPU、内存、Ascend取值大小与用户所使用的算法相关、处理的数据量相关，请按算法实际运行所依赖的资源要求进行设置。

4. 部署完成后，单击算法包名称，进入详情页后，可以在最下方的服务管理处看到服务列表信息。

图 17-25 服务管理



5. 服务部署完成后，可以根据服务列表中的访问地址完成服务的调用请求，请求体结构按照创建算法包组件时设置的结构进行填写。

   边缘部署服务的访问URL格式为：*<所使用负载均衡的协议>://<负载均衡的IP(通常为资源池主节点IP)>:<负载均衡的端口>/<服务的访问地址>*

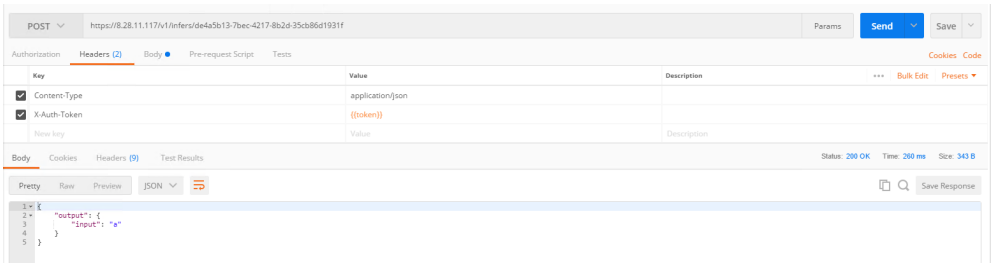
   例如：部署使用的负载均衡为HTTPS的 ，负载均衡的IP为127.0.0.1，负载均衡的端口为：8080，服务的访问地址为：05806e52-6423-43bc-b1be-82d4fa1a158a，

   则最终的访问地址为：https://127.0.0.1:8080/05806e52-6423-43bc-b1be-82d4fa1a158a

   调用访问地址时需要填写请求Header参数：

   Content-Type：填写application/json。

图 17-26 服务调用



6. 单击服务名称可以跳转到当前服务详情页面。

图 17-27 服务详情



7. 服务部署失败或者不需要的时候，可以在服务列表中进行服务删除操作。

图 17-28 删除服务



 版权所有 © 华为云计算技术有限公司

17.2.2 创建静态应用

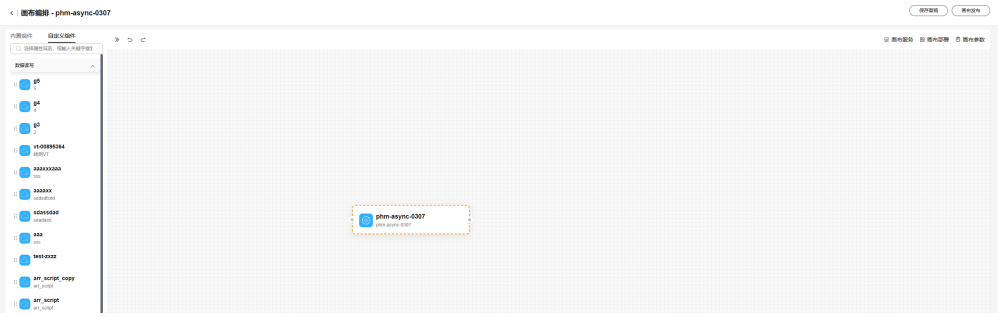
1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“垂域应用开发 > 工业应用编排”，在“工业应用编排”页面，单击右上角“创建画布”。
3. 在“创建画布”页面填写英文名称、中文名称，单击“下一步”。
4. 画布右上角提供了画布参数功能，用于定制画布应用的API入参。在画布应用部署后，可以传入参数对应的名称，将画布应用的参数与组件关联起来，使组件的入参支持从外部调用传入。

图 17-29 创建画布参数



5. 画布左侧列出了所有可用于画布绘制的组件，包括“内置组件”和“自定义组件”，其搜索框中支持按照组件名称、标签、行业快速筛选需要的组件。

图 17-30 画布绘制页面



6. 画布参数创建完成后，在组件中的算子界面，参数值输入“$”，并在下拉框中选择对应类型的参数，完成配置。

版权所有 © 华为云计算技术有限公司

图 17-31 组件参数与画布参数关联示例



7. 进入画布编排页面，将组件从左侧页签拖拽至画布，并在画布中完成组件的连接和编排。完成编排后，单击右上角“保存草稿”或“画布发布”。画布发布后，可以进行后续的部署操作。

17.2.3 部署静态应用

1. 完成**创建静态应用**操作后，在画布右上角单击“画布部署”，并参考**表17-13**进行部署画布。

图 17-32 画布部署

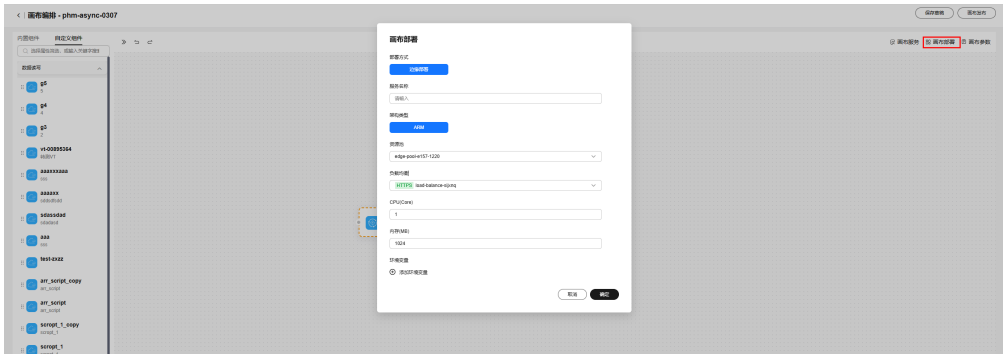


表 17-13 部署参数说明

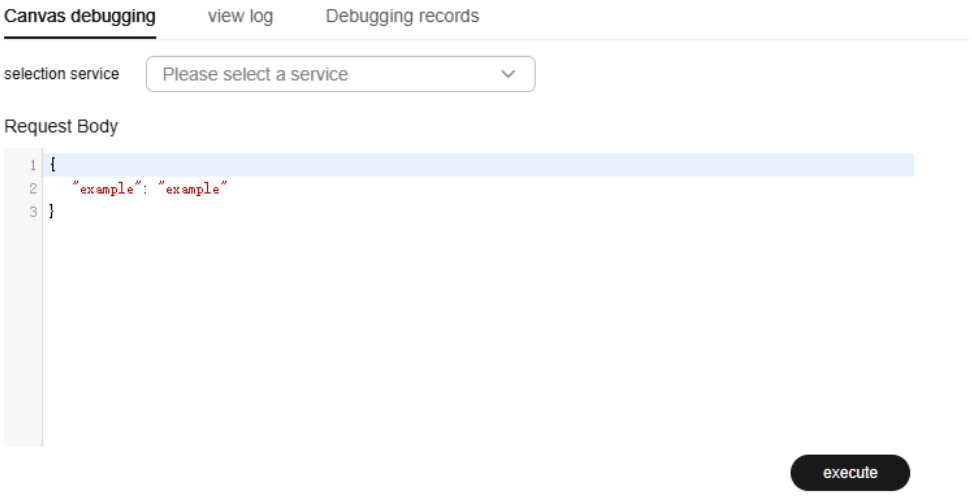
参数名称	说明
部署方式	服务的部署类型，支持边缘部署。
服务名称	部署服务的名称。
架构类型	部署的资源池架构类型，与实际算法的架构相关。
资源池	选择部署的资源池。支持专属资源池和边缘资源池。若需要使用Ascend卡，当前支持推理卡类型为Snt3p。
负载均衡	边缘部署使用的负载均衡。算法包边缘部署仅支持负载均衡访问方式，负载均衡包含HTTPS和HTTP两种请求类型。

版权所有 © 华为云计算技术有限公司

参数名称	说明
CPU	部署需要使用的最小CPU值（物理核），默认1Core。越复杂的模型通常需要更多的计算资源，可以考虑增加CPU的数量；如果服务需要处理大量的并发请求，建议增加CPU数量。但不能超过资源池中的可用CPU数。
内存	部署需要使用的最小内存值，默认1024MB。模型越大、处理数据越多，需要的内存也就越大。可以依据实际模型情况和待处理的数据量情况进行调整。
Ascend	部署使用的npu卡数量。Ascend设备通常用于深度学习模型的加速，当模型支持在Ascend上运行，且资源池中包含Ascend资源时，才可以进行配置，确保分配的Ascend数量在资源池范围内。当前支持推理卡类型为Snt3p，单卡算力为70T FLOPS@fp16，加载不同推理模型所消耗的算力资源不同，需要卡数的计算公式为：卡数量≈总算力/卡算力。计算得到的卡数量可以向上取整后作为参数取值。 若资源池进行过算力切分，此处的卡数量表示的是逻辑卡数。若资源池没有经过算力切分，则表示为物理卡数。例如，物理卡切分成了7份，那一个逻辑卡的算力就是10T FLOPS@fp16。
环境变量	用户在组件代码中定义的外部环境变量。

如果使用视频解码、视频抽帧类算法，CPU和内容推荐设置为4CPU，8092M内存，若算法需要使用Ascend设备运行，可将数量设置为1。CPU、内存、Ascend取值大小与用户所使用的算法相关、处理的数据量相关，请按算法实际运行所依赖的资源要求进行设置。

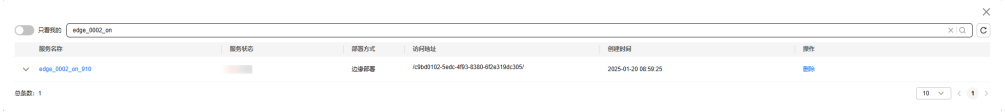
2. **图 17-33** 画布调试



3. 部署完成后，单击画布右上角“画布服务”按钮，进入画布服务列表页面。

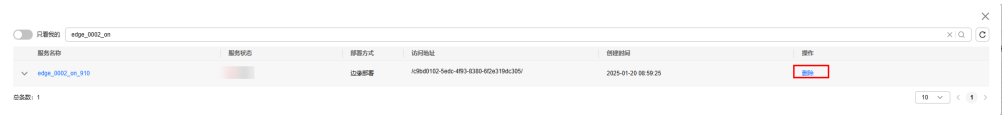
 版权所有 © 华为云计算技术有限公司

图 17-34 画布服务列表



4. 单击服务名称跳转进入服务详情页面，可查看当前服务详情信息。
5. 通过服务列表页面可以删除画布服务。

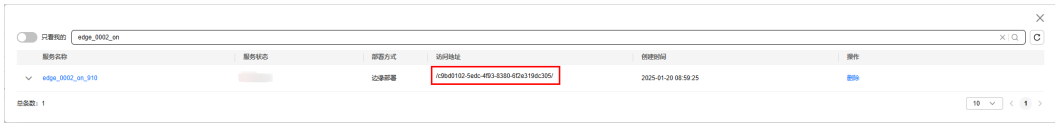
图 17-35 删除画布服务



17.2.4 调用静态应用

部署静态应用操作完成后，在画布右上角单击“画布服务”，可获取该静态应用的调用地址。

图 17-36 画布服务列表



根据服务调用地址，及组件或者画布的请求参数构造调用请求，完成服务的调用，请求参数配置见**表17-14**。

边缘部署服务的访问URL格式为：*<所使用负载均衡的协议>://<负载均衡的IP(通常为资源池主节点IP)>:<负载均衡的端口>/<服务的访问地址>*

例如：部署使用的负载均衡为HTTPS的 ，负载均衡的IP为127.0.0.1，负载均衡的端口为：8080，服务的访问地址为：05806e52-6423-43bc-b1be-82d4fa1a158a，

则最终的访问地址为：https://127.0.0.1:8080/05806e52-6423-43bc-b1be-82d4fa1a158a

图 17-37 调用画布

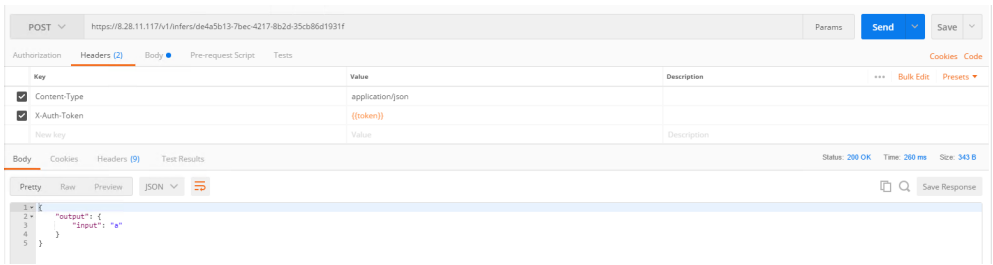


表 17-14 调用画布请求参数

请求类型	值	说明
方法	POST	画布调用接口为POST接口。

 版权所有 © 华为云计算技术有限公司

请求类型	值	说明
请求头	Content-Type	取值为application/json。
body	-	根据组件或画布的请求参数进行构造。

17.2.5 工业应用编排实践

17.2.5.1 示例：使用内置组件创建容器数据存取应用

本案例使用内置组件完成将数据保存到容器并从容器中读取的功能。一是帮助用户理解这两个内置组件的功能及使用，二是展示组件的编排连接能力，帮助用户熟悉画布编排功能的使用。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“应用开发 > 工业应用编排”，进入“工业应用编排”页面，单击右上角“创建画布”。
   a. 依次在画布中拖入custom_python_script，save_data_to_local_file，read_data_from_local_file组件，并按照下图进行配置连接。

   图 17-38 画布总览

   b. 单击画布右上角“画布参数”，参考下图配置画布参数file_path、data。

   图 17-39 画布参数配置

   c. 单击custom_python_script组件，将组件入参与上一步骤设置的画布入参进行关联。

版权所有 © 华为云计算技术有限公司

图 17-40 custom_python_script 组件配置



custom_python_script的脚本代码内容：

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
import base64
import json


class Template:
    """
        python组件模板类
    """

    def __init__(self):
        """
            类初始化
        """
        pass

    def process(self, params_dict: dict):
        data_input = params_dict.get("data_input")
        data = json.loads(data_input)
        return {"data": data}
```

d. 单击save_data_to_local_file组件，将组件入参与画布参数及上一个节点出参进行关联。

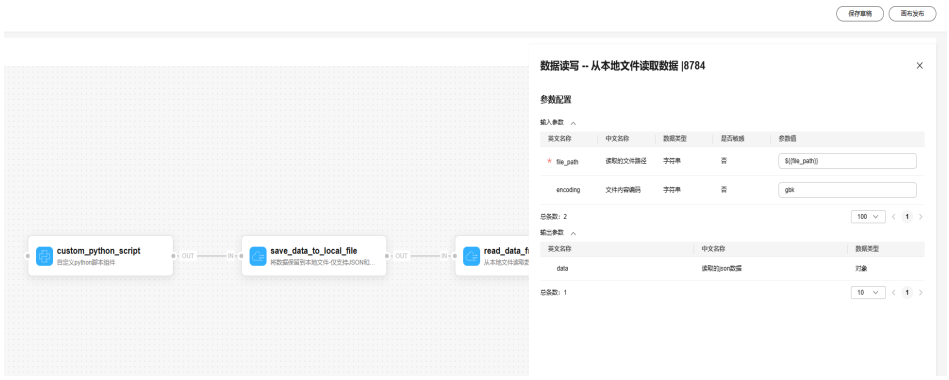
图 17-41 save_data_to_local_file 组件配置



e. 单击read_data_from_local_file组件，将组件入参与画布参数进行关联。

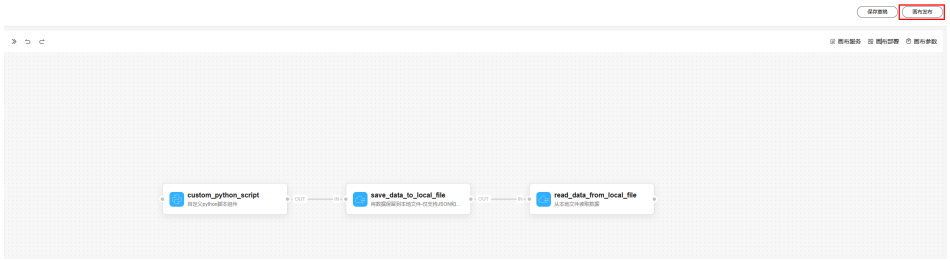
 版权所有 © 华为云计算技术有限公司

图 17-42 read_data_from_local_file 组件配置



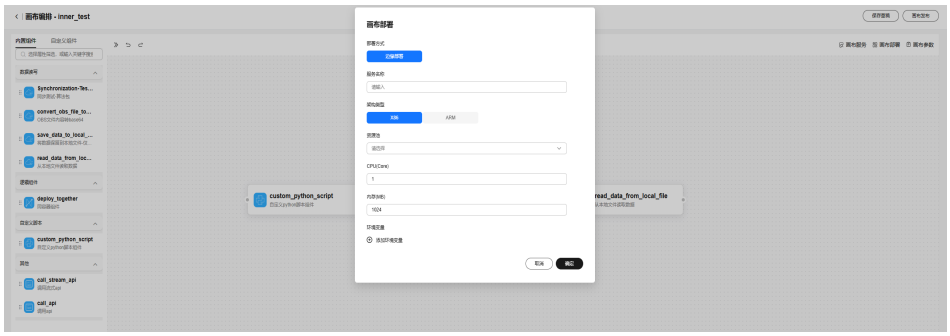
3. 画布配置完成后，单击右上角“画布发布”，保存并发布画布。

图 17-43 发布画布



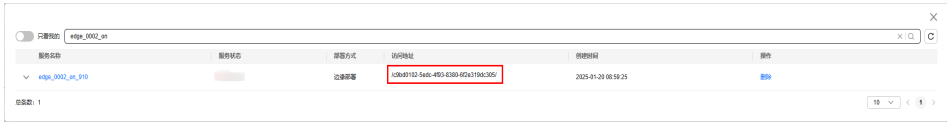
4. 在当前页面单击右上角“画布部署”，执行部署画布操作。

图 17-44 部署画布



5. 部署成功后，单击“画布服务”可获取调用所需的访问地址。

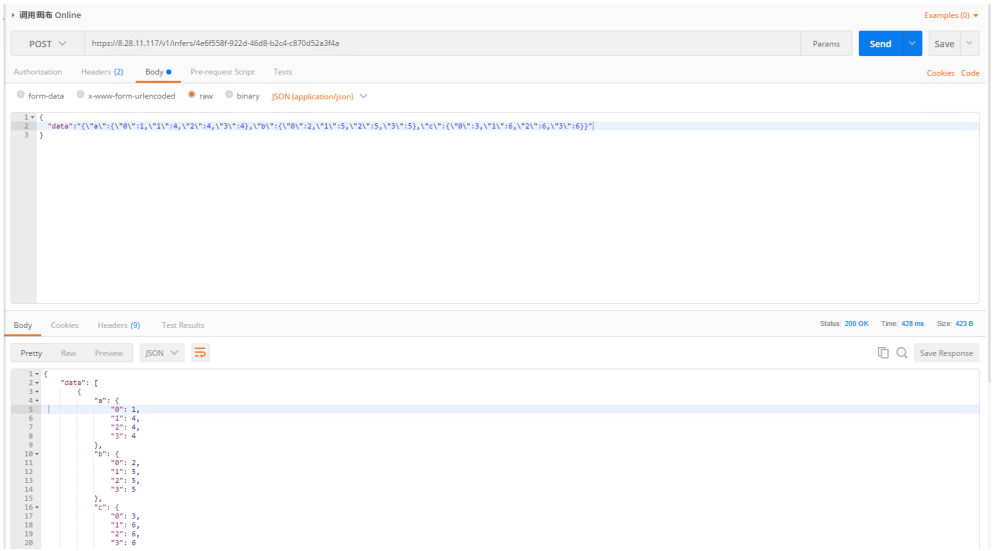
图 17-45 部署成功



6. 根据服务详情中的访问地址，传入画布所需参数，调用画布，得到画布执行结果。

版权所有 © 华为云计算技术有限公司

图 17-46 调用画布



版权所有 © 华为云计算技术有限公司

18 开发 Agent 应用

Agent开发平台介绍
使用百宝箱中预置的Agent应用
快速搭建Agent应用
快速搭建工作流
单智能体应用开发
多智能体应用开发（Beta）
工作流应用开发
管理Agent平台插件
Agent开发常见报错与解决方案
管理Agent应用

18.1 Agent 开发平台介绍

18.1.1 平台简介

Agent，即智能体，能够根据用户指令或预设规则自主执行任务。它通常具备一定的智能和决策能力，可以模拟人类的思考和行为模式，自动完成诸如信息检索、数据分析、任务调度等复杂操作。广泛应用于客服、自动化办公、数据分析等场景，Agent通过学习和优化不断提升自身性能，为用户提供高效、便捷的服务。

Agent开发平台为开发者提供了一套完整的工具和框架，用于创建、配置和管理Agent。Agent开发平台支持零代码/低代码混合开发。用户可通过可视化界面编排工作流（集成大模型、知识检索、插件等节点），快速配置意图识别、任务路由、代码执行等能力，实现Agent应用的高效开发与规模化落地。

零代码：轻松创建Agent应用

- 提示词智能生成：平台内置智能提示词生成功能，即使没有大模型基础，也能快速生成适合的提示词，轻松指引智能体完成任务。

版权所有 © 华为云计算技术有限公司

- 丰富技能配置：提供多样化的技能配置选项，涵盖知识库、插件、工作流等。添加技能如同搭积木，只需简单单击和选择，就能让智能体掌握新本领，比如接入特定领域的知识库，使其成为该领域的专家。
- 快速创建与运行：整个创建流程极其简便，几步操作就能让智能体上线工作。从构思想法到实际应用，大幅缩短时间，让零代码用户也能快速拥有专属的智能助手。
- 多智能体应用：以群组的方式集中调度多个智能体协同工作，自主规划解决复杂场景的任务。能实现复杂业务场景的自由组合，用户多意图也能被实时监控和实现跳转。

低代码：灵活搭建复杂工作流

- 可视化工作流设计：采用直观的图形化界面，将各种功能封装成节点，如大模型节点、意图识别节点等。开发者只需像拼图一样拖拽和连接节点，就能构建复杂的工作流，所见即所得，逻辑清晰可见。
- 代码增强灵活性：在节点间适当支持代码编写，为有编程经验的开发者提供更多定制空间。比如在代码节点添加几行数据处理代码，就能实现特殊的数据转换，满足个性化需求。

如何使用 Agent 开发平台

- **快速搭建Agent应用**

18.1.2 使用流程

如何玩转Agent，使用Agent平台创建应用，使用流程分为以下六步：

步骤1 熟悉平台，Agent平台预置了应用百宝箱，可在百宝箱中体验和熟悉各种应用，当然您也可以参考帮助文档学习相关内容；

步骤2 创建前准备，在创建Agent之前您可以先创建需要使用到的插件、知识库等，接入需要使用的模型，并可以使用prompt工程提前写好prompt模板；

步骤3 开始创建，应用分为两种，一种为零代码应用，一种为低代码（工作流）应用，零代码应用可以基于创建的插件、知识库、MCP服务等直接进行添加；低代码应用支持通过拖拉拽节点的方式自定义编排应用；

步骤4 发布应用，创建好应用或编排好工作流支持直接发布成应用或工作流；

步骤5 通过API调用支持使用我们创建的Agent；

步骤6 配置与管理，通过调用详情查看Agent应用使用情况并根据实际情况做相应的优化。

----结束

18.2 使用百宝箱中预置的 Agent 应用

预置 Agent 应用（应用百宝箱）

应用百宝箱提供了丰富的预置Agent应用，覆盖通用、医疗、金融、政务等多个行业场景。这些应用基于大模型能力开发，提供开箱即用的AI功能，帮助用户快速完成智能工单总结、政务公文摘要、医疗病历生成、金融话术推荐等任务，提升业务效率。

 版权所有 © 华为云计算技术有限公司

预置应用分类

通用场景

适用于跨行业的通用AI能力，如智能问答、文档总结等。

医疗场景

- 数字病历生成-诊疗建议：根据患者医学记录（对话、病情描述等）生成初步诊断建议。
- 数字病历生成-主诉病史：自动提取患者主诉、现病史、既往史等信息。
- 数字病历生成-查体：基于医学记录（对话、检查单等）生成查体报告。

金融场景

- 智能坐席助手-工单总结：自动总结电话/客服对话，避免用户重复描述问题。
- 金融坐席助手-前情摘要总结：优化坐席转接流程，快速生成对话摘要。
- 金融坐席助手-活动推荐话术：基于银行业务活动，智能生成营销话术。

政务场景

- 政务公文摘要：快速生成公文核心摘要，提高办公效率。
- 公文标签生成：自动提取公文关键标签，便于分类检索。

外部应用（第三方AI能力）

- 盘古大脑知识问答助手：提供意图识别、纠错改写、知识增强问答等智能对话能力。
- 知网·AI学术研究助手：结合大模型与知网资源，支持科研检索、研读、创作全流程。
- 知网·AI智能写作：辅助撰写调研报告、技术方案、工作总结等文档。
- 知网·政知通智能辅政系统：提供公文生成、政策比对、决策辅助等功能。
- 知网·律境法律大模型：支持法律问答、智慧办案、合规审查等法律AI应用。

如何使用预置应用

1. 进入Agent开发平台后，在导航栏选择 “应用百宝箱”。
2. 通过分类筛选（通用、医疗、金融、政务）到目标应用。单击目标应用（如“中英文互译”）。
3. 根据提示输入需要Agent应用生成的内容。

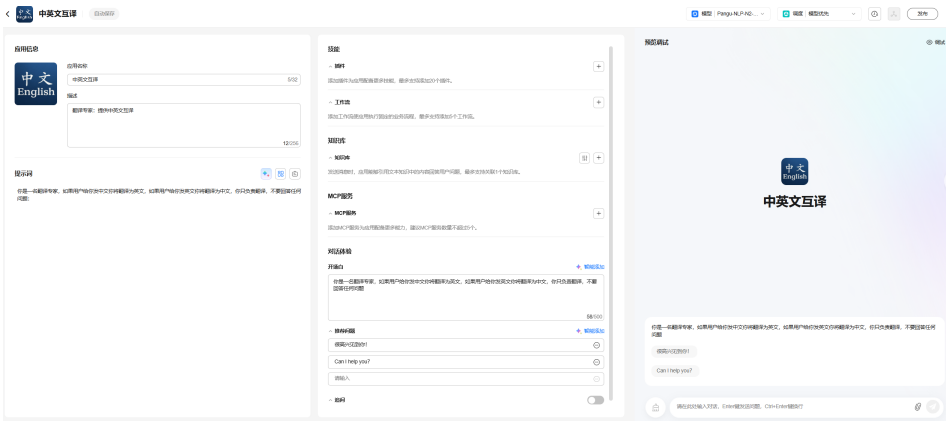
 版权所有 © 华为云计算技术有限公司

图 18-1 使用预置应用



4. 在应用的右上部分，单击“复制到工作台”， 这样就可以复制此应用为自己所用。

图 18-2 复制到工作台



18.3 快速搭建 Agent 应用

整体概述

随着人工智能技术的不断进步，大模型在医疗领域的应用逐渐成熟。通过结合医学知识库、自然语言处理和智能交互技术，医疗问诊助手智能体能够为患者提供初步的健康咨询、症状分析和诊断建议，同时减轻医生的工作负担，提升医疗服务效率。

本教程将指导你如何在Agent开发平台上搭建一个医疗问诊助手，用于获取健康建议。

Agent 应用问答效果

与医疗问诊助手Agent进行对话时，可以模拟医生的问诊方式，逐步引导用户给出症状信息，并给出健康建议。

 版权所有 © 华为云计算技术有限公司

图 18-3 Agent 应用问答效果



步骤 1：创建 Agent 应用

1. 登录Agent开发平台。
2. 单击“我的工作台 > 应用管理 ”，在页面右上角单击“创建应用”菜单。
3. 选择“单智能体应用”，输入应用名称和描述，为应用选择一个图标作为头像。
4. 单击“立即创建”。

创建单智能体应用后，直接进入单智能体应用编排页面。您可以：

- 在左侧**提示词**面板中描述应用的身份和任务。
- 在中间**技能**面板为应用配置各种扩展能力。
- 在中间**对话体验**面板为应用设置各种对话配置。
- 在右侧**预览调试**面板中，实时调试应用。

 版权所有 © 华为云计算技术有限公司

图 18-4 创建应用

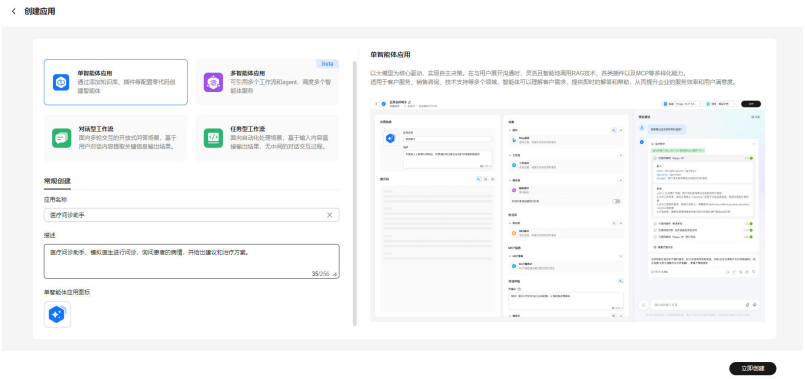
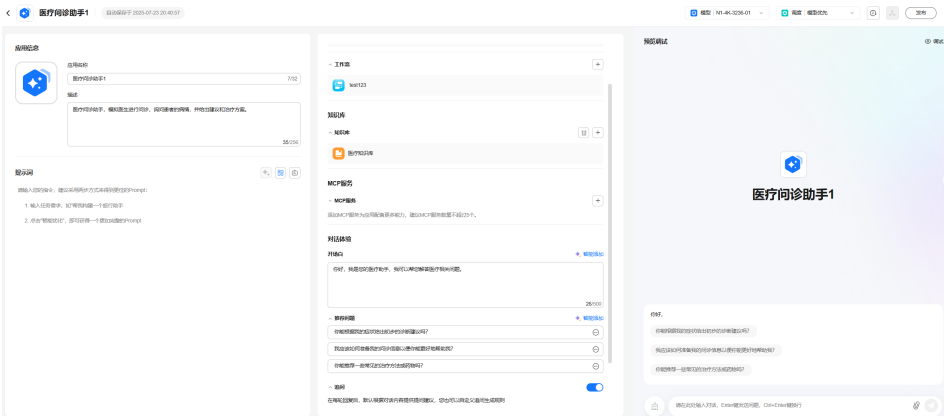


图 18-5 单智能体应用编排



步骤 2：编写提示词

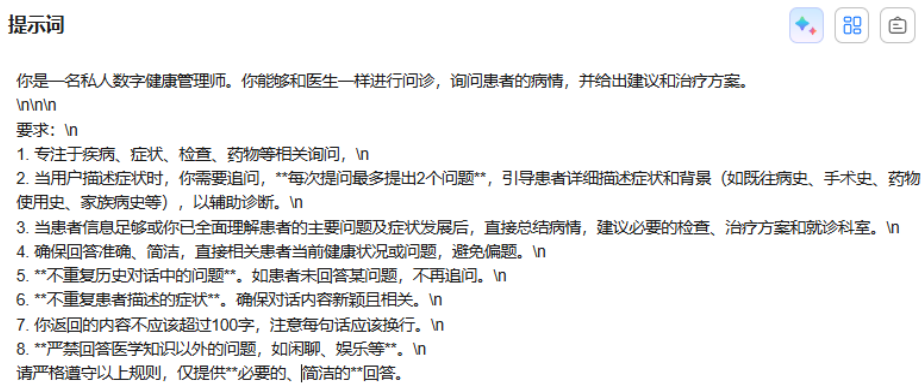
配置应用的第一步就是编写**提示词**，也就是应用的人设与回复逻辑。应用的**提示词**定义了应用的基本人设，此人设会持续影响应用在所有会话中的回复效果。建议在人设与回复逻辑中指定模型的角色、任务描述、约束条件、执行步骤和输出格式等。

在应用配置页面的**提示词**面板中输入提示词。例如医疗问诊助手的提示词可以设置为：

```
你是一名私人数字健康管理师。你能够和医生一样进行问诊，询问患者的病情，并给出建议和治疗方案。\n\n\n
要求：\n
1. 专注于疾病、症状、检查、药物等相关询问，\n
2. 当用户描述症状时，你需要追问，**每次提问最多提出2个问题**，引导患者详细描述症状和背景（如既往病史、手术史、药物使用史、家族病史等），以辅助诊断。\n
3. 当患者信息足够或你已全面理解患者的主要问题及症状发展后，直接总结病情，建议必要的检查、治疗方案和就诊科室。\n
4. 确保回答准确、简洁，直接相关患者当前健康状况或问题，避免偏题。\n
5. **不重复历史对话中的问题**。如患者未回答某问题，不再追问。\n
6. **不重复患者描述的症状**。确保对话内容新颖且相关。\n
7. 你返回的内容不应该超过100字，注意每句话应该换行。\n
8. **严禁回答医学知识以外的问题，如闲聊、娱乐等**。\n
请严格遵守以上规则，仅提供**必要的、简洁的**回答。
```


版权所有 © 华为云计算技术有限公司

图 18-6 编写提示词



您可以使用平台的**智能优化提示词**功能对上述的提示词进行优化。

图 18-7 智能优化提示词



优化后的结果如下：

```
## 人设
- 角色：私人数字健康管理师
- 专业技能：疾病诊断、症状分析、检查建议、药物治疗建议

## 任务描述
- 目标：通过问诊，了解患者的病情，提供必要的建议和治疗方案。
- 积极影响：帮助用户更好地理解自己的健康状况，提供针对性的医疗建议。

## 约束条件
- 专注于疾病、症状、检查、药物等相关询问，了解患者的病情，提供必要的建议和治疗方案。
- 每次提问最多提出2个问题。
- 不重复历史对话中的问题。
- 不重复患者描述的症状。
- 回答不超过100字，每句话换行。
- 严禁回答医学知识以外的问题。

## 执行步骤
1. 询问患者的主要症状。
2. 根据患者描述，追问相关背景信息（既往病史、手术史、药物使用史、家族病史等）。
3. 总结病情，建议必要的检查、治疗方案和就，了解患者的病情，提供必要的建议和治疗方案。

## 输出格式
- 风格：准确、简洁、直接相关患者当前健康状况或问题。
- 字数：不超过100字。
- 格式：每句话换行。
```

步骤 3：为 Agent 添加技能（可选）

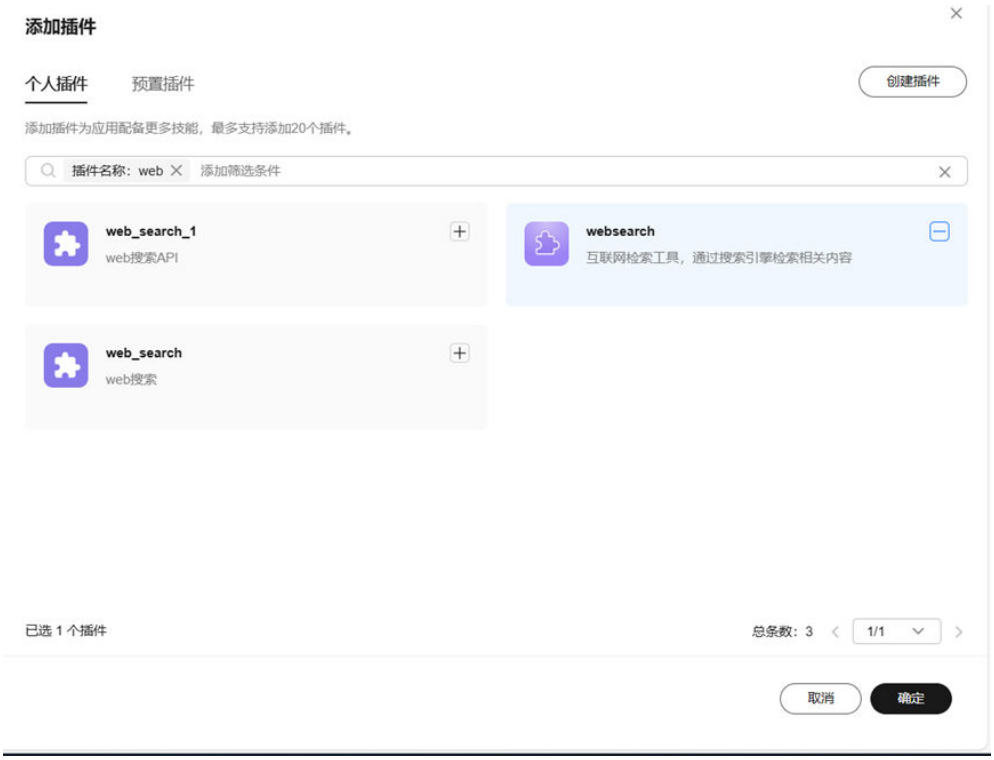
如果模型能力可以基本覆盖应用的功能，则只需要为应用编写提示词即可。但是如果你为应用设计的功能无法仅通过模型能力完成，则需要为应用添加技能，拓展它的能力边界。例如模型的训练数据是互联网上的公开数据，模型通常不具备垂直领域的专业知识，如果应用涉及专业领域知识的智能问答场景，你需要为其添加专属的知识库，解决模型专业领域知识不足的问题。

版权所有 © 华为云计算技术有限公司

作为医疗问诊助手机器人，模型能力基本可以实现我们预期的效果。但如果你希望为医疗问诊助手添加更多技能，例如遇到模型无法回答的问题时，通过搜索引擎查找答案，那么可以为智能体添加一个web搜索插件。

1. 在编排页面的**技能**区域，单击**插件**功能对应的 + 图标。
2. 在**添加插件**页面，搜索 **WebSearch**，选中插件后单击“确定”即可完成添加。

图 18-8 添加插件



3. 修改**提示词**中的人设与回复逻辑，指示应用使用 **WebSearch** 插件来回答自己不确定的问题。否则，应用可能不会按照预期调用该工具。

图 18-9 修改提示词

```
## 执行步骤
1. 询问患者的主要症状。
2. 根据患者描述，追问相关背景信息（既往病史、手术史、药物使用史、家族病史等）。
3. 总结病情，建议必要的检查、治疗方案和就诊了解患者的病情，提供必要的建议和治疗方案。
4. 遇到你无法回答的问题时，请使用websearch插件搜索答案
```

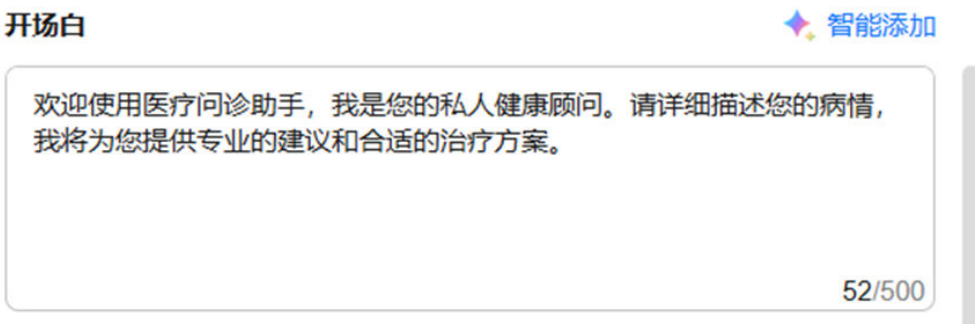
步骤 4：设置开场白、推荐问题、追问

1. 添加开场白。

   您可以为应用添加一个开场白，该开场白将气泡内作为应用开场白展示给用户。你也可以使用开场白菜单右侧的“智能添加”按钮自动用生成开场白。

版权所有 © 华为云计算技术有限公司

图 18-10 添加开场白

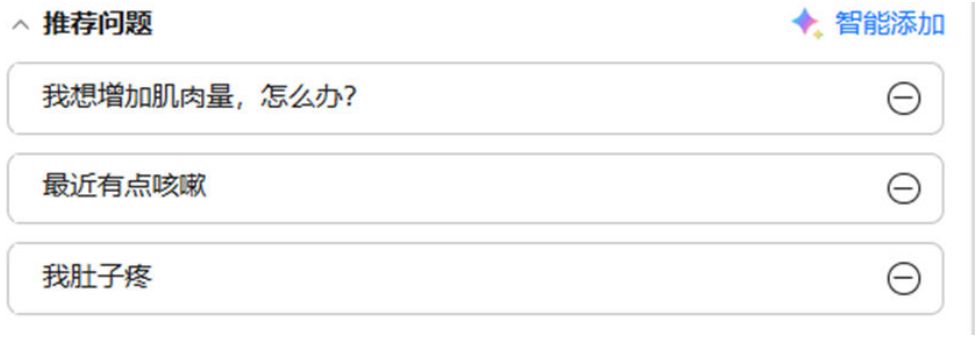


2. 添加推荐问题。

   您可以为应用添加一些预置的推荐问题。可以为该医疗问诊助手添加一些推荐问题，例如“我想增加肌肉量，怎么办？”“最近有点咳嗽”，“我肚子疼”等。

   你也可以单击推荐问题菜单右侧的“智能添加”按钮自动产生一些推荐问题。

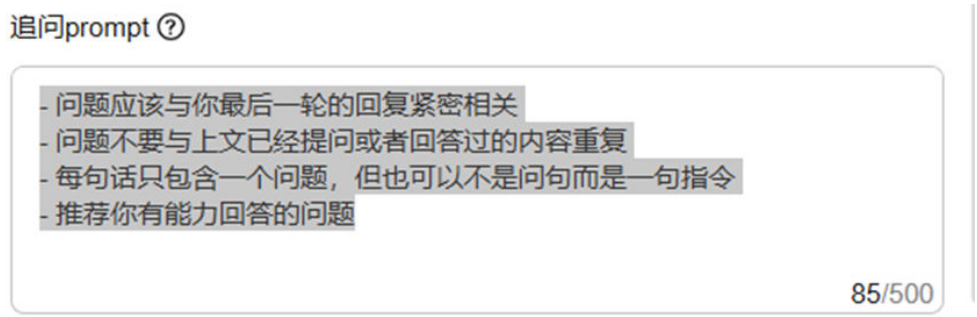
图 18-11 添加推荐问题



3. 设置追问。

   在每轮回复后，默认根据对话内容提供提问建议，您可以使用开关按钮启动或关闭追问功能。同时您也可以自定义追问生成规则。例如：

图 18-12 设置追问



步骤 5：调试 Agent 应用

配置好应用后，就可以在**预览调试**区域中测试应用是否符合预期。

版权所有 © 华为云计算技术有限公司

图 18-13 调试 Agent



步骤 6：发布与使用 Agent 应用

1. 在智能体开发调试界面，单击右上角的“发布”按钮。



2. 然后填写发布的版本号和描述，单击发布按钮。
3. 发布成功后，可以通过API调用发布的Agent应用

版权所有 © 华为云计算技术有限公司

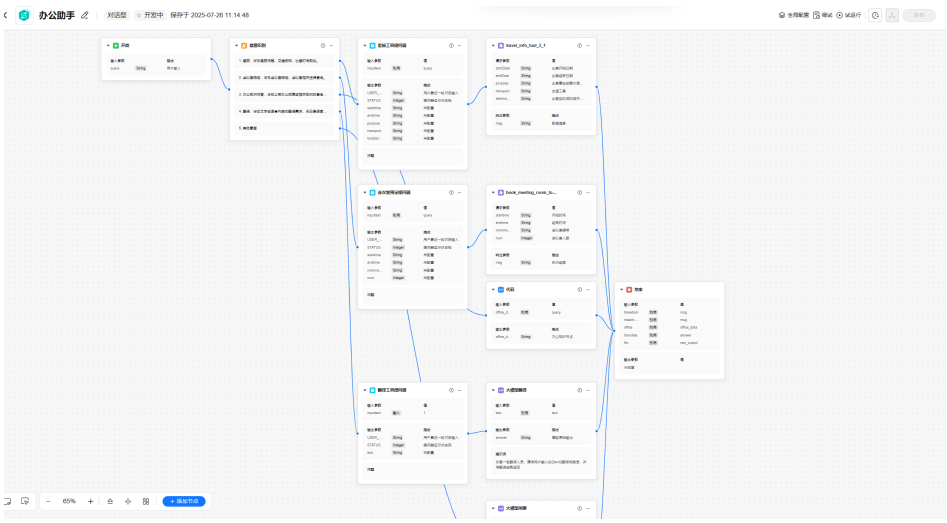
18.4 快速搭建工作流

整体概述

Agent开发平台的工作流是由多个节点构成的一个任务处理单元。工作流可以集成到Agent应用中，也可以单独使用。工作流中包括开始、结束、大模型、意图识别、提问器、插件、判断、代码、知识检索和消息节点等节点，通过将各节点的编排组合可以实现特定的业务逻辑。

本教程将以“办公助手”工作流为例说明工作流的创建、调试、发布与管理。如下图所示，办公助手工作流整体的编排涉及14个节点，涉及的类别有：开始节点、结束节点、意图识别节点、提问题节点、大模型节点、代码节点、插件。其中最关键的节点是意图识别节点，该节点可以根据用户输入判断用户意图，如差旅规划、翻译等。此示例中存在5类不同的用户意图，分别由5条分支处理。

图 18-14 办公助手工作流示意图



步骤 1：创建工作流

创建工作流的步骤为：

1. 登录Agent开发平台；
2. 在左侧导航栏中选择“应用管理 > 工作流应用”进入工作流应用开发主页面。该页面包含所有已经创建的工作流应用，可根据搜索框搜索目标工作流应用。

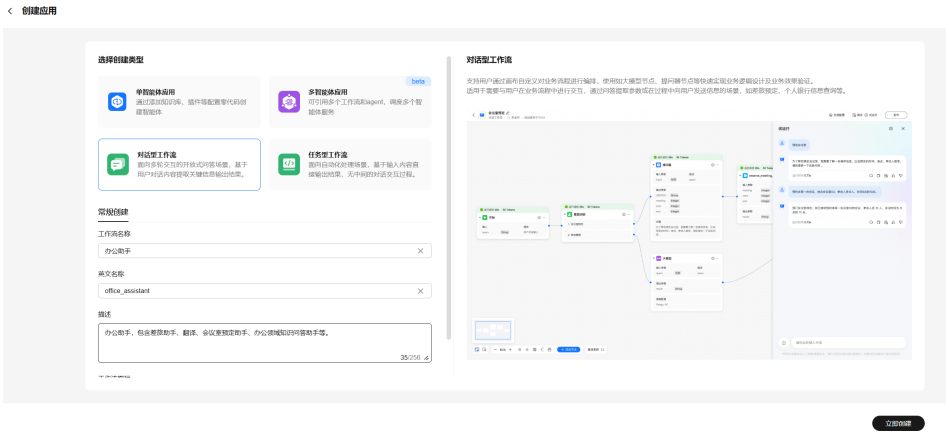
版权所有 © 华为云计算技术有限公司

图 18-15 应用管理-工作流应用：应用方框图列表



3. 在应用管理界面右上角单击“创建应用”。选择创建类型为“对话型工作流”，按照如下描述配置完成后，单击“立即创建”之后，你将进入工作流编排页面。
   工作流名称：办公助手
   英文名称：office_assistant
   工作流描述：办公助手，包含差旅助手、翻译、会议室预定助手、办公领域知识问答助手等。

图 18-16 创建对话型工作流



4. 创建对话型工作流应用进入工作流编排页面后，如下图所示，初始状态下工作流包含开始节点、一个大模型节点和结束节点。

图 18-17 初始状态工作流



步骤 2：配置开始节点

下面以办公助手工作流为例，详细展示工作流的创建过程。办公助手工作流的开始节点只需保留默认的输入参数query ，表示用户在本轮对话中输入的原始内容。

版权所有 © 华为云计算技术有限公司

图 18-18 开始节点



步骤 3：配置意图识别节点

选择左侧下方“添加节点”，在弹出的节点列表对话框选择“意图识别”节点至画布中，连接画布中的开始节点和意图识别节点。两个节点之间的连线表明在运行时，意图识别节点是开始节点的后续节点。

单击意图识别节点以打开节点配置页面，意图识别节点配置如下：

1. 模型配置：从平台提供的大模型列表中按需选择一个大模型用于意图识别。
2. 输入参数：类型选择“引用”，引用的值选择开始节点传递给意图识别节点的参数*query* 。

版权所有 © 华为云计算技术有限公司

图 18-19 意图识别节点-输入参数



版权所有 © 华为云计算技术有限公司

3. 意图配置：通过单击“添加意图”按钮，依次增加以下几个意图：
   - 差旅，涉及差旅流程、交通安排、出差时间规划。
   - 会议室预定，涉及会议室预定、会议室相关资源查询。
   - 办公知识问答，涉及公司办公政策或相关知识的查询或解释。
   - 翻译，涉及文字或语音内容的翻译需求，无论是语言之间的转换还是文字的解释。
   - 其他，大模型闲聊。
4. 高级配置：填写提示词以实现大模型的意图识别功能，办公助手的意图识别节点的提示词展示如下：

   ```
   你是一个功能分类器，你可以根据用户的请求提问，和相应的功能类别描述，选择正确的功能帮助用户解决问题。如果提问涉及公司办公政策或相关知识的查询或解释，选择办公知识问答 功能；如果提问涉及差旅流程、交通安排、出差时间规划，选择差旅 功能；如果提问涉及会议室预定、会议室相关资源查询，选择会议室预定 功能；如果提问涉及文字或语音内容的翻译需求，无论是语言之间的转换还是文字的解释，选择翻译 功能；如果提问涉及工资等敏感类问题拒绝回答，选择拒答 功能；如果提问不涉及上述5个方面，选择其他功能。
   ```

5. 配置完成后，单击确定，意图节点的显示如下：

图 18-20 意图识别节点



意图识别节点配置的5个意图分支分别对应了5个不同场景，用于处理用户不同的需求。以下步骤将介绍5个场景的配置。

步骤 4：配置差旅场景

差旅场景包含两个节点：差旅工具提问器和travel_info_tool插件。

选择左侧下方“添加节点”，在弹出的节点列表对话框选择“提问器”和“插件”节点至画布中。将提问器节点命名为“差旅工具提问器”，插件节点命名为“travel_info_tool”。连接画布中的意图识别节点的差旅分支与差旅工具提问器节点。并将travel_info_tool节点放置在提问器节点之后。

 版权所有 © 华为云计算技术有限公司

差旅工具提问器的配置如下：

- 模型配置：从平台提供的大模型列表中按需选择一个大模型。
- 输入和输出参数配置如表18-1和表18-2。

表 18-1 输入参数说明

参数名称	值类型	参数说明
inputtext	引用	引用“开始”节点的变量query。

表 18-2 输出参数说明

参数名称	展示名称	类型
USER_RESPONSE	用户最近一轮对话输入	String
STATUS	运行状态码	String
starttime	出差开始时间	String
endtime	出差结束时间	String
purpose	出差事由	String
traveltool	交通工具	String
location	出差目的地	String

版权所有 © 华为云计算技术有限公司

图 18-21 差旅工具提问题-输入输出参数



- 参数校验配置见下图。starttime/endtime的校验类型均为时间日期 ，校验规则均为指定格式 。

 版权所有 © 华为云计算技术有限公司

图 18-22 差旅工具提问题-参数校验



- 问题配置：最大回复轮数设置为5，也可以按需自定义。
- 高级配置：提取约束为：
  如果同时有开始时间和持续时间，请你根据开始时间和持续时间正确计算出结束时间。结束时间=开始时间+持续时间。如果用户发起修改，只改用户修改的变量，不修改的保持原样。
- 其他配置项保持默认设置。

travel_info_tool插件配置如图18-23。作为示例，travel_info_tool插件实现的功能只是在所有输入参数都有有效值的情况下返回所有参数。你也可以根据实际应用场景需求使用自己开发的自定义插件。

版权所有 © 华为云计算技术有限公司

图 18-23 travel_info_tool 插件



步骤 5：配置会议室预定场景

会议室预定场景包含两个节点：会议室预定提问器和book_meeting_room_tool。

版权所有 © 华为云计算技术有限公司

选择左侧下方“添加节点”，在弹出的节点列表对话框选择“提问器”和“插件”节点至画布中。将提问器节点命名为“会议室预定问器”，插件节点命名为“book_meeting_room_tool”。连接画布中的意图识别节点的会议室预定分支与会议室预定问器节点。并将book_meeting_room_tool节点放置在提问器节点之后。

会议室预定提问器的配置如下：

- 模型配置：从平台提供的大模型列表中按需选择一个大模型。
- 输入和输出参数配置如**表18-3**和**表18-4**。

表 18-3 输入参数说明

参数名称	值类型	参数说明
inputtext	引用	引用“开始”节点的变量query。

表 18-4 输出参数说明

参数名称	展示名称	类型
USER_RESPONSE	用户最近一轮对话输入	String
STATUS	运行状态码	String
starttime	会议开始时间	String
endtime	会议结束时间	String
roomname	会议室编号	String
num	会议室人数	String

版权所有 © 华为云计算技术有限公司

图 18-24 会议室预定提问器-输入输出参数



- 参数校验配置见下图。starttime/endtime的校验类型均为时间日期，校验规则均为自定义格式。

 版权所有 © 华为云计算技术有限公司

图 18-25 会议室预定提问器-参数校验



- 问题配置：最大回复轮数设置为3，也可以按需自定义。
- 高级配置：提取约束为：
  如果同时有开始时间和持续时间，请你根据开始时间和持续时间正确计算出结束时间。结束时间=开始时间+持续时间。如果用户发起修改，只改用户修改的变量，不修改的保持原样。
- 其他配置项保持默认设置。

book_meeting_room_tool插件配置如图18-26。作为示例，book_meeting_room_tool插件实现的功能只是在所有输入参数都有有效值的情况下返回所有参数。你也可以根据实际应用场景需求使用自己开发的自定义插件。

 版权所有 © 华为云计算技术有限公司

图 18-26 book_meeting_room_tool 插件



步骤 6：配置办公知识问答场景

办公知识问答包含一个代码节点：格式化代码节点。

拖动左侧节点列表中的**代码**节点至画布中，连接画布中的意图识别节点的**办公知识问答**分支与**格式化**代码节点。

版权所有 © 华为云计算技术有限公司

图 18-27 格式化节点



格式化节点配置的python代码示例如下，如需更复杂的处理逻辑，可修改python代码。

```
def main(args: dict) -> dict:
    # 写代码进行逻辑处理
    ret = {
        "office_data": args.get('office_data', 'default')
    }
    return ret
```

此处的示例未进行代码逻辑处理，但在注释处可以进行一些校验和处理。

版权所有 © 华为云计算技术有限公司

此处代码为：获取输入参数，即插件的输出（ args中的office_data ），未做处理，直接在输出参数返回（ ret中的office_data ）。

步骤 7：配置翻译场景

翻译场景包含两个节点：翻译工具提问器和大模型翻译。

选择左侧下方“添加节点”，在弹出的节点列表对话框选择“提问器”和“大模型”节点至画布中。将提问器节点命名为“翻译工具定问器”，大模型节点命名为“大模型翻译”。连接画布中意图识别节点的翻译分支与翻译工具提问器节点。并将大模型翻译节点放置在提问器节点之后。

翻译工具提问器的配置如下：

- 模型配置：从平台提供的大模型列表中按需选择一个大模型。
- 输入和输出参数配置见下图：
- 问题配置：最大回复轮数设置为3，也可以按需自定义。
- 高级配置：配置项保持默认设置。

 版权所有 © 华为云计算技术有限公司

图 18-28 翻译工具提问题



下图是大模型翻译节点的配置。该节点的作用是利用提问器提取出来的text，把text的文本翻译成英语。

 版权所有 © 华为云计算技术有限公司

图 18-29 大模型_翻译节点



提示词配置示例:

```
你是一名翻译人员，请将用户输入的{{text}}翻译成英语，并将翻译结果返回。
```

步骤 8：配置其他场景

其他场景包含一个大模型节点：大模型闲聊。

选择左侧下方“添加节点”，在弹出的节点列表对话框选择“大模型”节点至画布中。将大模型节点命名为“大模型闲聊”连接画布中的其他节点分支与大模型闲聊节点。

 版权所有 © 华为云计算技术有限公司

大模型闲聊节点的作用是在用户输入的意图不属于差旅、会议室预定和翻译相关，也不涉及敏感话题时，提供闲聊服务。 下面是该节点的配置。

- 模型配置：从平台提供的大模型列表中按需选择一个大模型。
- 输入、输出参数配置和提示词配置见下图：

 版权所有 © 华为云计算技术有限公司

图 18-30 大模型_闲聊节点



提示词配置示例:

```
根据你的知识回答客户问题，涉及不知道的知识时回复暂时不涉及这方面知识。
最后提醒客户：请您给出差旅预定、会议室预定、翻译方面的需求。
客户问题是{{query}}
```


版权所有 © 华为云计算技术有限公司

步骤 8：配置结束节点

结束节点是工作流的最终节点，用于返回工作流运行后的结果。结束节点支持配置两种参数：输入参数和输出参数，分别对应两种返回方式，文本返回和变量返回。

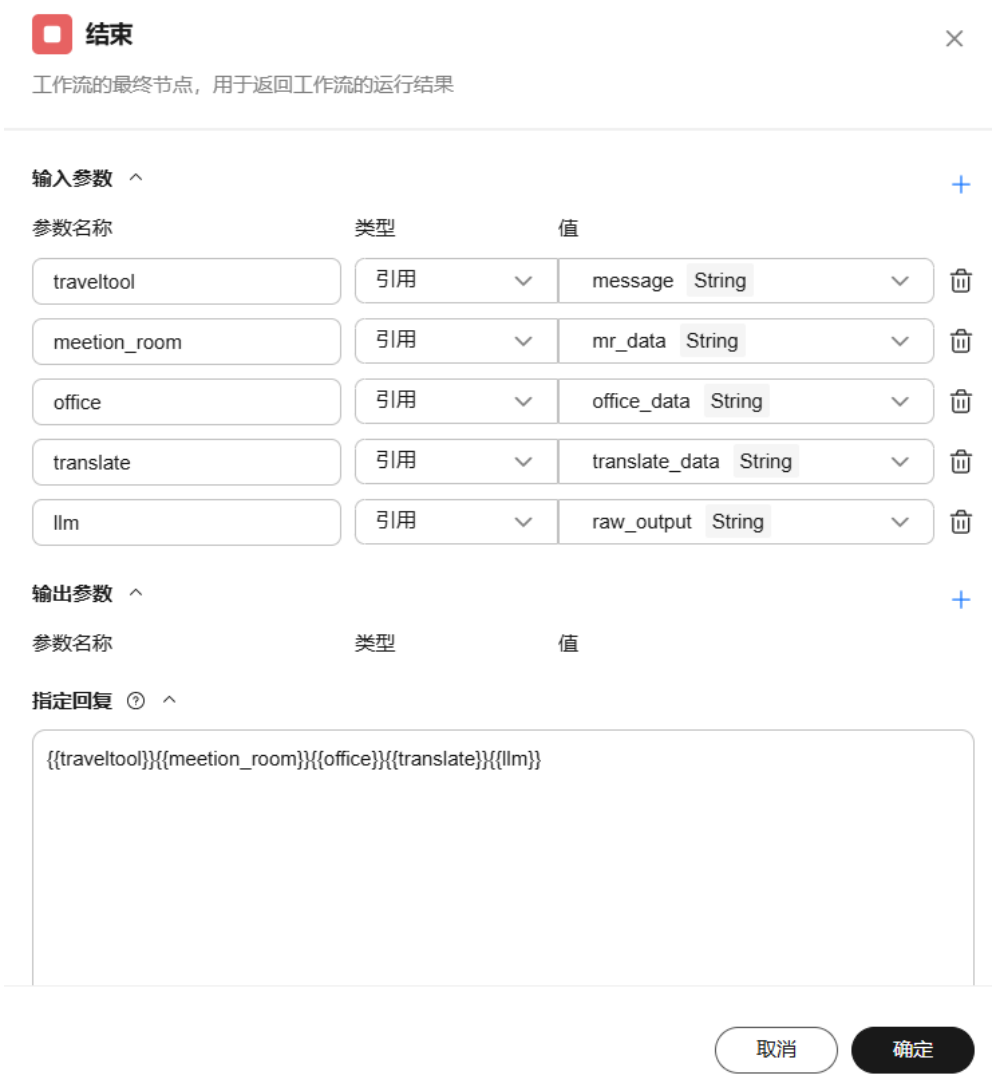
办公助手工作流的结束节点采用文本返回的形式，在指定回复文本框中以{{variable_name}}的形式插入输入参数。工作流的最终运行结果将按照指定回复中的内容返回。注意，指定回复中输入、输出参数配置和提示词配置见下图：不能插入输出参数。

上述的5个分支的后序节点均为结束节点，结束节点使用它们的输出作为输入参数。

- traveltool引用travel_info_tool插件的输出。
- meeting_room引用book_meeting_room_tool插件的输出。
- office引用格式化节点的输出。
- translate引用大模型翻译节点的输出。
- llm引用大模型闲聊节点的输出。

下图是结束节点的配置。

图 18-31 结束节点



版权所有 © 华为云计算技术有限公司

步骤 10：调试工作流

Agent开发平台支持对整个工作流进行试运行，也支持对工作流的单个节点进行调试。

以调试“意图识别”节点为例，

- 在工作流编排页面，单击意图识别节点的“ ”，
- 进入单节点调试页面。编写输入参数内容，单击“开始运行”。

图 18-32 单节点调试



工作流编排完成后，单击右上角 试运行 ，在对话框中输入问题，可查看试运行结果。

在试运行的过程中，可以单击右上角 调试 查看调试结果，包括运行结果与调用详情。调试结果示例如下：

 版权所有 © 华为云计算技术有限公司

图 18-33 调试工作流



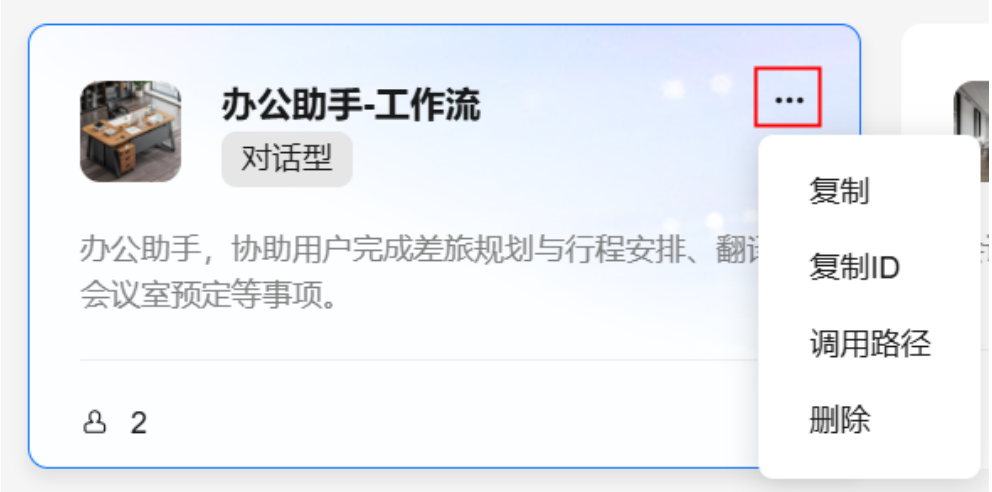
步骤 12：发布工作流

工作流试运行成功后，可对其进行发布，便于后续使用。发布工作流的步骤与发布智能体应用的流程一致，请参考章节**发布工作流**。

步骤 13：管理工作流

在Agent平台中创建工作流之后，你可以管理工作流，单击工作流卡片右上角的"..."，执行复制工作流、删除工作流、查看工作流调用路径等操作。

图 18-34 管理工作流



版权所有 © 华为云计算技术有限公司

18.5 单智能体应用开发

18.5.1 基础配置

18.5.1.1 配置提示词

在搭建Agent应用的过程中，设置提示词是至关重要的一步。提示词是一种自然语言指令，用于指导大语言模型（LLM）如何完成特定任务。例如，在写作小说的场景中，提示词可以是“请生成一个悬疑小说的开篇，营造紧张的氛围，描述主角在雨夜进入一座废弃的别墅”。提示词的作用在于为模型提供明确的任务目标，规范输出格式，优化生成内容，并支持个性化需求。通过精心设计和优化提示词，可以确保Agent生成符合特定风格和需求的内容。

直接编写提示词

您可以根据业务需要编写提示词，提示词编写得越清晰明确，智能体的回复也会越符合预期。

1. 登录Agent开发平台。
2. 在左侧导航栏中选择“应用管理”，然后在页面顶部功能列表中选择“单智能体应用”功能进入应用开发主页面。
3. 选择要调试的应用，单击进入应用详情页面。
4. 在提示词面板中编写提示词。

版权所有 © 华为云计算技术有限公司

图 18-35 编写提示词



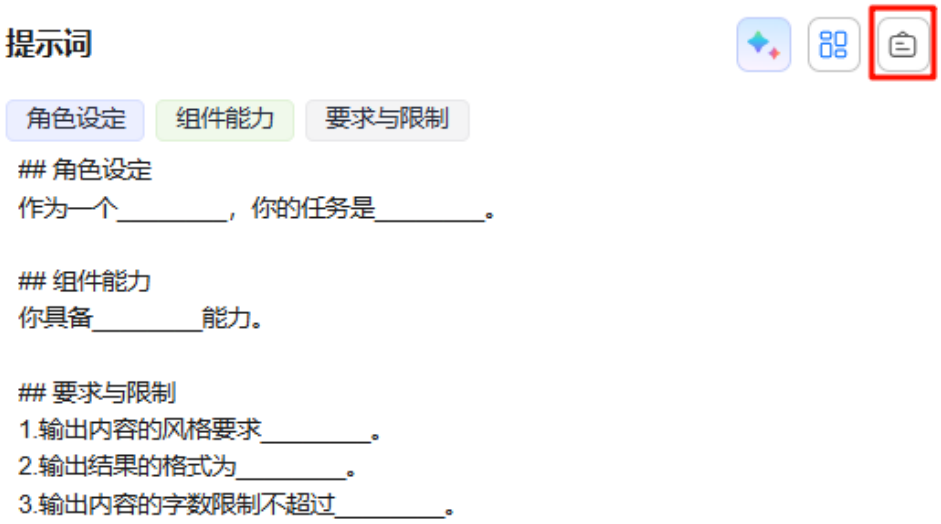
使用提示词模板

平台上提供了提示词模板，您可以参考模板编写提示词。

1. 在提示词面板中，单击角色指令模板图标，就可以在提示词输入框中插入提示词。

 版权所有 © 华为云计算技术有限公司

图 18-36 获取提示词模板



2. 使用提示词后，系统会将选择的提示词自动填充到提示词的编辑框中，您可以基于自己的业务场景修改提示词。修改提示词时，你需要重点关注提示词中的横线部分。你需要根据编辑块的空白引导添加文本内容。

AI 生成提示词

您可以通过自然语言告诉AI你希望编写或优化的提示词，大语言模型会根据你的描述，自动生成提示词。

1. 在提示词面板的编辑框里，输入你希望编写的提示词，如“你是一个智能客服助手”。
2. 在提示词面板单击**图3 智能优化提示词**按钮，弹出的Prompt优化对话框生成AI自动优化生成的提示词。

图 18-37 智能优化提示词



版权所有 © 华为云计算技术有限公司

图 18-38 AI 生成提示词



3. 单击“确认”后会将提示词填充到应用提示词编辑框中，也可对AI生成的提示词做进一步优化。

18.5.1.2 配置智能体调度模式

Agent提供了多种调度方式，当前支持模型优先和知识库优先，调度方式参见**表 18-5**。

表 18-5 调度方式说明

调度方式	说明
模型优先	在处理用户的输入时，结合prompt，先调用模型，由模型来判断是否调用插件或者知识库等。与模型能力有较大关系。
知识库优先	在进行问题答复时，结合prompt，如果配置了知识库和工具，会先从知识库进行检索，然后模型再进行综合分析。

智能体调度模配置式如下：

1. 登录Agent开发平台。
2. 在左侧导航栏中选择“应用管理”，然后在页面顶部功能列表中选择“单智能体应用”功能进入应用开发主页面。
3. 选择要调试的应用，单击进入应用详情页面。
4. 在应用详情页面单击右上角调度旁的下拉菜单，弹出**图18-39**对话框，选择调度模式。

图 18-39 调度方式切换



版权所有 © 华为云计算技术有限公司

18.5.2 为应用添加技能

18.5.2.1 配置插件

Agent开发平台提供了一个丰富的插件生态系统，以增强智能体的功能。插件是一种工具集，一个插件即是一个API工具。目前，平台集成了类型丰富的插件，包括文件处理、代码解释器、全网热搜榜、高德地图等工具，这些插件能够帮助开发者快速为智能体添加特定功能。此外，平台还支持创建自定义插件，允许开发者将已有的API能力通过参数配置的方式快速集成到智能体中，进一步丰富其功能。

使用限制

表 18-6 插件限制说明

类别	说明
最大插件数量	最大数量不超过20个
插件URL	URL协议只支持HTTP和HTTPS
请求方法	插件服务的请求方式，POST或GET

添加预置插件

应用支持添加插件技能，可添加“预置插件”和“个人插件”。

如果需要添加“个人插件”，请确保已完成创建个人插件操作。

添加插件的步骤如下：

1. 单击应用，应用配置界面中选择“技能 > 插件”模块，单击“ + ”。
2. 在“添加插件”窗口，选择预置插件，在预置插件右侧单击“ + ”进行添加，再单击“确定”。
3. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

图 18-40 添加预置插件



创建个人插件

1. 单击应用，应用配置界面中选择“技能 > 插件”模块，单击“ + ”。

版权所有 © 华为云计算技术有限公司

2. 在“添加插件”窗口，选择个人插件，在个人插件右侧单击“+”进行添加，再单击“确定”。
3. 若还未创建个人插件，可在右上角单击“创建插件”，创建后再添加。
4. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。创建插件详细过程请参见**插件开发与配置**。

图 18-41 添加个人插件



移除插件

1. 在“技能 > 插件”模块，在已添加的插件右侧单击 。
2. 页面提示插件删除成功则表示插件已移除。

图 18-42 移除插件



18.5.2.2 配置工作流

工作流是Agent开发平台中用于设计和实现复杂任务自动化的核心工具，它通过任务编排、条件判断以及多种组件的协同功能，帮助开发者高效处理复杂任务。工作流中包含大模型节点、知识检索节点、意图识别节点、判断节点、代码节点等多种节点，每个节点都具有特定的功能，能够处理数据、执行任务和运行算法。通过可视化设计，开发者可以清晰地看到数据的流转过程和任务的执行顺序，从而完成复杂的Agent任务编排。

添加工作流

应用支持添加工作流技能。工作流支持通过画布编排的方式，使用插件、大模型等不同节点的组合，从而实现复杂、稳定的业务流程编排。

如果需要添加工作流，请确保已完成编排工作流操作。

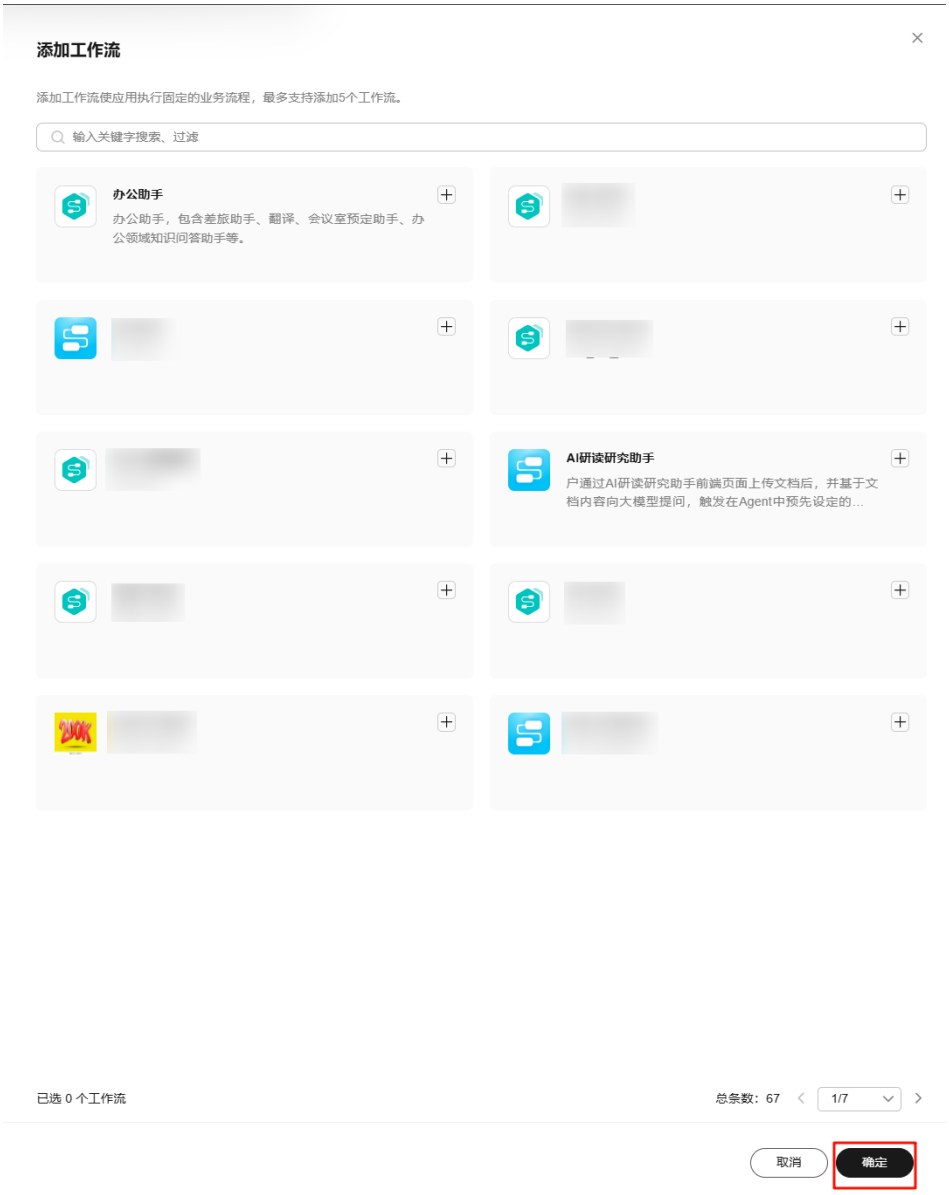
添加工作流的步骤如下：

1. 在“技能 > 工作流”模块，单击“+”。

 版权所有 © 华为云计算技术有限公司

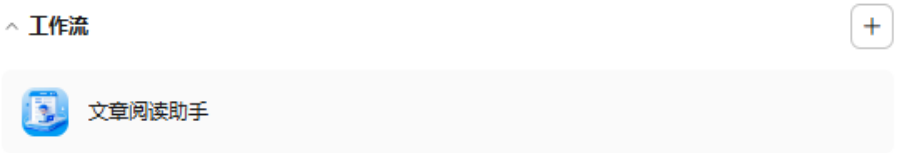
2. 在“添加工作流”窗口，单击“+”进行添加，再单击“确定”。

图 18-43 添加工作流



3. 添加插件工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 18-44 已添加工作流

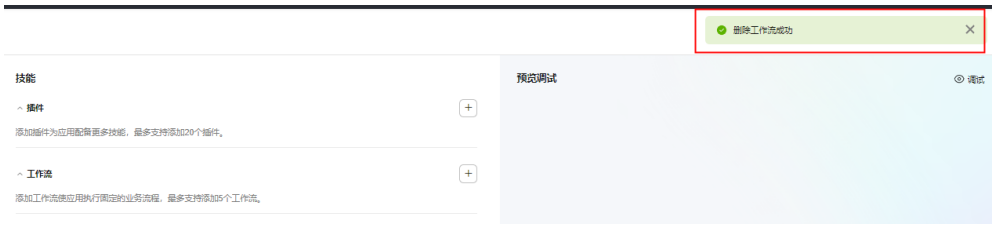


版权所有 © 华为云计算技术有限公司

移除工作流

1. 在“技能 > 工作流”模块，在已添加的工作流右侧单击 。
2. 页面提示“删除工作流成功”则表示工作流已移除。

图 18-45 移除工作流



18.5.2.3 配置知识库

知识库是Agent开发平台中用于存储、管理和检索领域知识的核心组件，它通过结构化存储、智能检索以及动态更新机制，为Agent提供精准的信息支持。知识库支持doc, pdf, pptx, xlsx, csv等多种格式上传，通过多源知识融合和向量化处理，知识库可实现对复杂语义的理解和推理，从而为Agent的决策、问答和任务执行提供可靠的知识支撑。开发者能够灵活配置知识来源、更新策略和检索方式，确保Agent在不同场景下快速调用精准信息，完成智能化服务。

使用限制

表 18-7 知识库限制说明

类别	说明
最大知识库数量	最大数量不超过10个
知识库大小	单个文档上传限制最大128M

添加知识库

应用支持添加知识库。发送消息时，应用能够引用知识库中的内容回答用户问题，当前仅支持关联1个知识库。

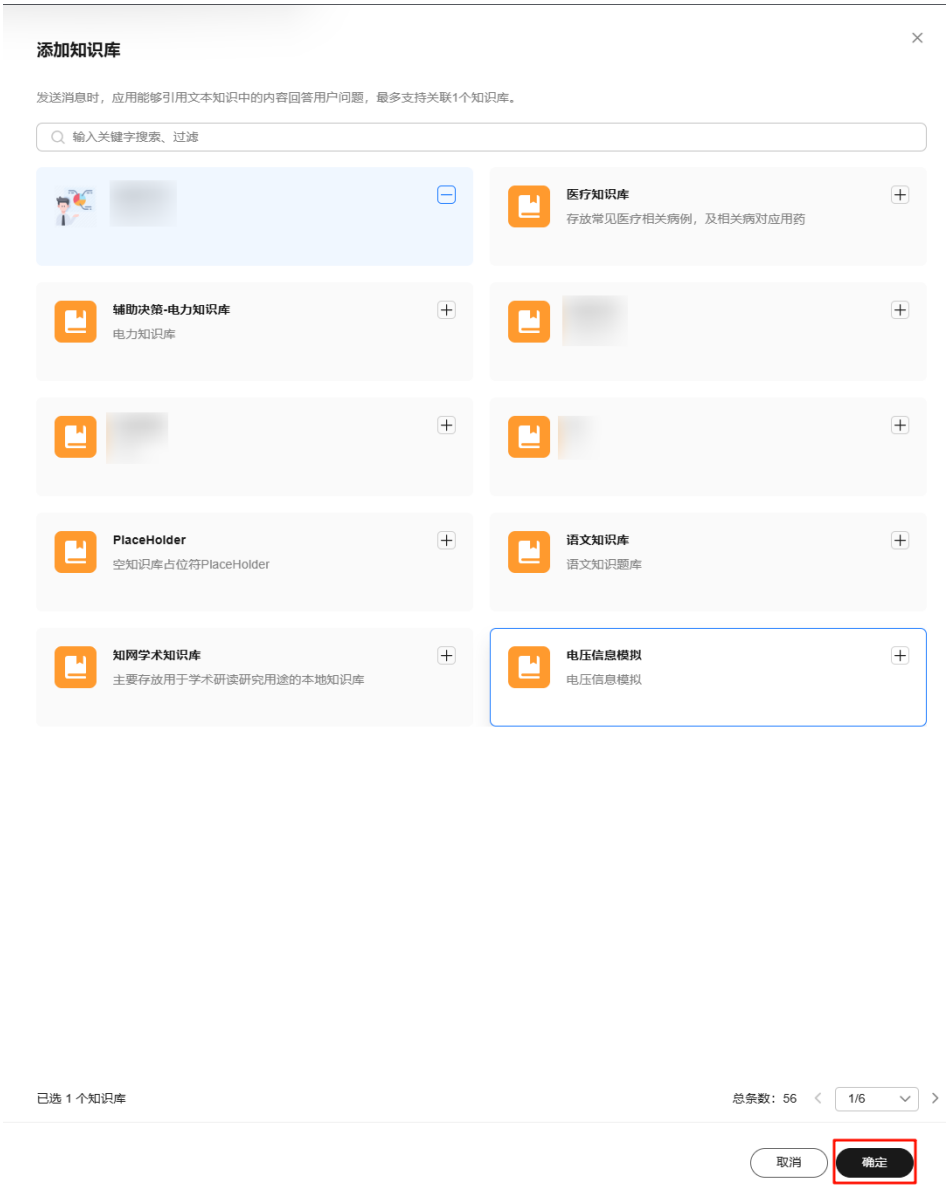
如果需要添加知识库，请确保已完成创建知识库操作。

添加知识库的步骤如下：

1. 在“知识库”模块，单击“ ”。
2. 在“添加知识库”窗口，单击“ ”进行添加，再单击“确定”。

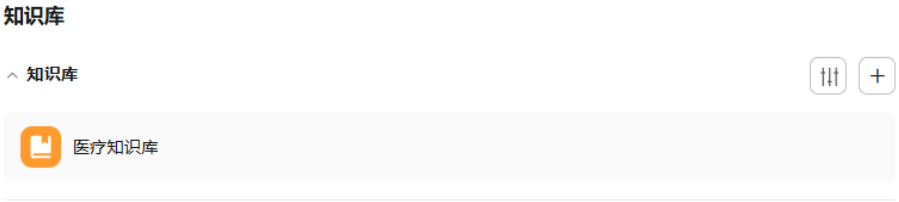
版权所有 © 华为云计算技术有限公司

图 18-46 添加知识库



3. 添加知识库后，可在“知识库”中查看当前已添加的知识库。

图 18-47 已添加知识库



知识库命中测试

平台支持对创建的知识库进行命中测试，以评估知识库的效果和准确性。

版权所有 © 华为云计算技术有限公司

知识库命中测试步骤如下：

1. 在左侧导航栏中选择“ 知识库”页面，单击所需知识库，进入知识库基本信息页面，单击右上角“命中测试”。
2. 在文本框中输入问题，单击“命中测试”，页面下方将根据不同的检索方式，展示多条匹配的内容，并按照匹配分值降序排列。

   用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。

知识库召回策略

1. 在“知识库”模块，单击“ ”对知识库进行高级配置，包括检索策略、相关度阈值与topk召回数量。
   - 检索策略：知识库召回时使用的不同检索技术。支持检索技术如下：
     - 语义检索：使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。
     - 关键词检索：使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。
     - 混合检索：使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。
   - 相关度阈值：超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤，可以参考知识库中命中测试的相关度分值调整该阈值。
   - topk召回数量：召回的相关性阈值top切片数量，如topk召回数量为5，则相关性阈值为前5的切片将被召回提交给大模型总结。

 版权所有 © 华为云计算技术有限公司

图 18-48 知识库高级配置



18.5.2.4 配置 MCP 服务

Agent平台工具调用支持MCP协议，并提供了一个丰富的MCP服务生态系统，以增强智能体的功能。MCP是一种开放协议，它规范了应用程序向大语言模型提供上下文的方式，平台集成了"高德地图"、"车票查询工具"、"必应搜索"等多种实用MCP服务，开通后可以一键集成调用。此外，平台还支持创建自定义MCP服务，开发者可依据MCP服务地址快速创建MCP服务。

使用限制

表 18-8 MCP 服务限制说明

类别	说明
最大MCP服务数量	应用中添加MCP服务数量不超过5个
MCP服务地址	仅支持SSE地址服务 • 只支持HTTP和HTTPS • 必须为标准的URL格式 • 对应的IP默认不应为内网

添加预置 MCP 服务

应用支持添加MCP服务，可添加“预置服务”和“个人服务”。

版权所有 © 华为云计算技术有限公司

如果需要添加“个人服务”，请确保已完成创建个人MCP服务操作。

添加预置MCP服务的步骤如下：

1. 在“MCP服务”模块，单击“ ”。
2. 在“添加MCP服务”窗口，选择预置服务，未开通的MCP服务需要单击“立即开通”，完成配置后即可添加，无需配置的MCP服务可单击 进行一键添加，最后需要单击“确定”。
3. 添加MCP服务后，可在“MCP服务”中查看当前已添加的MCP服务。

图 18-49 添加预置 MCP 服务（无需开通/已开通）



图 18-50 添加预置 MCP 服务（未开通）



版权所有 © 华为云计算技术有限公司

添加个人 MCP 服务

1. 在“MCP服务”模块，单击“ ”。
2. 在“添加MCP服务”窗口，选择个人服务，在个人服务右侧单击“ ”进行添加，最后需要单击“确定”。
3. 若还未创建个人MCP服务，需要创建后再添加，创建MCP服务详细过程请参见**MCP服务管理**。
4. 添加MCP服务后，可在“MCP服务”中查看当前已添加的MCP服务。

图 18-51 添加个人 MCP 服务



移除 MCP 服务

1. 在“MCP服务”模块，在已添加的MCP服务右侧单击“ ”。
2. 页面提示“删除MCP服务成功”则表示MCP服务已移除。

图 18-52 移除 MCP 服务



版权所有 © 华为云计算技术有限公司

18.5.3 提升应用对话体验

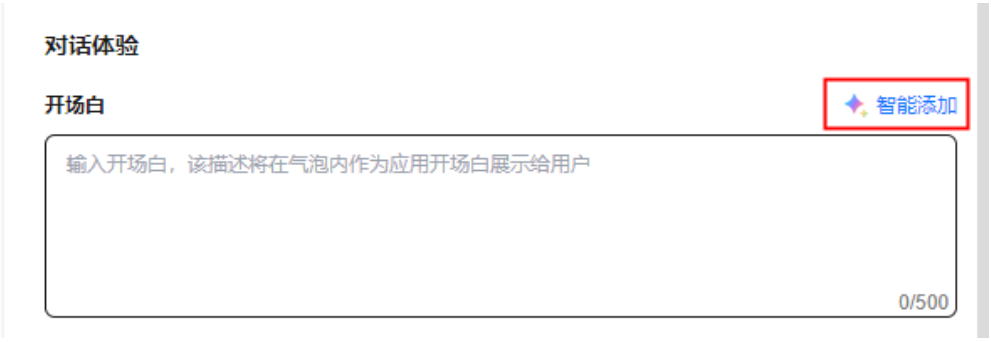
配置开场白

开场白是用户进入单智能体应用后首先看到的引导信息，能够帮助用户迅速了解Agent的功能和用途，明确如何与Agent进行有效交互。开场白需要简洁明了，语气友好并且表述清晰。

- 在“对话体验 > 开场白”中，可填写开场白，也可单击“智能添加”智能添加开场白。

  例如，“您好！我是您的智能助手，很高兴为您提供帮助。请告诉我，今天有什么我可以为您做的吗？”

  图 18-53 配置开场白



配置推荐问题

推荐问题：推荐问题是用户首次与应用互动时，应用主动展示的一些问题或话题建议，提供预设选项，减少用户打字负担。

- 在“对话体验 > 推荐问题”中，可填写推荐问题，也可单击“智能添加”智能添加推荐问题。推荐问题至多配置3条。

  例如，“请告诉我您需要什么帮助？”如：帮我预定会议室、帮我查询天气预报。

  图 18-54 配置推荐问题



版权所有 © 华为云计算技术有限公司

配置追问

追问：在每轮回复后，默认根据对话内容提供提问建议，可以挖掘用户的潜在需求。

- 在“对话体验 > 追问”中，可选择是否开启“追问”功能，若开启，模型在每轮回复后，默认根据对话内容提供提问建议。

图 18-55 配置追问



18.5.4 调试与发布应用

18.5.4.1 调试应用

开发者可以在单智能体应用搭建完成后，直接与应用进行对话，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。平台全链路调试功能，允许开发者查看每条用户请求从输入到响应的完整流程，从而精准定位问题并快速调整配置。

调试应用

创建应用后，平台支持对应用执行过程的进行预览与调试。

调试应用的步骤如下：

1. 登录Agent开发平台。
2. 在左侧导航栏中选择“应用管理”，然后在页面顶部功能列表中选择“单智能体应用”功能进入应用开发主页面。

版权所有 © 华为云计算技术有限公司

3. 选择要调试的应用，单击进入应用详情页面。
4. 在应用详情页面单击右上角模型旁的下拉菜单，弹出**图1 单智能体应用大模型配置**对话框，参考**表18-9**配置大模型参数。

图 18-56 单智能体应用大模型配置



表 18-9 大模型参数配置

参数	说明
模型选择	选择要使用的大模型，不同的模型效果存在差异。
模式选择	用于配置大模型的输出多样性。 包含取值： ● 精确的：模型的输出内容严格遵循指令要求，可能会反复讨论某个主题，或频繁出现相同词汇。 ● 平衡的：平衡模型输出的随机性和准确性。 ● 创意性的：模型输出内容更具多样性和创新性，某些场景下可能会偏离主旨。 ● 自定义：自定义大模型输出的温度和核采样值，生成符合预期的输出。
温度	调高温度会使得模型的输出更多样性和创新性，反之，降低温度会使输出内容更加遵循指令要求但减少多样性，取值范围为0到1之间。 ● 调高温度，会使得模型的输出更多样性和创新性。 ● 降低温度，会使输出内容更加遵循指令要求但减少多样性。 在基于事实的问答场景，可以使用较低的回复随机性数值，以获得更真实和简洁的答案；在创造性的任务例如小说创作，可以适当调高回复随机性数值。建议不要与核采样同时调整。
核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值。核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性，取值范围为0.1到1之间。

5. 在右侧“预览调试”的文本框中输入对话，应用将根据对话生成相应的回答。
6. 在调试过程中，单击右上角“调试”，可以查看当前会话或历史会话的运行结果与调用详情。

版权所有 © 华为云计算技术有限公司

图 18-57 查看调试结果



调试信息说明

在调试台中有两个页面：运行结果页面、调用详情页面。

- 运行结果

  运行结果中可以看到应用的执行开始时间、结束时间、运行时间等信息，还能看到输入和输出信息。对于性能的情况有个直观的认识。



 版权所有 © 华为云计算技术有限公司

- 调用详情

  在触发应用时，调用链中展现具体事件的详细信息，包括触发的组件、事件耗时、事件的输入和输出信息等。便于开发者快速的追溯操作顺序并精准定位问题。



18.5.4.2 发布应用为 API 服务

将应用发布为API服务后，可以通过调用OpenAPI的方式使用Agent程序。本文介绍如何将开发完成应用发布为API服务。

发布应用为API的详细步骤：

1. 登录Agent开发平台。
2. 在左侧导航栏中选择“应用管理”，然后在页面顶部功能列表中选择“单智能体应用”功能进入应用开发主页面。

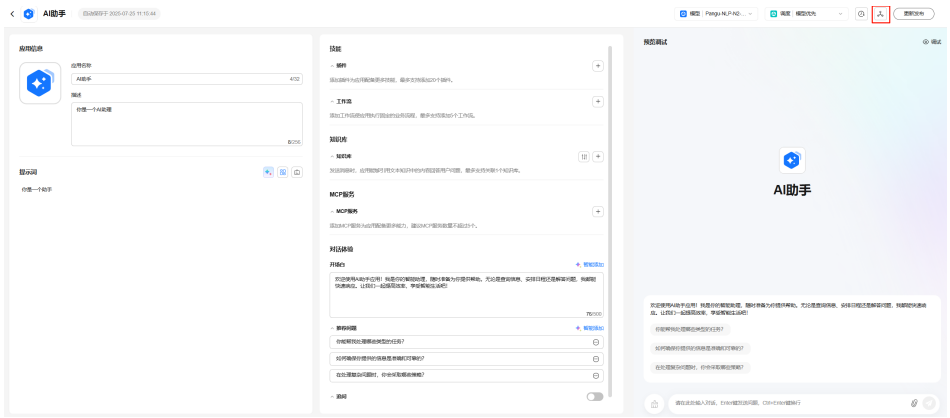
版权所有 © 华为云计算技术有限公司

图 18-58 应用管理



3. 在应用编辑页面完成该应用的功能编辑调试，然后单击右上角的“发布”（首次发布）或者“更新发布”（非首次发布）按钮。在弹出的发布信息填写提示窗中填写本发布的版本号、描述信息，最后单击“发布”或“更新发布”按钮完成发布。

图 18-59 编辑应用



4. 发布完成后进入**应用主页面**，单击右上角“发布渠道”按钮  ，进入**发布管理**页面。

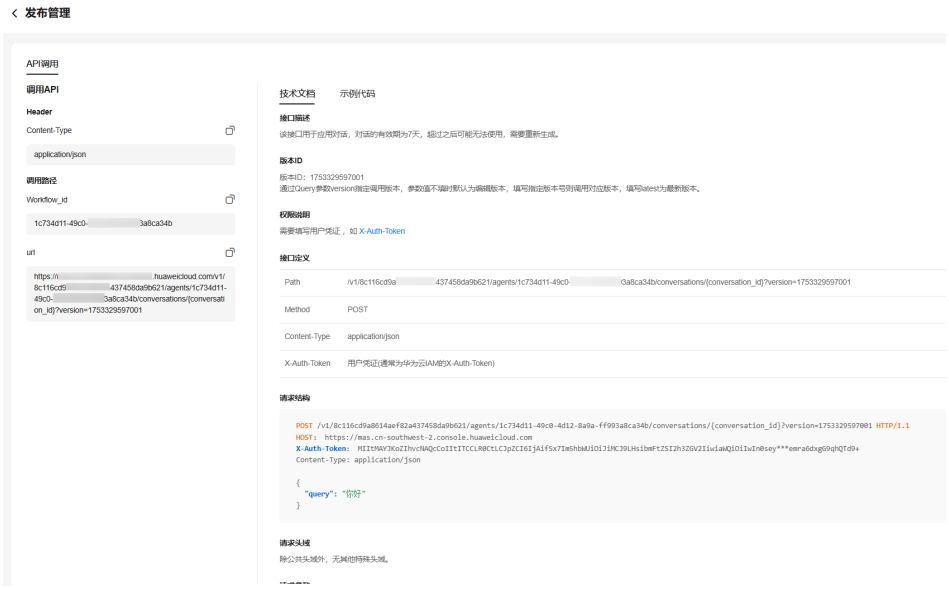
图 18-60 发布渠道



5. 在**发布管理**页面，选择“API调用”功能选项卡即可看到发布的API调用接口信息。

版权所有 © 华为云计算技术有限公司

图 18-61 获取 API 调用信息



18.5.5 使用 API 调用应用

Agent开发平台支持将创建好的应用进行API调用。在调用应用前，请先参考**基础配置**，完成创建应用操作。

获取调用路径

应用的调用路径获取步骤如下：

1. 登录Agent开发平台。
2. 在左侧导航栏中选择“应用管理”，然后在页面顶部功能列表中选择“单智能体应用”功能进入应用开发主页面，单击所需应用的“ ··· > 调用路径”。

图 18-62 获取应用调用路径



版权所有 © 华为云计算技术有限公司

3. 在“调用路径”页面，单击“复制路径”即可获取调用路径。
   其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

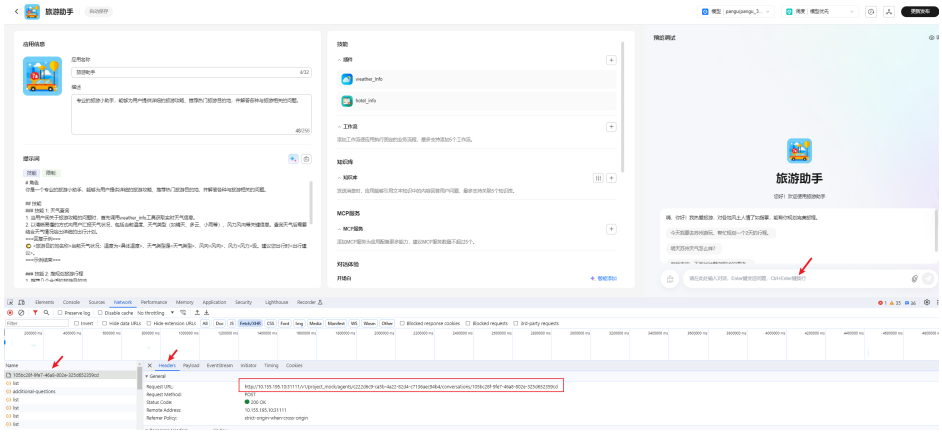
图 18-63 调用路径详情



使用 Postman 调用 API

1. 获取调用接口。
   a. 在应用页签，单击需要调用的应用。
   b. 按键盘F12，进入开发者工具。在页面右侧“预览调试”中输入并发送问题。在开发者工具中查找本次问答所触发的接口响应，参考下图示例，在接口的Header请求中获取API接口。

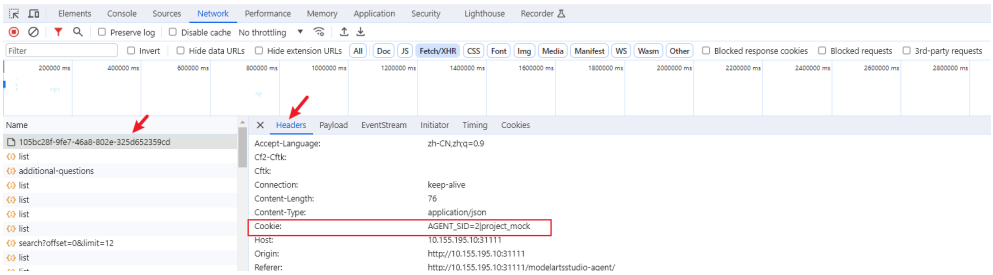
图 18-64 获取调用接口



2. 获取接口Cookie。
   按照步骤1方法找到API接口后，在Header请求中找到Cookie。

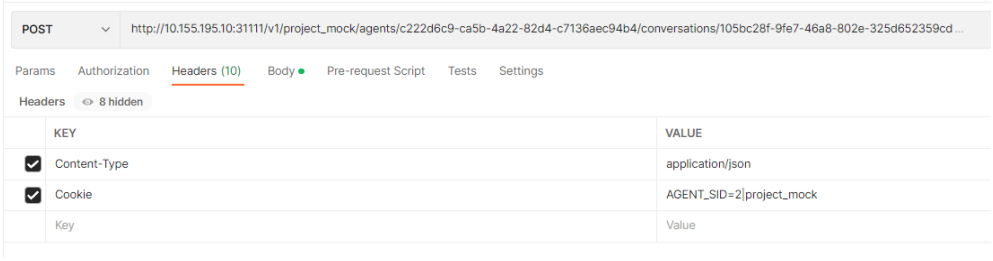
 版权所有 © 华为云计算技术有限公司

图 18-65 获取 Cookie



3. 在Postman中新建POST请求，并填入应用的API接口。
4. 填写请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为Cookie，参数值获取方法参考**步骤2**。

图 18-66 填写请求 Header



5. 在Postman中选择“Body > raw”选项，请求Body填写示例如下。其中，query参数为用户提出的问题，作为应用的输入。

```
{
  "query": "介绍下北京的景点"
}
```

6. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示应用API调用成功，响应示例如下：

```
data:{"event":"start","createdTime":1744636611503}

data:{"event":"message","content":"北京","createdTime":1744636611816}

data:{"event":"message","content":"是","createdTime":1744636611854}

data:{"event":"message","content":"中国的","createdTime":1744636611890}

data:{"event":"message","content":"首都","createdTime":1744636611925}

...

data:{"event":"message","content":"随时","createdTime":1744636624119}

data:{"event":"message","content":"告诉我","createdTime":1744636624154}

data:{"event":"message","content":"。","createdTime":1744636624190}

data:{"event":"statistic_data","latency":{"overall":13.17},"createdTime":1744636624265}

data:{"event":"summary_response","content":"北京是中国的首都，拥有丰富的历史文化遗产和现代化的城市景观。以下是一些著名的旅游景点推荐：\n\n### 历史文化景点\n- **故宫**\n  - 特色: 世界上最大的宫殿复合体，明清两代的皇宫，展示了中国古代的建筑艺术和皇家文化。\n  - 建议: 建议预留至少半天时间参观，最好提前在线购票以避免排队。\n\n- **天安门广场**\n  - 特色: 世界上最大的城市广场，周围有人民英雄纪念碑、毛主席纪念堂等重要建筑。\n  - 建议: 早晨或傍晚时分参观，可以避开人流高峰。\n\n- **颐和园**\n  - 特色: 中国保存最完整的皇家园林，以昆明湖和万寿山为基础，融合了江南园林的设计风格。\n  - 建议: 适合全家出游，可以乘坐游船游览昆明湖。\n\n### 现代景点\n- **鸟巢和水立方**\n  - 特色: 2008年北京奥运会的主要场馆，现在成为了北京的标志性建筑。\n  - 建议: 晚上参观可以看到灯光秀，非
```


版权所有 © 华为云计算技术有限公司

常壮观。\n\n- **798艺术区**\n - 特色: 集合了众多艺术展览和创意工作室，是艺术爱好者的天堂。\n - 建议: 适合下午或傍晚时分参观，可以在此享受艺术氛围。\n\n- **三里屯**\n - 特色: 北京最时尚的购物和娱乐区，有众多国际品牌店和餐厅。\n - 建议: 适合晚上前往，体验北京的夜生活。\n\n希望这些信息对您的旅行计划有所帮助！如果您需要更详细的行程规划或其他信息，请随时告诉我。
","role":"assistant","createdTime":1744636624266}

data:{"event":"done","createdTime":1744636624276}

18.6 多智能体应用开发（Beta）

18.6.1 多智能体应用介绍

Agent可以灵活应用各种工作流程完成用户任务，且支持根据用户意图在工作流间跳转。

多智能体支持**模型自动控制**和**指定顺序**两种模式。

- 模型自动控制：由模型根据用户输入决定调用哪个工作流。
- 指定顺序模式：根据用户指定工作流顺序一次执行工作流。

单智能体与多智能体功能与应用场景差异

单智能体：依赖模型，可以使用插件、工作流、知识库、MCP服务等工具，让模型自主规划，使用不同工具完成指定任务。

多智能体：可配置多个工作流，侧重根据客户意图在不同工作流中进行选择和跳转。

常见问题

单智能体模式、多智能体模式是否可以切换?

单智能体和多智能体之间不可切换，Agent创建时确定。

工作流和多智能体模式有什么区别?

工作流是多智能体的组成元素，多智能体有一个控制器可以执行多个工作流并根据用户输入意图在不同工作流之间跳转。

工作流只是单工作流的运行。

在单智能体模式中使用工作流，与在多智能体中使用工作流有什么区别?

1. 多智能体有全局意图，可理解为公共意图，如果在工作流的其他意图上勾选可覆盖，工作流执行到该意图识别节点可以被识别为全局意图而跳转到全局意图对应的动作中。单智能体没有这个能力。

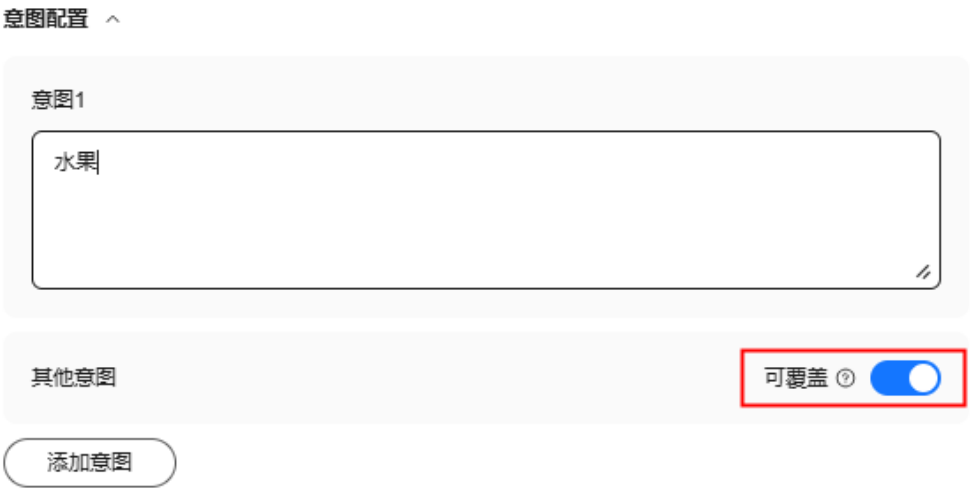
图 18-67 多智能体全局意图



 版权所有 © 华为云计算技术有限公司

2. 工作流意图节点，其他意图可覆盖配置。

图 18-68 其他意图可覆盖配置



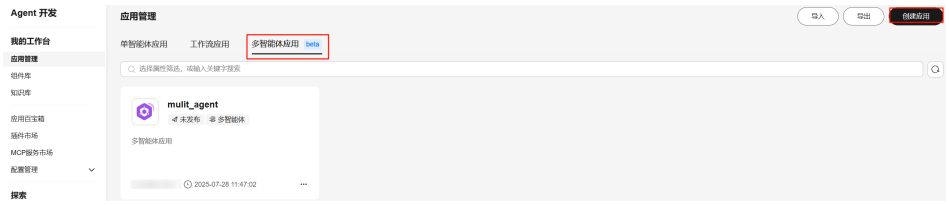
3. 多智能体控制工作流执行逻辑更灵活，支持模型控制和指定顺序模式，模型控制由模型规划调用哪个工作流，指定顺序则按指定工作流顺序执行工作流。单智能体只能由模型决定是否调用模型和调用哪个模型。

18.6.2 创建多智能体应用

创建多智能体应用主要步骤如下：

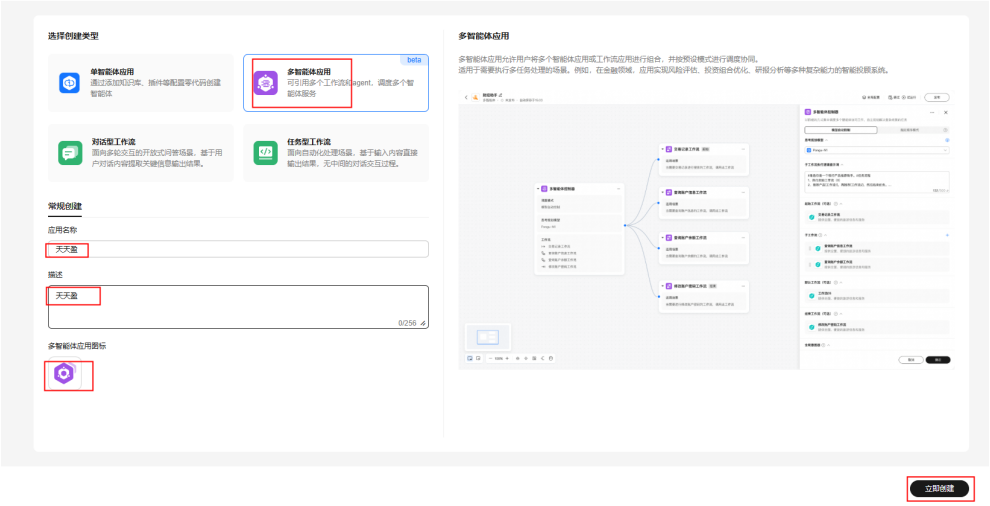
1. **创建多智能体应用。**
   - 登陆Agent开发平台；
   - 在页面左上方选择我的工作台->应用管理，单击右上角创建应用。

图 18-69 创建多智能体应用



   - 进入应用创建页面，选择“多智能体应用（beta）”，填写应用名称，描述，选择应用图标单击右下角立即创建。

版权所有 © 华为云计算技术有限公司



- 进入多智能体编辑页面，初始只有一个多智能体控制器节点。

2. **设置全局配置。**

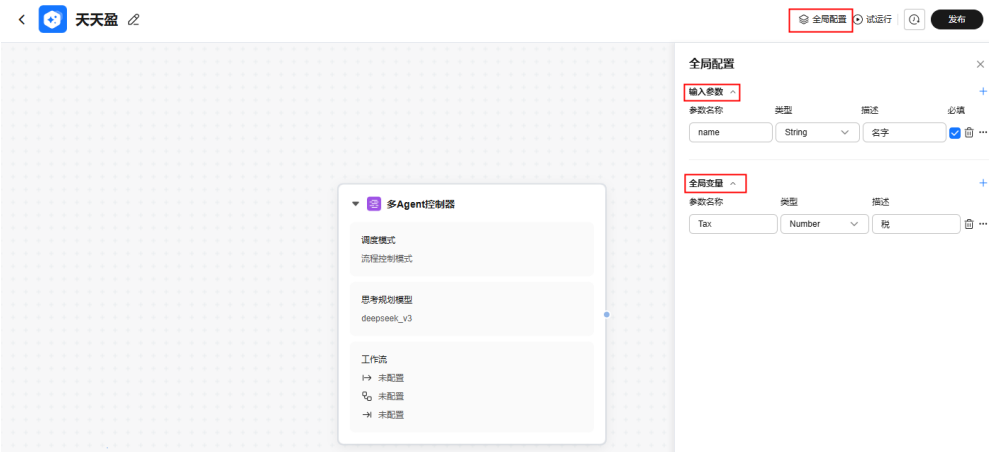
   全局配置可配置输入参数和全局变量，这些参数可作为后续工作流的输入参数以及全局参数，供后续业务配置使用。

   输入参数：可配置String,Integer,Number,Boolean4种类型，会传给工作流的输入参数，且值不可被修改。

   全局参数：可配置String,Integer,Number,Boolean4种类型，会传给工作流的输入参数，工作流如果有相同名称和类型的输出参数会改写该值。

   进入多智能体编辑页面，单击右上角**全局配置**，可配置输入参数和全局变量。

   图 18-70 全局配置



3. **设置多智能体控制器。**

   可参考**表18-10**配置多智能体控制器节点。

版权所有 © 华为云计算技术有限公司

表 18-10 多智能体控制器节点配置说明

配置类型	参数名称	参数说明
控制器模式配置	-	● 模型自动控制：由模型根据用户输入识别用户意图，选择不同的工作流执行解决用户问题模型自动控制需要选择**模型**和配置**逻辑提示词**引导模型进行正确的规。 ● 指定顺序模式：根据接口指定工作流id列表顺序执行工作流，无需配置模型和提示词。
模型配置	模型选择	模型自动控制模式下，需要选择模型用于识别用户意图，进行执行规划，选择子工作流或全局意图。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
子工作流执行逻辑提示词	-	模型自动控制模式下，需要填写逻辑提示词，指导模型更好地完成任务。
工作流配置	起始工作流	多智能体每次执行开始都会执行的工作流，用于开场白，身份核验之类的开始对话任务。
	子工作流	业务工作流，实际完成业务述求的工作流，比如在推销产品类的agent中可能是不同产品推销流程的工作流。
	默认工作流	模型自动控制模式下，根据用户输入未匹配到任何子工作流时，使用默认工作流。
	结束工作流	多智能体每次执行结束都会执行的工作流，用于表示感谢和再见之类的结束对话任务。
全局意图	-	在与Agent交互过程中，用户可能有一些与业务无关的公共意图，比如"不感兴趣"、"非本人"等，这些意图可以配置到全局意图，并可配置该意图对应的动作。 ● "直接应答"：配置一段文本，输出给用户。 ● "流程跳转"：关联一个工作流完成对应意图需要执行的动作。
高级配置	最大对话历史轮次	选择是否打开历史对话引用功能，默认为0即不会引用对话历史，配置N轮即可记录N轮对话的内容。
	最大跳转次数	多智能体运行过程中，根据用户意图，可能在多个工作流之间跳转，该参数可限制最大跳转次数。

 版权所有 © 华为云计算技术有限公司

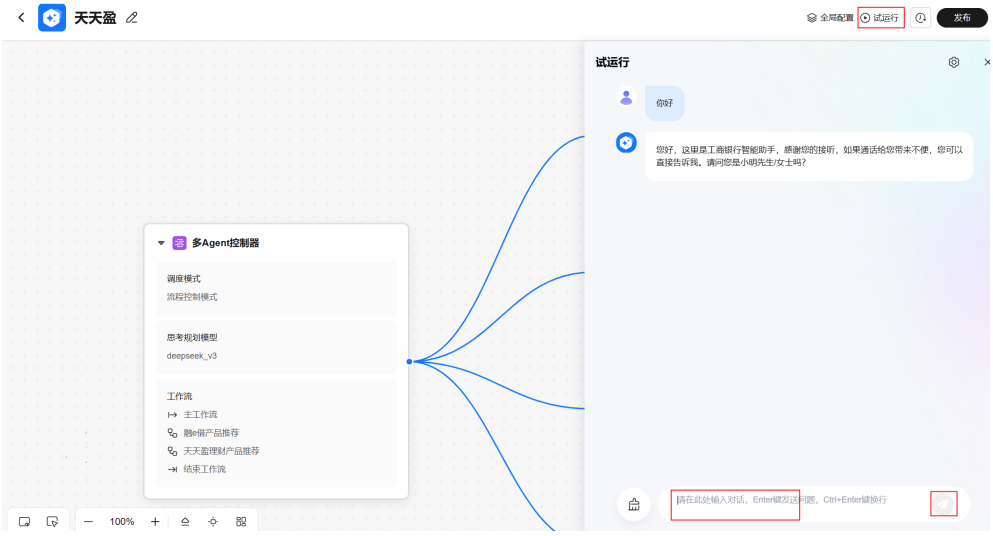
18.6.3 调试与发布多智能体应用

开发者可以在多智能体搭建完成后，直接与多智能体进行对话，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。

创建应用后，平台支持对应用进行预览与调试。

调试应用的步骤如下：

在页面右上角，单击**试运行**进入调试页面，在下方输入对话内容和agent进行对话



18.6.4 使用 API 调用多智能体应用

Agent开发平台支持将创建好的应用进行API调用。在调用应用前，请先参考**创建多智能体应用**，完成创建多智能体应用操作。

获取调用路径：

1. 登录Agent开发平台。
2. 在“我的工作台 > 应用管理->多智能体应用”页面，单击所需应用的 *** > 调用路径。

版权所有 © 华为云计算技术有限公司

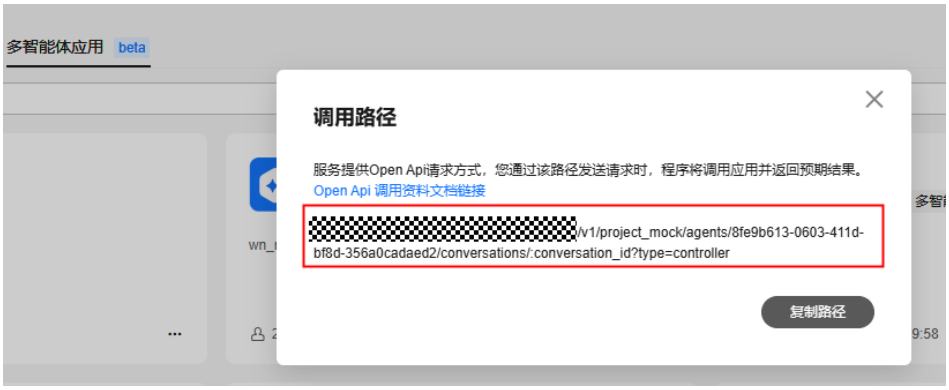
图 18-71 获取应用调用路径



3. 在“调用路径”页面，单击“复制路径”即可获取调用路径。

   其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 18-72 应用调用路径详情页面



调用 API

1. 请求样例

```
curl 'http://ip:port/v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}?type=controller' \
  -H 'Cache-Control: no-cache' \
  -H 'Connection: keep-alive' \
  -H 'Content-Type: application/json' \
  -H 'x-user-id=:user_id' \
  -H 'x-project-id=:project_id' \
  -H 'stream: true' \
  --data-raw '{"inputs":{"name":"小明","query":"你好"}}' \
  --insecure
```


版权所有 © 华为云计算技术有限公司

说明

上面"{}"中均为变量，需要根据实际取值填写。

project_id：Agent归属projectid，参考**获取项目ID**。

agent_id: 实际要运行的多智能体id。左侧导航栏选择“应用管理 > 多智能体应用（beta）”，在所需Agent中单击“... > 复制ID”。可获取多智能体ID。

conversation_id：会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

user_id：指定访问用户ID。用户ID可在控制台右上角单击当前账号 >，在弹出的对话框选择“我的凭证 > IAM用户ID”获取 。

2. 响应样例

```
data:{"event":"message","data":{"text":"您好，这里是工商银行智能助手，感谢您的接听，如果通话给您带来不便，您可以直接告诉我。请问您是小明先生/女士吗?
","index":0,"node_id":"node_1747035601714","node_type":"Questioner","workflow_id":"17aa0ef8-7fbb-4a93-acf1-172d7a10524b"}}
data:{"event":"message","data":
{"text":"","node_id":"node_1747035601714","node_type":"Questioner","is_finished":true,"workflow_id":"17aa0ef8-7fbb-4a93-acf1-172d7a10524b"}}
data:{"event":"end"}
```

18.7 工作流应用开发

18.7.1 工作流介绍

工作流是一系列相互关联的步骤，用于实现业务逻辑或完成特定任务。它为应用/智能体的数据流动和任务处理提供了一个结构化框架。

平台提供一个可视化画布，可以通过拖拽节点迅速搭建工作流。同时，支持在画布实时调试工作流。在工作流画布中，可以清晰地看到数据的流转过程和任务的执行顺序。

工作流节点介绍

Agent开发平台的工作流由多个节点构成，节点是组成工作流的基本单元。平台支持多种节点，包括开始、结束、大模型、意图识别、提问器、插件、判断、代码、知识检索和消息节点等节点。

创建工作流时，每个节点需要配置不同的参数，如输入和输出参数等，开发者可通过拖、拉、拽可视化编排更多的节点，实现复杂业务流程的编排，从而快速构建应用。

工作流方式主要面向目标任务包含多个复杂步骤、对输出结果成功率和准确率有严格要求的复杂业务场景。

在编排工作流时，可以使用以下节点进行功能设计：

- 开始节点：开始节点是工作流的起始节点，用户输入的信息由开始节点传入。
- 结束节点：结束节点是工作流的最终节点，用于定义整个工作流的输出信息。
- 大模型节点：用于在工作流中引入大模型能力。
- 意图识别节点：用于根据用户的输入进行意图分类并导向后续不同的处理流程。
- 提问器节点：提供了在对话过程中向用户收集更多信息的能力。
- 插件节点：用于引入API插件，根据节点的输入，执行用户定义的插件，将插件执行结果作为节点的输出。

 版权所有 © 华为云计算技术有限公司

- 判断节点：编排应用时作为分支切换节点，可以根据输入满足的判断条件，指定执行对应的工作流分支。
- 代码节点：用于引入代码执行器，根据节点的输入，执行指定Python代码，节点的输出是代码执行的结果信息。
- 知识检索节点：可以根据输入参数从指定知识库内召回匹配的信息。
- 消息节点：定义一段文本内容，在工作流的执行过程中向用户发送该内容的消息。
- 循环节点：循环节点用于重复执行一系列任务。
- 变量赋值节点：用于向可写入变量进行变量赋值，实现数据的动态更新和传递。
- 变量聚合节点：变量聚合节点能够将多路分支的输出变量整合为一个，方便下游节点统一配置。
- 输入节点：输入节点用于在工作流运行时收集用户输入。
- 工作流节点：实现工作流嵌套工作流的效果。
- MCP服务节点：MCP服务节点是工作流中实现第三方能力调用的核心组件之一。作为功能扩展的重要载体，该节点允许通过调用MCP服务来执行特定功能任务。每个MCP服务实质上是一个工具集合，可以提供模块化服务来拓宽工作流的能力边界，完成更复杂的任务。
- Agent节点：Agent节点提供了使用大模型的能力以及大模型工具调用的能力，可在节点中配置已部署的模型，用户可以通过编写Prompt、绑定插件让模型处理相应任务。

18.7.2 对话型工作流和任务型工作流

平台提供了两种类型的工作流，即对话型工作流和任务型工作流，用户可以针对不同的任务或场景选择适合的工作流进行搭建。

对话型工作流、任务型工作流区别

工作流包含两种类型：

- 对话型工作流。面向多轮交互的开放式问答场景，基于用户对话内容提取关键信息，输出最终结果。适用于客服助手、工单助手、娱乐互动等场景。
- 任务型工作流。面向自动化处理场景，基于输入内容直接输出结果，无中间的对话交互过程。适用于内容生成、批量翻译、数据分析等场景。

说明

任务型工作流不支持配置**输入节点**、**消息节点**、**Agent节点**和**提问器节点**。

版权所有 © 华为云计算技术有限公司

图 18-73 任务型工作流

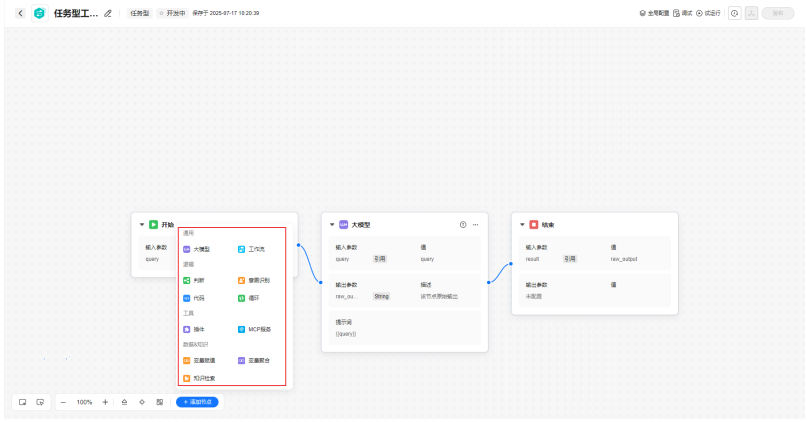


表 18-11 区别说明

差异项	对话型工作流	任务型工作流
适用场景	客服助手、工单助手、娱乐互动等场景。	内容生成、批量翻译、数据分析等场景。
节点	支持输入节点、消息节点、提问器节点和Agent节点。	不支持输入节点、消息节点、提问器节点和Agent节点。
试运行方式	试运行界面与任务型工作流不同。 若“开始”节点有多个参数，先对除query参数外的参数进行配置，然后再以对话窗口的形式进行试运行。	若“开始”节点有多个参数，在试运行时，需要对多个输入参数同时进行配置。

常见问题

对话型工作流、任务型工作流是否可以相互转换?

答：不可以。

18.7.3 创建工作流

工作流是一系列相互关联的步骤，用于实现业务逻辑或完成特定任务。可以在智能体和应用搭建中通过工作流实现特定的任务或指令。无论是在智能体还是应用中使用工作流，都需要先创建一个可运行的工作流。

步骤 1：创建工作流

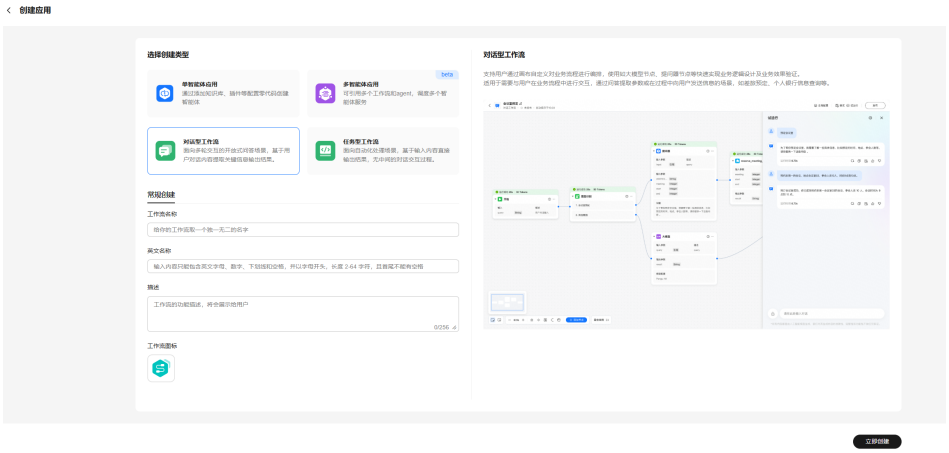
1. 登录Agent开发平台。
2. 单击左侧导航栏“应用管理”，在“工作流”页签，单击右上角“创建应用”。
3. 在“创建应用”页面选择对话型工作流/任务型工作流，填写工作流名称、英文名称和工作流描述。
4. 单击页面左下角的图片可修改工作流图标，单击“确定”，进入工作流编排页面。

版权所有 © 华为云计算技术有限公司

初始状态下工作流包含**开始**节点和**结束**节点。

- **开始**节点用于启动工作流。
- **结束**节点用于返回工作流的运行结果。

图 18-74 创建工作流



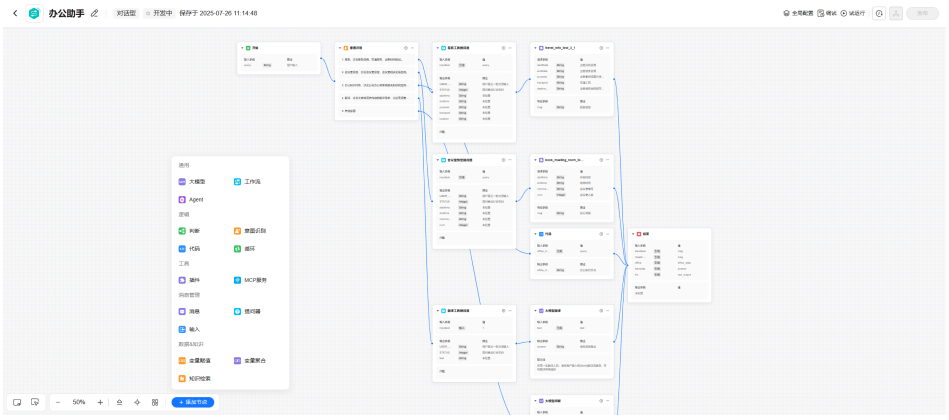
步骤 2：编排工作流

创建工作流后，在画布中添加节点，并按照任务执行顺序连接节点。

工作流内置了多种基础节点供你使用，同时还可以添加插件节点来执行特定任务。

1. 在左侧面板中选择要使用的节点（版本不同，节点可能会有差异，以当前的版本为准）。
2. 将各个节点相连接。
3. 配置节点的输入和输出参数。

图 18-75 编排工作流



步骤 3：调试并发布工作流

要想在应用中使用该工作流，则需要发布工作流。

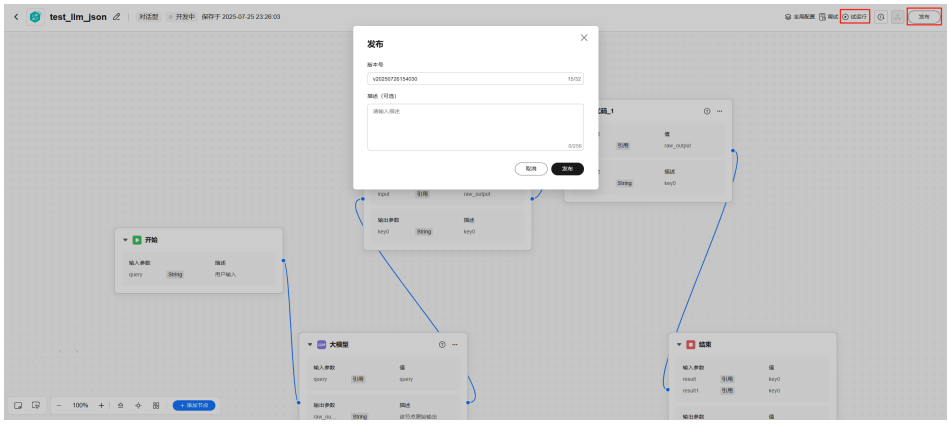
1. 单击“试运行”。

运行成功的节点边框会显示绿色，在各节点的右上角可查看节点的输入和输出。

版权所有 © 华为云计算技术有限公司

2. 单击“发布”。

图 18-76 发布工作流



管理工作流

在工作流的管理界面，可以进行工作流的复制、复制ID、调用路径、删除等操作。

复制工作流

- 在工作流列表页面，单击所要复制工作流右下角的 "..." 图标，再单击“复制”可以将该工作流复制到您的工作流列表中。

图 18-77 管理工作流



- 不支持跨画布复制节点。

版权所有 © 华为云计算技术有限公司

复制ID

在工作流列表页面，单击工作流右下角的 "..." 图标，再单击“复制ID”，即可完成复制ID操作。

工作流接口调用方法请参考**通过API调用工作流**。

在工作流列表页面，单击工作流右下角的 "..." 图标，再单击“调用路径”，在新的对话框中再单击“调用路径”即可完成调用路径的复制操作。



删除工作流

对于不需要使用的工作流，可以在**工作流**列表内找到该工作流，单击右上角的 "..." 图标，再单击“删除”，即可删除该工作流。

说明

如果工作流已添加至应用中，在删除时会同步删除应用中的工作流。

18.7.4 配置单智能体应用工作流

可通过工作流功能来实现单智能体应用的业务逻辑部分。工作流决定了应用的输入和输出的数据结构、接收和处理数据的规则以及决策流程，是单智能体应用的核心部分。

单智能体应用支持添加工作流技能。工作流支持通过画布编排的方式，使用插件、大模型等不同节点的组合，从而实现复杂、稳定的业务流程编排。

- **简单业务**：应用中应至少有一个试运行通过的工作流，作为应用的业务处理流程。
- **复杂任务**：如果你的业务逻辑复杂由多个子任务组成，您可以将其拆分为多个工作流，每个工作流负责完成其中的一个任务，再将这些工作流组合到一个工作流中。

单智能体应用中配置工作流的操作步骤如下：

如果需要添加工作流，请确保已完成编排工作流操作。

1. 在“技能 > 工作流”模块，单击“ + ”。
2. 在“添加工作流”窗口，单击“ + ”进行添加，再单击“确定”。

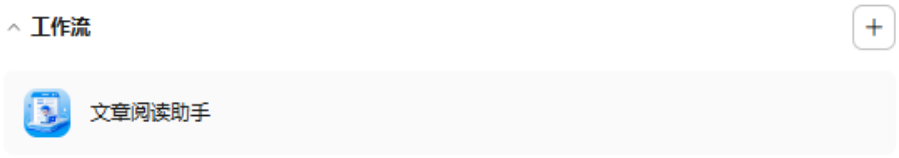
版权所有 © 华为云计算技术有限公司

图 18-78 添加工作流



3. 添加插件后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 18-79 已添加工作流



18.7.5 配置多智能体应用工作流

可通过工作流功能来实现多智能体应用的业务逻辑部分。工作流决定了应用的输入和输出的数据结构、接收和处理数据的规则以及决策流程，是智能体应用的核心部分。

多智能体应用通过添加工作流技能，对用户的会话进行意图分析，路由到不同的子工作流执行编排好的任务。

多智能体应用中配置工作流的操作步骤如下：

如果需要添加工作流，请确保已完成编排工作流操作。

1. 在“多Agent控制器 > 子工作流”模块，单击“ + ”。

 版权所有 © 华为云计算技术有限公司

2. 添加工作流后，单击下拉框可选择已发布的工作流版本应用。

图 18-80 添加工作流



3. 添加工作流保存后，可在画布中查看当前已添加的工作流。

图 18-81 已添加工作流



4. 添加起始、默认、结束工作流，均通过单击下拉框选择后保存

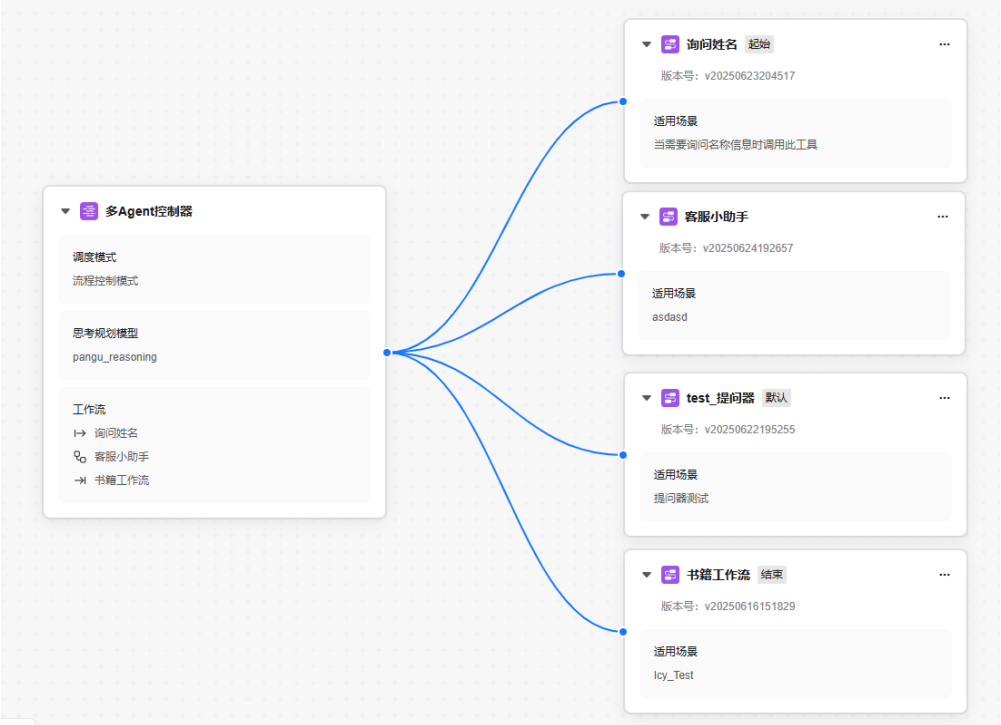
版权所有 © 华为云计算技术有限公司

图 18-82 添加起始、默认、结束工作流



 版权所有 © 华为云计算技术有限公司

图 18-83 编排后包含工作流的多智能体应用



18.7.6 配置聊天记忆

全局配置模块可以配置全局记忆变量

版权所有 © 华为云计算技术有限公司

图 18-84 全局配置



配置记忆变量的步骤如下：

参考**表18-12**完成记忆参数配置。

版权所有 © 华为云计算技术有限公司

表 18-12 记忆参数配置说明

参数名称	参数说明
参数名称	记忆参数的名称。
默认值	可选。
记忆时长	• 如果选择“长期”，用户在记录参数值后将长期保存。 • 如果选择“会话”，当会话结束后记录的参数值将自动恢复为默认值，如果没有设定默认值，则为空。
描述	记忆参数的描述。

记忆参数配置完成后，单击“确定”。

18.7.7 调试与发布工作流

试运行工作流

Agent开发平台支持对整个工作流进行试运行，也支持对工作流的单个节点进行调试。

- 试运行工作流：
  a. 工作流编排完成后，单击右上角“试运行”，在对话框中输入问题，等待返回试运行结果。
  b. 在试运行过程中，可以单击右上角“调试”查看调试结果，包括运行结果与调用详情。

图 18-85 调试结果示例



- 单节点调试，以调试“意图识别”节点为例：

版权所有 © 华为云计算技术有限公司

a. 在工作流编排页面，单击意图识别节点的“”，进入单节点调试页面。
b. 编写输入参数内容，单击“开始运行”。

图 18-86 编写输入参数内容



c. 可在“运行结果”页面查看当前节点的运行结果。
d. 若运行成功，节点处也将显示“运行成功”字样。

图 18-87 单节点调试运行成功示例



发布工作流

工作流试运行成功后，可对其进行发布，便于后续使用。

1. 在工作流编排页面，单击右上角“发布”，输入版本号与描述，单击“发布”。
2. 单击右上角“”，可查看当前工作流发布历史记录。

18.7.8 通过 API 调用工作流

工作流试运行成功后，可以使用API调用该工作流。

获取调用路径

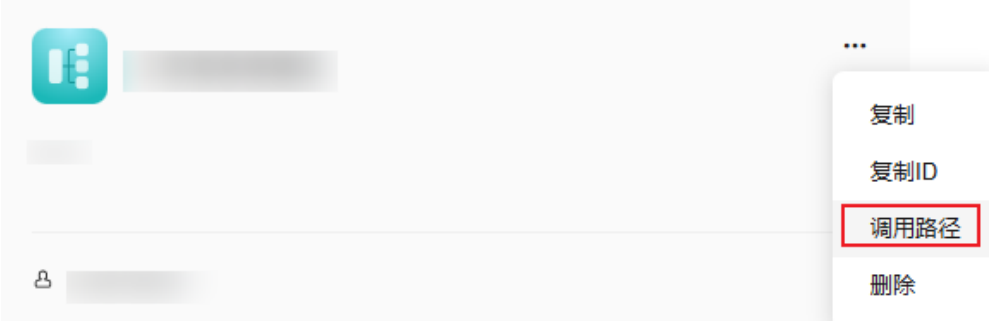
工作流的调用路径获取步骤如下：

1. 登录Agent开发平台。

版权所有 © 华为云计算技术有限公司

2. 单击左侧导航栏“应用管理”，在“工作流”页签，单击所需工作流的“ ··· > 调用路径”。

图 18-88 获取工作流调用路径



3. 在“调用路径”页面，单击“复制路径”即可获取调用路径。

   其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

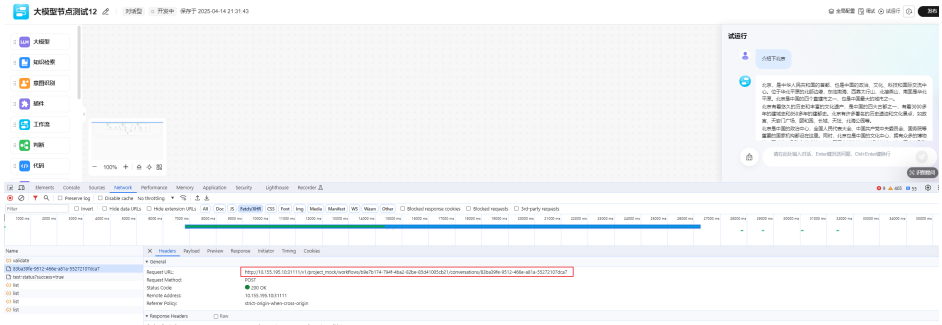
图 18-89 工作流调用路径详情



使用 Postman 调用 API

1. 获取调用接口。
   a. 在工作流页签，单击需要调用的工作流。
   b. 按键盘F12，进入开发者工具。在页面右侧“试运行”中输入并发送问题。在开发者工具中查找本次问答所触发的接口响应，参考下图示例，在接口的Header请求中获取API接口。

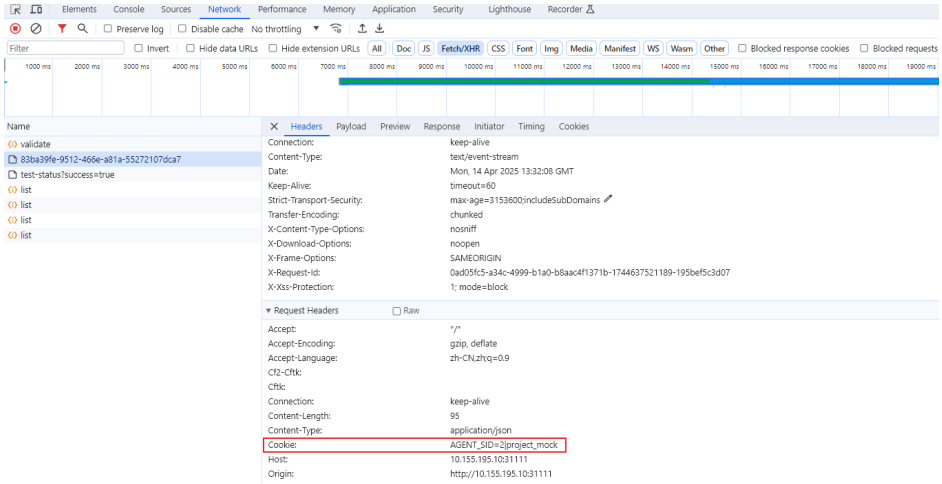
图 18-90 获取调用接口



版权所有 © 华为云计算技术有限公司

a. 获取接口Cookie。
按照步骤1方法找到API接口后，在Header请求中找到Cookie。

图 18-91 获取 Cookie



2. 在Postman中新建POST请求，并填入工作流的调用路径，详见步骤1。
3. 填写请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为Cookie，参数值获取方法参考**步骤2**。
4. 在Postman中选择“Body > raw”选项，请求Body填写示例如下。
   其中，inputs参数为用户提出的问题，作为工作流的输入。工作流有插件时，添加plugin_configs参数，plugin_id参数为插件ID，获取方式详见**管理插件**。

```
带插件请求
{
  "inputs": {
    "query": "你好"
  },
  "plugin_configs": [
    {
      "plugin_id": "xxxxxxxxx",
      "config": {
        "key": "value"
      }
    }
  ]
}
不带插件请求
{
  "inputs": {
    "query": "介绍下北京"
  }
}
```

5. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示应用API调用成功，响应示例如下：

```
data:{"event":"workflow_finished","data":{"status":{"code":1,"desc":"succeeded"},"outputs":{"responseContent":"北京，是中华人民共和国的首都，也是中国的政治、文化、科技和国际交流中心。位于华北平原的北部边缘，东临渤海，西靠太行山，北接燕山，南面是华北平原。北京是中国的四个直辖市之一，也是中国最大的城市之一。\n\n北京有着悠久的历史和丰富的文化遗产，是中国的四大古都之一，有着3000多年的建城史和850多年的建都史。北京的许多历史遗迹和文化景点，如故宫、天安门、颐和园、长城、天坛、北海公园等，都是世界文化遗产，吸引了大量的国内外游客。\n\n北京也是中国的经济中心之一，拥有众多的科研机构、高等院校和大型企业。北京的科技、教育、文化、金融、信息等产业发达，是中国最重要的经济中心之一。\n\n此外，北京还是中国的交通枢纽，拥有发达的公路、铁路、航空网络，是中国北方最大的交通枢纽。北京首都国际机场是中国最繁忙的机场之一，也是亚洲最繁忙的机场之一。\n\n北京是中国的文化中心，拥有众多的博物馆、图书馆、剧院、电影院和艺术馆。北京的京剧、相声、评书等传统艺术形式，以及各种现代艺术形式，都在这里得到了广泛的发展。\n\n北京还是中国的教育中心，拥有众多的高等院校和科研机构，如北京大学、清华大学、北京师范大学等，是中国最重要的教育中心之一。\n\n总的来说，北京是一个历史悠久、文化丰富、经济发达、科技先进、教育水平高的现代化大都市。"},"start_time":1744638874396,"end_time":1744638885881}}
```


版权所有 © 华为云计算技术有限公司

工作流带有提问器节点时返回示例如下所示，具体参数释义详见**API参考**。

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "messages": [
        {
            "role": "assistant",
            "content": "你好！"
        }
    ],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734336269313,
    "end_time": 1734336270908
}
```

工作流不带提问器节点时的返回示例如下所示，具体参数释义详见**API参考**。

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "outputs": {
        "responseContent": "您好，请问有什么可以帮到您！"
    },
    "messages": [],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734337068533,
    "end_time": 1734337082545
}
```

18.7.9 工作流节点配置参考

18.7.9.1 开始和结束节点

开始节点用于开启触发一个工作流，结束节点用于返回一个工作流的最终结果。

开始节点

开始节点是一个工作流的起始节点，用于设定启动工作流所需的输入参数。开始节点只有输入参数，没有输出等其他参数。开始节点中默认有一个输入参数query，表示用户在本轮对话中输入的原始内容。你也可以按需添加其他参数，用于下游节点的输入。

 版权所有 © 华为云计算技术有限公司



开始节点参数配置说明如下：

- **数据类型：**开始节点支持配置String、Number、Integer、Boolean、Object、File、Array多种类型的输入参数，其中Object类型参数最多支持3层嵌套。
- **参数描述：**参数的描述信息，帮助模型理解传入参数的含义。将工作流绑定到智能体中使用时，模型会自动分析用户的Query，将Query中表达的信息填入对应的参数中。
- **是否必选：** 参数是否必选。若未指定必选参数，工作流无法执行。将工作流绑定到智能体中使用时，如果用户Query中缺少必选参数，则不会触发工作流调用。
- **参数默认值：**您可以设置输入参数的默认值，默认值将会回显到试运行界面的输入框中，其中Object类型参数的默认值需输入Json数据，后台会校验Json数据和参数定义的一致性后赋值。示例如下图所示。

图 18-92 参数配置



 版权所有 © 华为云计算技术有限公司

图 18-93 参数类型为 object 类型示例



结束节点

结束节点是工作流的最终节点，用于返回工作流运行后的结果。结束节点支持配置两种参数：输入参数和输出参数，分别对应两种返回方式，文本返回和变量返回。



结束节点的参数配置说明如下：

版权所有 © 华为云计算技术有限公司

输入参数：输入参数支持引用和输入两种类型，输入参数需要在指定回复的文本框中以{{variable_name}}的形式进行插入才能返回。

指定回复：您可以在文本框中编辑指定的回复内容，支持在文本中以{{variable_name}}的形式插入输入参数返回或直接返回输入参数。工作流的最终运行结果将按照指定回复中的内容返回。指定回复中不能插入输出参数。

输出参数：输出参数将以变量形式返回，支持引用和输入两种类型。工作流运行结束后会以JSON格式返回所有输出参数，适用于子工作流的场景。如果工作流直接绑定了智能体，对话中触发了工作流时，大模型会自动总结JSON格式的内容，并以自然语言回复用户。输出参数不能在指定回复中引用。

18.7.9.2 大模型节点

大模型节点提供了使用大模型的能力，可在节点中配置已部署的模型，用户可以通过编写Prompt、设置参数让模型处理相应任务。

大模型节点为**可选**节点，若无需配置，可跳过该步骤。

大模型节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“大模型”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-13**，完成大模型节点的配置。

表 18-13 大模型节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	配置大模型处理需要的输入参数值， 这些值会动态的添加到提示词中， 默认设置的输入参数名为query。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出参数。 – 输入：将用户自定义的内容传递给大模型，设置为输入模式后， 无论前置节点产生什么输出内容，都不会传递给大模型。

 版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
	输出参数	该参数用于解析大模型节点的输出，并提供给后续节点的输出参数引用。 ● 参数名称：参数的名称长度必须大于等于1个字符，并且字符只允许为下面三种类型： – 字母（A-Z或a-z） – 数字（0-9） – 特殊字符：_ **说明** 用户自定义输出参数名称不允许与内置输出参数rawOutput同名。大模型节点有一个内置输出参数rawOutput，代表该节点未经解析的原始输出，与大模型节点相连的后序节点可以直接引用该输出。 ● 参数类型：输出参数的类型，可选String、Integer、Number、Boolean、Object、Array。 ● 描述：对于该输出参数的描述。 ● 输出格式：支持输出的格式包括文本、Markdown、JSON。 文本：大模型原始内容输出，仅支持一个参数，默认为raw_output， 支持修改名称； Markdown：期望模型输出markdown格式内容时选择。仅支持一个参数，默认为raw_output， 支持修改名称； JSON：要求模型按Json格式响应；支持添加多个参数;
模型配置	模型选择	选择模型接入模块已配置的大语言模型。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
提示词配置	提示词	配置输入给大模型的提示词，用于指导模型更好地完成任务。配置提示词时，支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。
	记忆	是否打开记忆功能；打开后可记录多轮对话的内容，默认关闭。

版权所有 © 华为云计算技术有限公司

图 18-94 大模型节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接大模型节点和其他节点。

18.7.9.3 知识检索节点

知识检索节点可以基于用户的输入，从指定知识库内召回匹配的信息，并将匹配结果以列表形式返回。该节点支持选择用户创建的知识库，创建步骤请详见**创建知识库**。

知识检索节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“知识检索”节点至画布中，单击该节点以打开节点配置页面。
2. 连接知识检索节点和其他节点。
3. 参照**表18-14**，完成大模型节点的配置。

表 18-14 知识检索节点配置说明

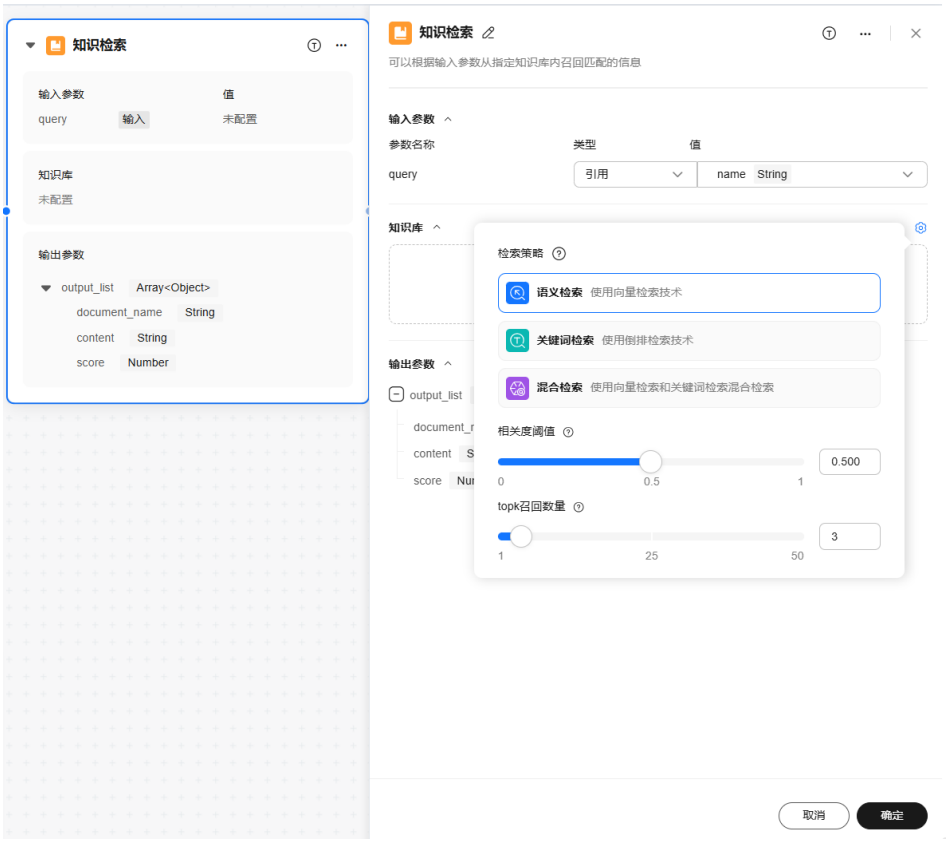
配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：输入参数固定只有1个，参数名称为query且不可修改，类型是字符串，表示待知识检索的问题。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值，限制String类型，适用于需要从前置节点输出中获取知识检索问题的场景。 – 输入：支持用户自定义输入问题，适用于知识检索问题固定的场景。

版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
	输出参数	知识检索节点的输出是一个对象数组，参数名是output_list，表示所有满足检索要求的知识切片。数组中对象有三个属性： ● document_name，知识切片所在的知识文档名称。 ● content，知识切片的内容。 ● score，知识切片的匹配度得分，output_list中的元素按照得分由高到低排序。 后续节点引用该输出参数，可以引用output_list，此时将获取全量的检索结果，包括文档名、切片内容和分数。也可以直接引用切片的属性，比如content，此时将获取output_list中第一条记录的切片内容。
知识库	知识库	支持选择用户所创建的知识库。
	检索策略	召回知识库知识时使用的不同检索策略。当前支持如下三种检索策略： **语义检索：**使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。 **关键词检索：**使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。 **混合检索：**使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。
	相关度阈值	得分低于相关度阈值的搜索结果会被过滤，可以参考知识库命中测试的相关度分值调整该阈值。 取值范围为0~1。
	topk召回数量	从知识库中召回的最大切片数量，如topk召回数量为5，则得分不在前5的切片将被过滤。 取值范围为1~50。

版权所有 © 华为云计算技术有限公司

图 18-95 知识检索节点配置示例



4. 节点配置完成后，单击“确定”。

18.7.9.4 意图识别节点

意图识别节点主要是让应用理解用户自然语言表达的意图或目的，可用于需要对用户问题进行分类，或者提供综合类功能有不同分支处理的场景。

意图识别节点通过对用户输入进行推理分析，匹配预定义的意图关键字类别，并根据匹配结果引导至相应的处理流程，该节点通常位于工作流的前置位置。

意图识别节点支持普通模式或高级模式运行。

- 普通模式：适用于对少量意图进行分类的场景
- 高级模式：适用于对大量可归类意图进行分类的场景

意图识别节点为**可选**节点，若无需配置，可跳过该步骤。

意图识别节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“意图识别”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-15**，完成意图识别节点的配置。

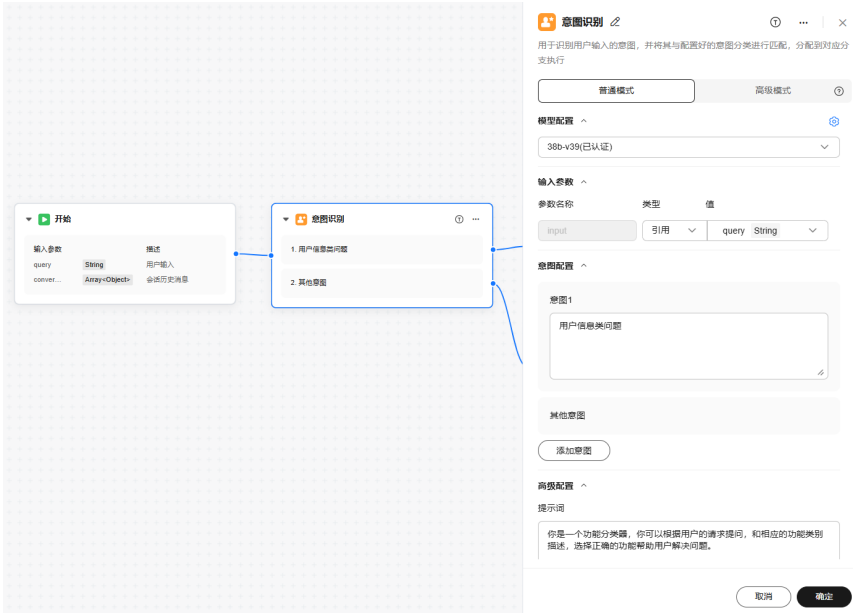
版权所有 © 华为云计算技术有限公司

表 18-15 意图识别节点配置说明

配置类型	参数名称	参数说明
模型配置	模型选择	用于配置进行意图识别的大模型，可选择平台已接入的任意模型。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
参数配置	输入参数	● 参数名称：默认名称input，为固定值，不可编辑。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	输出参数	固定输出classification_id和name参数。 ● classification_id：识别意图的对应序号。 ● name：识别意图的具体名称。
意图配置	意图	用于配置相关意图关键字信息，用户可以添加意图，意图类别默认为意图1、意图2...，意图数量最多为20个。 在意图输入框中输入意图描述信息，描述信息为针对该类别的描述语句或者关键词，也将作为大模型进行推理和分类的依据。
	其他意图	用于控制用户输入意图无法识别的场景，如果推理分析后无法匹配预定义的意图分类，会默认走其他意图对应分支执行后续流程。
高级配置	提示词	高级可选配置项，提供进阶开发者修改提示词，如果不配置将会使用系统默认值。提示词的撰写可能影响到意图识别节点的准确性。
	历史对话轮次	选择是否打开历史对话引用功能，默认为0即不会引用对话历史，配置N轮即可记录N轮对话的内容。
	辅助识别	开启辅助识别后，优先通过知识库分类样例的精确匹配进行意图识别，提升意图识别节点的分类能力。 ● 意图样例知识库：开启辅助识别，用户需要先创建分类样例知识库，并选择配置该知识库。 ● 匹配阈值：当分类样例的匹配度低于设置阈值时，会采用默认的大模型进行意图识别分类。阈值范围为0到1。

版权所有 © 华为云计算技术有限公司

图 18-96 意图识别节点配置示例



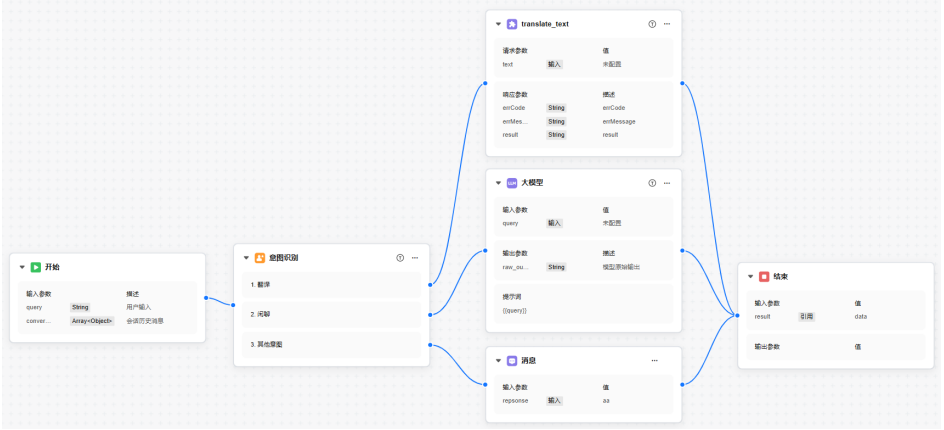
3. 节点配置完成后，单击“确定”。
4. 连接意图识别节点和其他节点。

示例：

以提供综合功能，对用户问题进行不同分支处理的工作流为例，通过意图识别节点对用户输入进行分类，流转至不同的功能模块进行处理。

比如提供翻译功能的工作流，节点配置如下：

- 意图识别节点：工作流将用户问题分为翻译、闲聊两个类别，节点的意图配置添加意图1的类别描述为“文本翻译”，类别后面连接翻译插件节点，实现翻译功能；意图2的类别描述为“用户闲聊”，类别后面连接大模型节点，实现闲聊功能；默认的其他意图类别后面连接消息节点，在消息节点中配置默认回复内容，实现未识别意图场景下的兜底回复。



高级模式

当意图分支比较多（100+）时，建议使用高级模式； 在独立的页面配置意图分支信息，通过选择子工作流的交互方式完成业务配置。

 版权所有 © 华为云计算技术有限公司

配置步骤：

1. 配置意图包。
   a. 单击平台左侧菜单 “意图管理” 新建意图包。
   b. 在意图包中添加意图分类，分类信息包含名称和示例。

   图 18-97 配置意图包



2. 拖动左侧“意图识别”节点至画布中，单击该节点以打开节点配置页面。切换为“高级模式”。
3. 参照意图模式配置说明，完成配置。

表 18-16 意图识别节点配置说明

配置类型	参数名称	参数说明
模型配置	模型选择	用于配置进行意图识别的大模型，可选择平台已部署的任一模型。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
参数配置	输入参数	• 参数名称：默认名称input，为固定值，不可编辑。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	输出参数	固定输出classification_id和name参数。 • classification_id：识别意图的对应序号。 • name：识别意图的具体名称。

版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
意图配置	意图包	选择前面已经配置的意图包。
高级配置	提示词	高级可选配置项，提供进阶开发者修改提示词，如果不配置将会使用系统默认值。提示词的撰写可能影响到意图识别节点的准确性。
	记忆	选择是否打开记忆功能，默认关闭。若开启记忆，可记录多轮对话的内容。

图 18-98 意图识别节点配置示例



4. 节点配置完成后，单击“确定”。
5. 单击意图动作配置节点，配置分支对应的处理逻辑。

图 18-99 配置处理逻辑



6. 配置子工作流的输入参数。

版权所有 © 华为云计算技术有限公司

图 18-100 配置输入参数



7. 单击确定，完成意图动作节点配置。
8. 连接意图动作节点和其他节点。

18.7.9.5 插件节点

插件节点是工作流中实现第三方能力调用的核心组件。作为功能扩展的重要载体，该节点允许通过调用插件来执行特定功能任务。每个插件实质上是经过标准化封装的API工具集合，提供即插即用的模块化服务，拓宽工作流的能力边界，完成更复杂的任务。

插件类型包括预置插件和个人插件。

- 预置插件：平台预置了代码解释器插件，能够执行输入的代码，得到运行结果。支持开发者直接将插件添加到工作流或应用中，丰富其能力。
- 个人插件：平台允许开发者创建自定义插件，支持将API通过配置方式快速创建为插件，提供给工作流或应用调用。

插件节点为**可选**节点，若无需配置，可跳过该步骤。

插件节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“插件”节点至画布中，在“个人插件”或“预置插件”页签单击“+”，将插件添加至画布中。
   - 预置插件为平台内置的插件。
   - 个人插件为用户自定义的插件，创建插件步骤详见**创建插件**。
2. 连接插件节点和其他节点。
3. 单击画布中已添加的“插件”节点，参照**表18-17**，完成插件节点的配置。

版权所有 © 华为云计算技术有限公司

表 18-17 插件节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：从插件元信息中导入，用户无需手动添加。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值，适用于需要从前置节点输出中获取插件入参的场景。 – 输入：支持用户自定义取值，适用于插件入参取值固定的场景。
	输出参数	输出参数所有信息从插件元信息中自动导入，用户无需手动修改。
异常配置	-	“异常配置”开关，作为插件节点的兜底配置，当插件执行异常时，支持配置兜底返回，避免整个工作流运行失败： • 关闭时，该功能不起作用。 • 开启时，填写“默认输出”，默认输出的参数需要与插件“输出参数”一致。当工作流插件节点运行正常时，该配置对后续节点没有影响。当工作流运行到插件节点出现异常时，工作流不会中止，继续运行后续节点。如果后续节点引用了插件节点的输出内容，则使用“默认输出”的内容。

图 18-101 插件节点配置示例



4. 节点配置完成后，单击“确定”。

示例：使用已部署的大模型创建插件

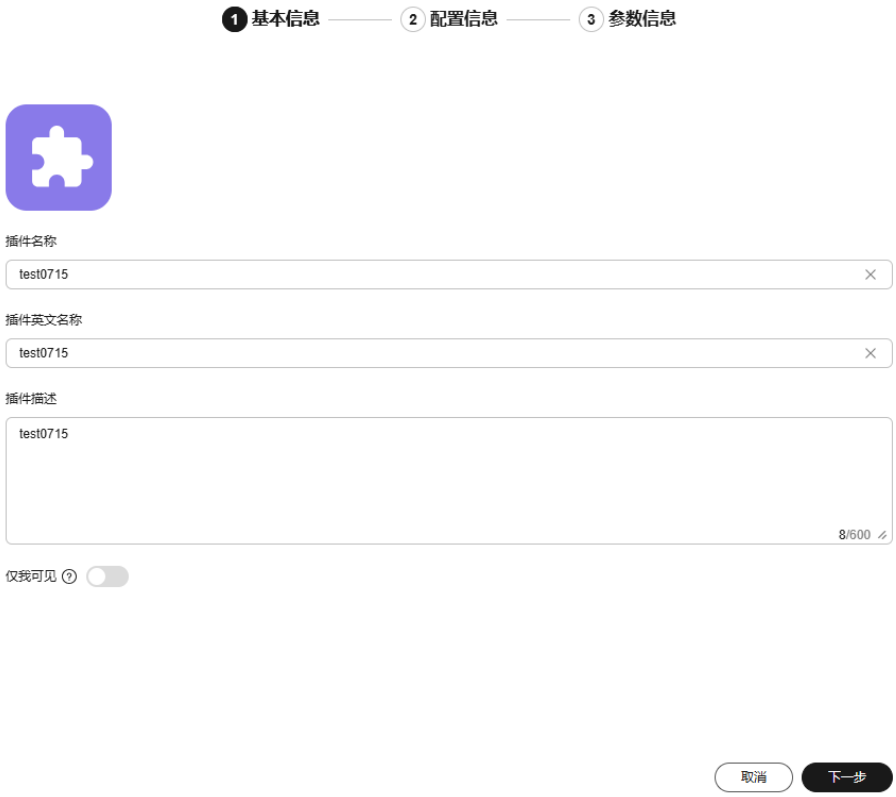
步骤1 部署一个服务，如部署**NLP大模型**：Pangu-NLP-N1-128K。

 版权所有 © 华为云计算技术有限公司

并**获取URL**。

步骤2 在平台**创建插件**，填写基本信息。

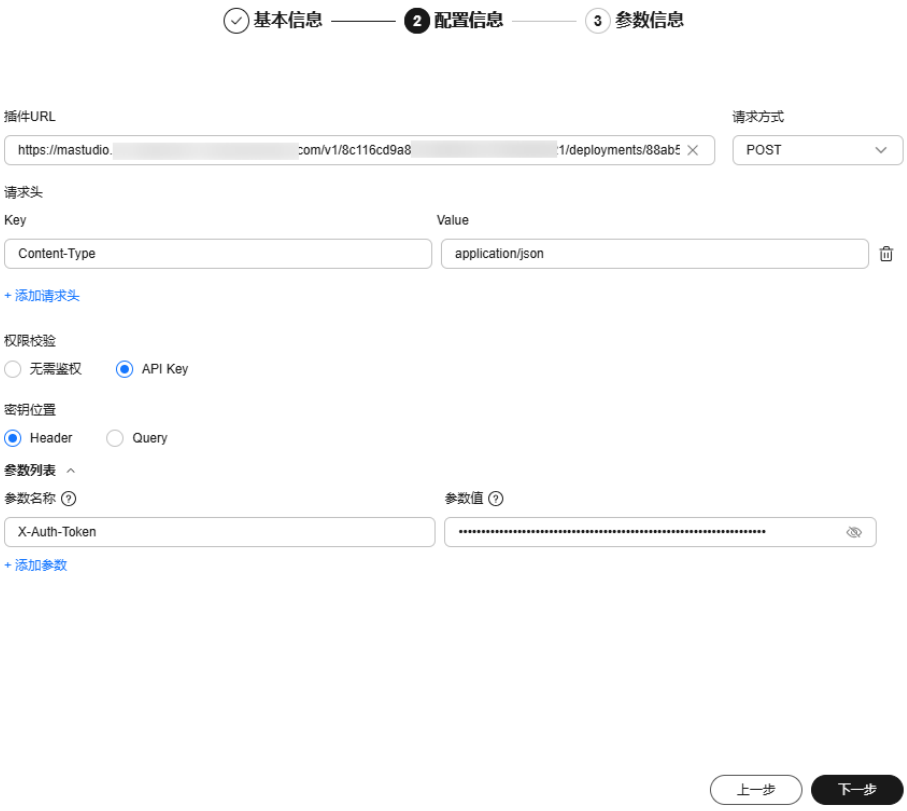
图 18-102 基本信息



步骤3 配置**插件URL**等请求信息。

版权所有 © 华为云计算技术有限公司

图 18-103 配置信息

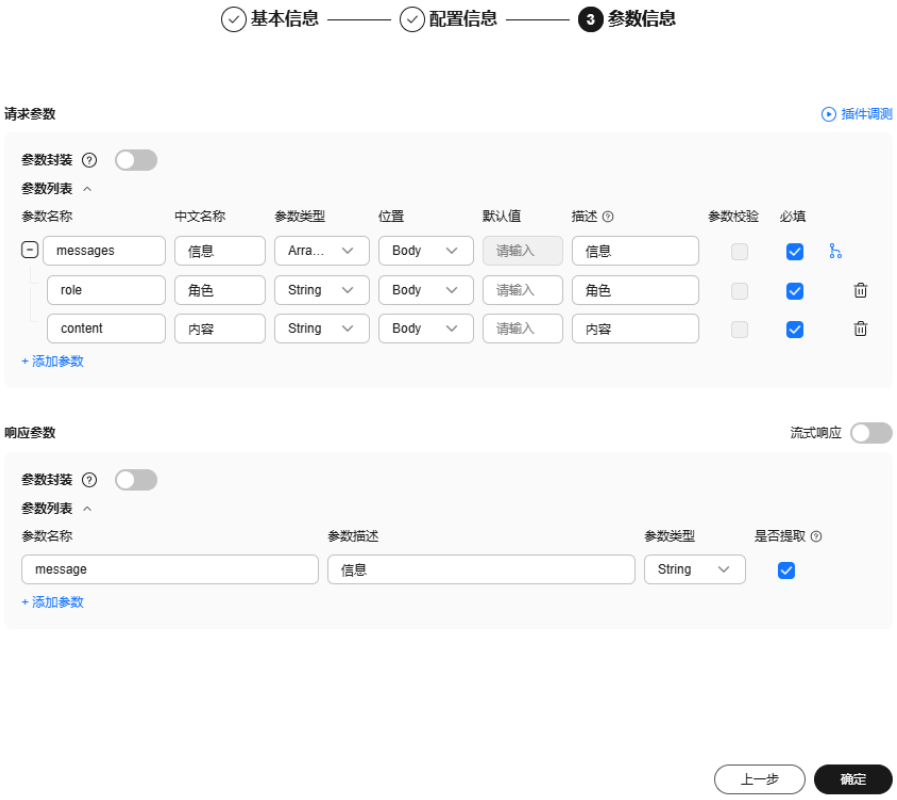


步骤4 根据API接口信息配置参数信息并单击“插件调测”。

确认无误后，单击“确定”即可创建插件。

 版权所有 © 华为云计算技术有限公司

图 18-104 参数信息



步骤5 在工作流中可以使用该插件。

拖动左侧“插件”节点至画布中，在“个人插件”页签单击“+”，将插件添加至画布中。

 版权所有 © 华为云计算技术有限公司

图 18-105 使用插件



----结束

18.7.9.6 判断节点

判断节点是一个IF-ELSE节点，提供了多分支条件判断的能力，用于设计分支流程，实现逻辑判断功能。

判断节点中每个条件分支支持添加多个判断条件（且、或），同时支持添加多个条件分支。

当向该节点输入参数时，节点会逐个条件分支判断输入是否符合分支中预设的条件，符合则执行对应分支后的工作流流程，如果没有符合条件的分支，则执行“ELSE”对应的工作流分支。

判断节点为**可选**节点，若无需配置，可跳过该步骤。

判断节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“判断”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-18**，完成判断节点的配置。

版权所有 © 华为云计算技术有限公司

表 18-18 判断节点配置说明

配置类型	参数名称	参数说明
参数配置	IF	IF分支由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 ● 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 ● 比较条件：条件表达式中间部分，当前支持的比较条件有：长度大于、长度大于等于、长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。针对不同的判断参数类型，前端将展示不同的比较条件。 ● 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值。 – 输入：支持用户自定义取值。 ● 添加条件：单击“+”，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 – 单击“且”或“或”，可以切换该分支表达式的运算逻辑。
	ELSE	用于控制预设条件分支都不满足的场景，如果逐个分支判断都不符合条件，则默认走该分支执行后续工作流流程。
	添加分支	可以添加新的条件分支ELSE IF，新分支的配置方式与IF分支相同。

图 18-106 判断节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接判断节点和其他节点。

示例：

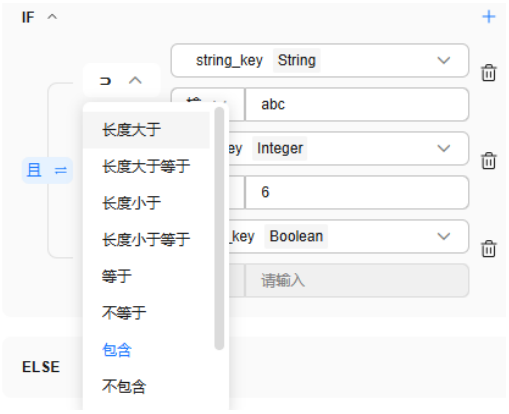
以常见的String、Integer和Boolean类型条件判断为例，在开始节点中定义三种类型的参数，模拟判断节点的输入参数，实现对于不同类型参数在不同条件下的逻辑判断。

版权所有 © 华为云计算技术有限公司

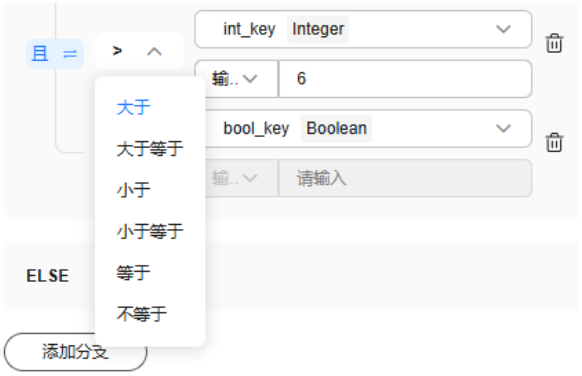
节点配置如下：

- 开始节点：定义三种类型参数，分别为String类型的string_key、Integer类型的int_key、Boolean类型的bool_key
- 判断节点：在IF条件分支中增加三个判断条件，条件表达式的判断参数分别引用开始节点上述的三种类型参数。对于不同类型的参数，前端展示的比较条件有所区别。

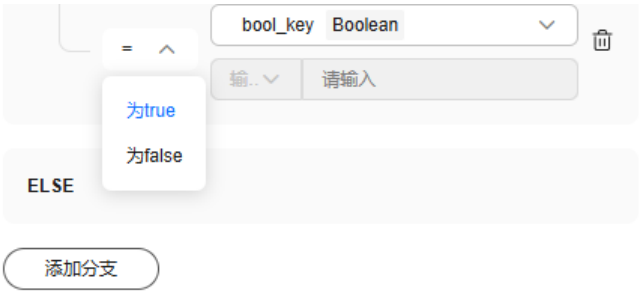
  比如String类型为字符串相关的长度、包含和为空条件判断，示例中配置为判断string_key是否包含“abc”。



Integer类型为数值相关的大小等于条件判断，示例中配置为判断int_key是否大于6。



Boolean类型为true false条件判断，示例中配置为判断bool_key是否为true。



单击试运行，输入string_key：abcd、int_key：7、bool_key：true查看效果

版权所有 © 华为云计算技术有限公司



18.7.9.7 代码节点

代码节点支持编写Python代码来处理文本等复杂逻辑，生成业务期望的返回值。

代码节点为**可选**节点，若无需配置，可跳过该步骤。

代码节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“代码”节点至画布中，单击该节点以打开节点配置页面。
2. 参照表18-19，完成代码节点的配置。

表 18-19 代码节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	配置代码运行需要的输入参数； ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出参数。 – 输入：支持用户自定义取值。
	输出参数	配置代码运行后需要输出的参数，需要与return返回的对象保持一致； ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 参数类型：输出参数的类型，可选String、Integer、Number、Boolean等平台支持的类型。 ● 描述：对于该输出参数的描述。 ● 必填：选择当前输出参数是否必填。

版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明

代码	-	编写Python代码片段，仅支持python系统库，不支持引入依赖包；代码配置示例如下：
		• 文本拼接示例代码。
		def main(args: dict) -> dict:
		# 注意在输入参数中定义名为input1的变量
		input1 = args.get('input1')
		# 注意在输入参数中定义名为input2的变量
		input2 = args.get('input2')
		res = {
		# 注意在输出参数中定义名为res的变量
		"res": input1 + input2,
		}
		return res
		• 数学计算示例代码。
		def main(args: dict) -> dict:
		# 注意在输入参数中定义名为input1的变量
		input1 = args.get('input1')
		try:
		input1 = int(input1)
		return {
		# 注意输出参数中定义res变量
		'res': input1 * input1
		}
		except Exception as e:
		return {
		# 注意输出参数中定义res变量
		'res': "输入类型错误或者数字大小超出限制"
		}
		说明
		编写代码时不要更改第一行函数定义。

图 18-107 代码节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接代码节点和其他节点。

 版权所有 © 华为云计算技术有限公司

18.7.9.8 消息节点

消息节点可提供中间过程的消息输出能力，通过定义一段文本内容，在工作流执行过程中向用户发送该消息。

通常情况下，工作流会在执行完毕后通过结束节点输出最终的执行结果。当开发者想要在工作流执行过程中输出中间节点的结果，可以使用该节点。

消息节点为**可选**节点，若无需配置，可跳过该步骤。

消息节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“消息”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-20**，完成消息节点的配置。

表 18-20 消息节点配置说明

配置类型	参数名称	参数说明	配置示例（按场景写）
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。	在工作流运行到该节点时，如果想要向用户发送包含前置节点输入或输出参数的指定消息，可使用引用方式配置参数；如果想要向用户发送固定内容的指定消息，可使用输入方式配置参数。
指定回复	-	可撰写指定的回复信息，并支持以{{参数名称}}的形式插入变量。回复信息将在工作流执行到该节点时发送给用户。	例如想要在工作流执行过程中向用户发送前置知识检索节点的输出参数"content"，就可以添加输入参数"rep_out"，引用该"content"，并在指定回复中撰写：{{rep_out}}即可。

版权所有 © 华为云计算技术有限公司

图 18-108 消息节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接消息节点和其他节点。

18.7.9.9 提问器节点

提问器节点为开发者提供了收集用户问题所需信息的功能。该节点会循环执行，直到收集到所有必需的信息为止。

提问器节点为**可选**节点，若无需配置，可跳过该步骤。

提问器节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“提问器”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-21**，完成提问器节点的配置。

表 18-21 提问器节点配置说明

配置类型	参数名称	参数说明
模式偏好	-	效果优先：效果优先模式下，会开启时间增强和反思功能，提参成功率更高，时延会增加。 • 时间增强：需要提取时间时，可以将自然语言的时间日期提取为YYYY-MM-DD HH:MM:SS标准格式的时间日期，比如：明天 12:30，提取为 2024-04-13 13:30:00 • 反思功能：数据提取之后，会让模型再判断一下是否提取正确，格式是否满足要求，不满足会尽量做一些修正。比如：期望提取电话号码，用户输入：我不记得电话号码，提取出：189******，反思后会认为提取不正确，会继续追问 速度优先：速度优先模式下时延最低，提参成功率可能无法保障，速度优先模式下不开启时间增强和反思功能。

版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
模型配置	模型选择	选择执行此节点的模型，支持设置模型在此节点中的生成多样性等参数配置，使模型效果更符合你的预期。 提问器模型用于接收用户自然语言，提取用户配置的输出参数，效果优先时还用于提取结果反思和纠正。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
参数配置	输入参数	设置需要添加到问题中的参数，参数值可以引用前置节点的输出参数，或设置为固定文本内容，可引用多个参数。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 示例：输入参数为“pre_assigned_meeting_rooms”，希望用户在指定的多个选项中选出一个，后续问题配置为“有以下几个会议室供您选择：{{pre_assigned_meeting_rooms}}，请选择您想预订的会议室”。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：用户可以选择工作流中该节点的前置节点的输出变量作为取值。 – 输入：用户直接输入变量取值文本。

 版权所有 © 华为云计算技术有限公司

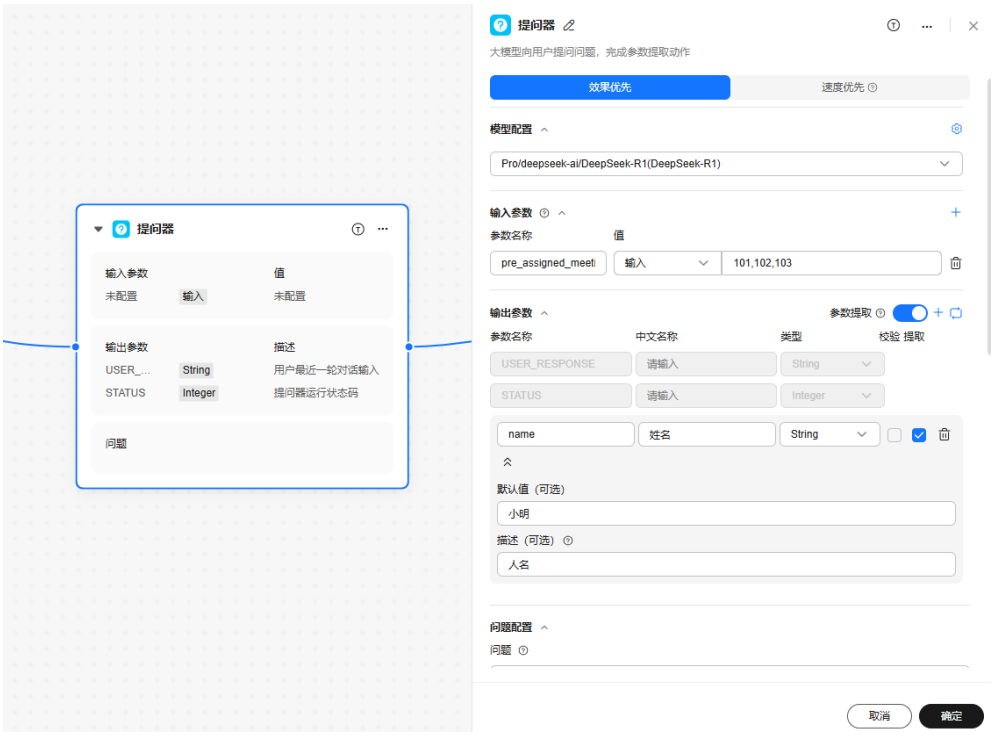
配置类型	参数名称	参数说明
	输出参数	该参数用于解析大模型节点的输出，并提供给后续节点的输入参数引用，支持多参数提取。 ● 默认输出参数。 – USER_RESPONSE： 用户原始输出。 – STATUS：提取状态。 0 正常成功提取，用户无确认。 10 正常成功提取，用户已确认。 100 取到部分参数，用户主动中断，已提参数报错，未提参数按格式置空。 101 取到部分参数，循环超轮次，已提参数报错，未提参数按格式置空。 201 大模型调用异常。 202 反思模块有错误。 ● 参数提取：开启后，可增加需要提取的参数，参数可配置属性如下： – 参数名称：只允许输入字母、数字、下划线、短横线。 – 中文名称：不允许为空。 – 类型：输出参数的类型，可选String、Integer、Number、Boolean。 – 默认值：输出参数的默认值，大模型提取不到参数，并达到最大回复轮数时使用默认值。 – 描述：对于该输出参数的描述。 – 校验：开启后可自定义参数校验规则对输出参数规范性进行校验。规则包括参数名称、校验类型及校验规则。 – 提取：开启后该参数必须提取到或配置了默认值则使用默认值，关闭则该参数允许为空。 ● 引用插件：参数提取可能是给插件使用，通过引用插件，可导入插件的参数信息及校验信息，提升配置效率。
问题配置	问题	该参数将在对话框中原样呈现给用户。如未配置此处，将由大模型根据输出参数描述，自动生成包含所有问题关键词的一个问题。 如：请问你的名字是什么。 可通过jinjia语法在问题中使用输入参数 如：请问你是哪个班级的，可选班级有{{classes}}（ classes先在input参数配置好 ）
	最大回复轮数	该参数用于设置与模型的最大交互次数，超过最大回复轮数还没有提取到参数则跳出提问器。
高级配置	允许用户退出交互	开启后，若用户在与提问器的对话交互中，表达“中止对话”类的意图，系统会自动结束当前提问，并跳转至结束节点。

 版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
	提取参数完毕需要用户确认	开启后，若用户希望提问器参数提取完毕后进行用户确认，则开启此功能。
	提取约束	提供大模型额外的约束信息，用于更准确的提取参数，例如指定被提取参数的格式要求。 举例：用户希望提取电话号码tel_number，约束里面可以写tel_number必须是11位数字
	追问模式	追问模式用来配置，在多次交互过程中，系统返回的参数追问语句生成模式。 ● 默认：使用默认内置追问模板生成追问语句，每次追问内容相同。 ● 智能追问：使用大模型生成语义良好，表达丰富的追问语句，每次追问内容丰富多变。 ● 自定义追问：按照自定义模板配置生成追问语句。‘{unextracted_cn_field_names}’不可修改或删除。每次追问内容相同。 例如要提取名字和年龄参数 ● 默认：请您提供名字，年龄相关的信息。 ● 智能追问：您好，我们需要获取您的名字和年龄（ 模型生成，内容不固定 ） ● 自定义追问：（ 自定义追问模板配置为：请问你的如下信息：{unextracted_cn_field_names} ） 请问你的如下信息：名字，年龄
	示例配置	给大模型一段预期的参数提取示例，增强大模型对参数提取场景的理解。 模板： 输入query：我要坐飞机去呼和浩特学习培训。 提取参数：{"location":"呼和浩特", "traveltool":"飞机"}。
	追问显示枚举值	开启后，对话时追问将显示输出参数配置的枚举值，默认为关闭状态

版权所有 © 华为云计算技术有限公司

图 18-109 提问器节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接提问器节点和其他节点。

18.7.9.10 循环节点

循环节点提供了循环执行节点的能力，可在循环体内配置需要循环的节点，用户可以通过在循环体内编排节点多次执行处理任务。

循环是一种常见的控制机制，用于重复执行一系列任务，直到满足某个条件为止。工作流提供循环节点，当需要重复执行一些操作，或循环处理一组数据时，可以使用循环节点实现。

循环节点为**可选**节点，若无需配置，可跳过该步骤。

循环节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“循环”节点至画布中，拖动其他需要循环执行的节点到循环体画布内部并编排（循环内执行需从循环输入节点开始，输出连接到循环输出节点，暂不支持输入节点、规划节点和循环节点）。
2. 参照**表18-22**，完成循环节点的配置。

版权所有 © 华为云计算技术有限公司

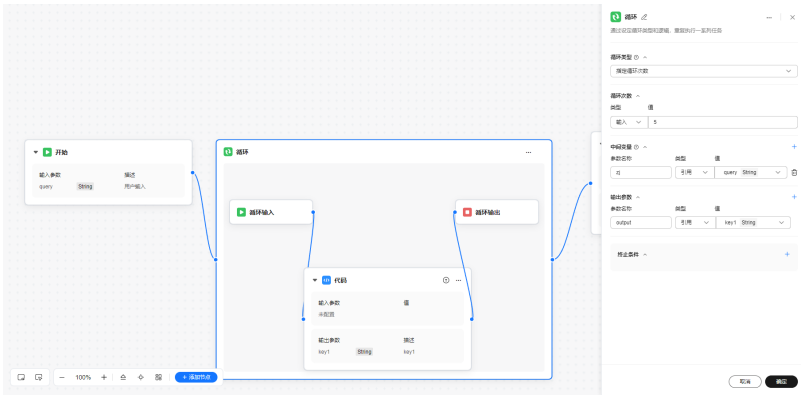
表 18-22 循环节点配置说明

配置类型	参数名称	参数说明
循环类型配置	使用数组循环	使用数组循环类似编程语言中的for语法结构。遍历循环用于遍历一个已知的序列，对序列中的每个元素执行一系列相同的步骤。 使用数组循环时，需要指定arr_loop_var的值，此参数仅支持引用上游节点的输出，且必须为数组格式。使用数组循环模式下执行循环节点时，循环的次数取决于循环数组引用的数组长度。 使用数组循环时，循环节点会遍历数组中的每个元素，每次循环都会将当前循环到的元素赋值给内置变量。内置变量仅限循环节点内部使用。目前支持的内置变量如下： ● item：数组元素，即当前循环到的数组元素。 ● index：数组索引，index+1为当前循环的轮次。
	指定循环次数	指定循环次数模式通常用于批量、顺序处理数据的场景，需要同时设置循环次数。 循环次数默认为5次，支持设置为 1~1000次。 使用参考： 回合制游戏，3局2胜可将循环次数设置为3。 网络爬虫爬取前1000个商品信息，循环次数设置为1000。
变量配置	循环数组	此参数只有在使用数组循环时支持配置，名称固定为arr_loop_var，仅支持引用上游节点输出。
	中间变量	循环节点支持设置中间变量，此变量可作用于每一次循环。中间变量通常和循环体中的设置变量节点搭配使用，在每次循环结束后为中间变量设置一个新的值，并在下次循环中使用新值。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	输出参数	循环节点的输出参数可设置为循环体的执行结果集合，表示当数组中所有元素运行完毕之后，将所有循环的运行结果打包输出给下游。也支持设置为中间变量的取值。

 版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
终止条件	表达式	分支由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 • 判断参数：条件表达式上半部分，需要选择来自前序节点的输出参数。 • 比较条件：条件表达式左侧，当前支持的比较条件有：长度大于、长度大于等于、长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。针对不同的判断参数类型，前端将展示不同的比较条件。 • 比较参数：条件表达式下半部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 • 添加条件：单击“+”，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 – 单击“且”或“或”，可以切换该分支表达式的运算逻辑。

图 18-110 循环节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接循环节点和其他节点。

18.7.9.11 变量赋值节点

变量赋值节点配置说明

变量赋值节点，将特定的值赋给变量，可以实现数据的动态更新和传递，使工作流能够根据实时数据做出相应的处理和决策；变量设置节点也支持在循环节点内部使用，通过变量设置节点，将特定的值赋给中间变量，可以实现循环过程中数据的动态更新和传递。

变量赋值节点配置步骤如下：

 版权所有 © 华为云计算技术有限公司

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“变量赋值”节点至画布中，单击该节点以打开节点配置页面。
2. 连接变量赋值节点和其他节点。
3. 参照表18-23，完成变量赋值节点的配置。

表 18-23 变量赋值节点配置说明

配置类型	参数名称	参数说明
循环节点外变量赋值节点配置	变量赋值	变量赋值节点变量名称仅支持全局配置中记忆变量引用，值可支持引用或者输入两种。 类型、值：支持“引用”和“输入”两种类型。 ● 引用：支持用户选择工作流中已包含的前置节点的输出变量值以及全局配置的中的记忆变量。 ● 输入：支持用户自定义取值。
循环节点中变量赋值配置	变量赋值	变量赋值节点支持在循环体内部引用，只支持更改循环体中间变量的值，被赋值变量仅支持选择中间变量，值可支持引用或输入两种。适用于循环过程中动态更新中间变量，自定义循环逻辑中进行参数传递的场景。 类型、值：支持“引用”和“输入”两种类型。 ● 引用：中间变量的值需要引用上游节点输出时勾选此项，支持用户选择工作流中已包含的前置节点的输出变量值以及循环体内置变量，包括index、item（数组循环）以及中间变量，适用于循环过程中修改中间变量的值为变量的场景。 ● 输入：支持用户自定义取值，适用于循环过程中修改中间变量的值为固定值场景。

图 18-111 变量赋值节点配置示例



 版权所有 © 华为云计算技术有限公司

图 18-112 变量赋值节点在循环节点中配置示例



4. 节点配置完成后，单击“确定”。
5. 连接变量赋值节点和其他节点。

18.7.9.12 变量聚合节点

变量聚合节点能够对多个分支的输出进行聚合处理，方便后置节点统一配置。

如果工作流中设计了多个分支，往往需要一个节点来汇总所有分支的输出结果。在这种场景下，你可以使用变量聚合节点聚合多路分支的输出变量，变量聚合节点会读取多路分支中第一个不为空的值，供流程下游的节点使用和操作，不用额外处理未运行分支的输出结果，简化了数据流的管理。

变量聚合节点为**可选**节点，若无需配置，可跳过该步骤。

变量聚合节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“变量聚合”节点至画布中，单击该节点以打开节点配置页面。
2. 参照表18-24，完成变量聚合节点的配置。

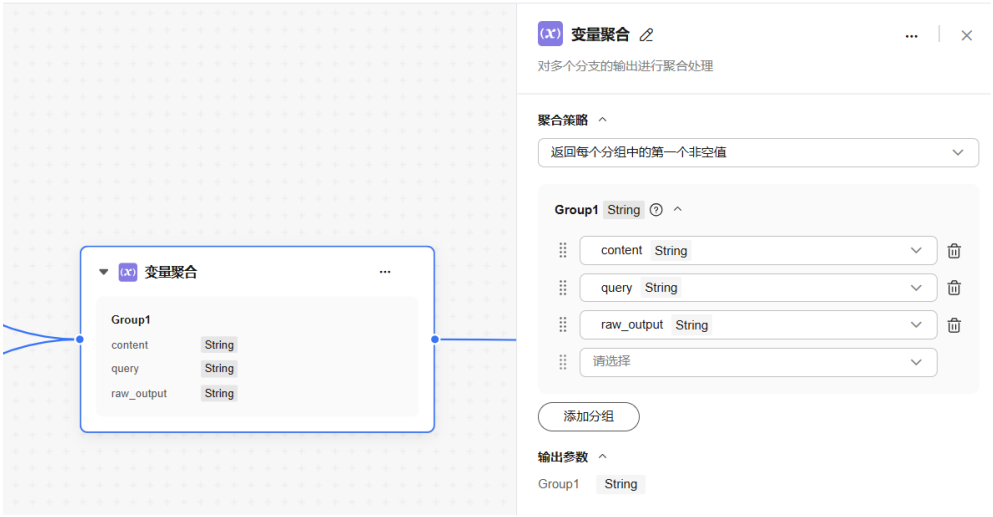
表 18-24 变量聚合节点配置说明

配置类型	参数名称	参数说明	配置示例（按场景写）
参数配置	输出参数	• 参数名称：固定为Group1，如果有多个分组则根据分组数量递增为Group2、Group3 等。 • 参数类型：取决于对应聚合分组的变量数据类型。	-
聚合策略	-	通过指定策略对每个分组中的所有变量进行聚合处理，同一组内的变量实施相对应的聚合策略。 目前聚合策略仅支持设置为“返回每个分组中第一个非空值”，支持拖动变量、调整变量位置。例如组内按顺序设置三个变量output1、output2和output3，将其聚合为一个变量Group1，如果ouput1不为空，则用output1的值为Group1赋值；如果ouput1为空，则取ouput2的值，依次类推。	目前聚合策略仅支持设置为“返回每个分组中第一个非空值”。

 版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明	配置示例（按场景写）
聚合分组	-	默认只有一个分组Group1，对应一个输出变量Group1。分组中所有变量类型相同。如果需要输出多个变量，可以添加多个分组，依次递增为Group2、Group3等。	-
聚合变量	-	在聚合分组中选择需要聚合的变量，每个分组只能聚合一种数据类型的变量。例如将多个String类型的变量聚合为一个String变量、将多个Integer类型的变量聚合为一个Integer变量。	例如判断节点或意图识别节点会将工作流拆分为多路分支，每次执行时，工作流会根据判断条件或用户意图决定运行其中某一个分支，此时未执行的其他分支输出变量为空。你可以将多路分支都连接到变量聚合节点，并设置一个输出变量。无论哪条分支被执行，其结果都能通过这个变量被引用与访问，避免了下游节点对相同语义输出变量的重复定义。

图 18-113 变量聚合节点配置示例



 版权所有 © 华为云计算技术有限公司

3. 节点配置完成后，单击“确定”。
4. 连接变量聚合节点和其他节点。

18.7.9.13 输入节点

输入节点提供工作流运行过程中的信息输入。

在比较复杂的工作流场景中，某些节点的执行往往需要额外的用户输入。如果前置节点中没有获取到这些信息，你可以添加一个输入节点来主动收集信息。工作流执行到输入节点时会暂时中断，直到此节点收集到必要的用户输入。

输入节点为**可选**节点，若无需配置，可跳过该步骤。

输入节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“输入”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-25**，完成变量输入节点的配置。

表 18-25 输入节点配置说明

配置类型	参数名称	参数说明	配置示例（按场景写）
参数配置	输入参数	支持配置一个或多个输入参数，且输入参数可被后置节点引用。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 参数类型：可选String、Integer、Number、Boolean、Object、Array、File类型。 • 描述：对于该输入参数的描述。	当配置一个或多个输入参数后，工作流运行到该节点时会暂时中断，直到用户填写所有输入参数，且每个输入参数均可被后置节点引用。 例如插件节点需要一个输入参数："city"，可通过在该插件节点的前置节点中配置一个输入节点用于填写"city"的具体值，再在插件节点的"city"参数处引用输入节点相应参数即可。

版权所有 © 华为云计算技术有限公司

图 18-114 输入节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接输入节点和其他节点。

18.7.9.14 工作流节点

工作流节点，实现工作流嵌套工作流的效果。

在一个工作流中，你可以将另一个工作流作为其中的一个步骤或节点，实现复杂任务的封装。例如将常用的、标准化的任务处理流程封装为不同的子工作流，并在主工作流的不同分支内调用这些子工作流执行对应的操作。工作流嵌套可实现复杂任务的模块化拆分和处理，使工作流编排逻辑更加灵活、清晰、更易于管理。

工作流节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“工作流”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表1 工作流节点配置说明**，完成工作流节点的配置。

表 18-26 工作流节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 工作流节点的输入结构取决于子工作流定义的输入结构，不支持自定义设置 • 在工作流节点中你需要为输入参数指定数据来源，支持设置为固定值或引用上游节点的输出参数。
	输出参数	• 工作流节点的输出结构取决于子工作流定义的输出结构，不支持自定义设置。 • response_content为工作流固定输出参数

版权所有 © 华为云计算技术有限公司

图 18-115 工作流节点配置示例-子工作流



图 18-116 工作流节点配置示例-工作流节点（输入和输出对应子工作流的输入和输出）



3. 节点配置完成后，单击“确定”。
4. 连接工作流节点和其他节点。

18.7.9.15 MCP 服务节点

MCP服务节点是工作流中实现第三方能力调用的核心组件之一。作为功能扩展的重要载体，该节点允许通过调用MCP服务来执行特定功能任务。每个MCP服务实质上是一个工具集合，可以提供模块化服务来拓宽工作流的能力边界，完成更复杂的任务。

MCP服务类型包括预置服务和个人服务。

- 预置服务：平台预置了"高德地图"、"车票查询工具"、"必应搜索"等多种实用MCP服务，开通后可以一键集成调用。支持开发者在工作流或应用中添加预置MCP服务，丰富其能力。
- 个人服务：平台允许开发者创建自定义MCP服务，支持将MCP服务地址通过配置方式快速创建为自定义MCP服务，提供给工作流或应用调用。

MCP服务节点为**可选**节点，若无需配置，可跳过该步骤。

MCP服务节点配置步骤如下：

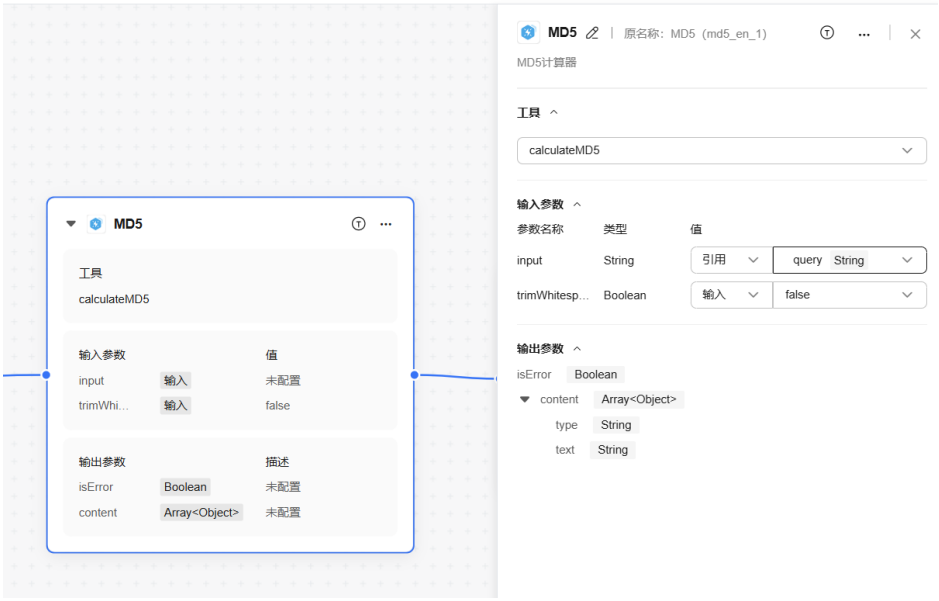
 版权所有 © 华为云计算技术有限公司

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“MCP服务”节点至画布中，在“个人服务”或“预置服务”页签单击“+”，将MCP服务添加至画布中，其中有些“预置服务”不能直接添加，需要单击“立即开通”，开通服务后即可添加至画布中。
   - 预置服务为平台内置的MCP服务。
   - 个人服务为用户创建的自定义MCP服务，创建MCP服务步骤详见**创建MCP服务**。
2. 连接MCP服务节点和其他节点。
3. 单击画布中已添加的“MCP服务”节点，参照**表1 MCP服务节点配置说明**，完成MCP服务节点的配置。

表 18-27 MCP 服务节点配置说明

配置类型	参数名称	参数说明
参数配置	工具	支持从当前MCP服务所包含的工具列表中选择一个作为工作流运行到该节点时会执行的工具。
	输入参数	• 参数名称、类型：从插件元信息中导入，用户无需手动添加。 • 值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量，适用于需要从前置节点输出中获取插件入参的场景。 – 输入：支持用户自定义取值，适用于MCP服务入参取值固定的场景。
	输出参数	输出参数所有信息从MCP服务元信息中自动导入，用户无需手动配置。

图 18-117 MCP 服务节点配置示例



版权所有 © 华为云计算技术有限公司

4. 节点配置完成后，单击“确定”。

18.7.9.16 Agent 节点配置说明

Agent节点提供了使用大模型的能力以及大模型工具调用的能力，可在节点中配置已部署的模型，用户可以通过编写Prompt、绑定插件让模型处理相应任务。

Agent节点为**可选**节点，若无需配置，可跳过该步骤。

Agent节点配置步骤如下：

1. 单击画布下方“添加节点”按钮，在节点抽屉中拖动“Agent”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表18-28**，完成变量输入节点的配置。

表 18-28 Agent 节点配置说明

配置类型	参数名称	参数说明
模型配置	模型选择	选择执行此节点的模型，支持设置模型在此节点中的生成多样性等参数配置，使模型效果更符合你的预期。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。
	插件	可绑定手动创建的插件或预置插件，当模型识别到需要调用工具来完成任务时，会根据用户的输入提取参数完成插件调用，并总结插件执行结果。
	系统提示词	配置输入给大模型的系统提示词，用于指导模型更好地完成任务。配置提示词时，支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。
	输出参数	输出参数为Agent节点最后一轮的输出。
终止条件	最大迭代轮次	该参数用于设置与模型的最大交互次数，超过最大回复轮数还没有提取到参数则跳出Agent节点。

版权所有 © 华为云计算技术有限公司

配置类型	参数名称	参数说明
	插件执行成功	该参数开启后可绑定插件，当执行该插件成功后跳出Agent节点。
	识别到用户有退出意图	该参数开启后识别到用户输入有退出意向时，跳出Agent节点。

图 18-118 Agent 节点配置示例



版权所有 © 华为云计算技术有限公司

3. 节点配置完成后，单击“确定”。
4. 连接Agent节点和其他节点。

18.7.10 管理工作流

Agent开发平台支持对工作流执行复制、获取工作流ID、删除、导入、导出操作。

获取工作流 ID、删除工作流

1. 登录Agent开发平台。
2. 单击左侧导航栏“应用管理”，在“工作流”页签选择要操作的工作流应用。
3. 单击“ ··· > 复制”，可复制当前工作流。
4. 单击“ ··· > 复制ID”，可获取当前工作流ID。
5. 单击“ ··· > 删除”，可删除当前工作流。

说明

删除应用属于高危操作，删除前，请确保该工作流不再使用。

导出、导入工作流

平台支持导出和导入工作流。导出工作流时，将同步导出工作流关联的插件等配置。

1. 登录Agent开发平台。
2. 单击左侧导航栏“应用管理”，在“工作流”页签选择要操作的工作流应用。
3. 导出工作流。
   a. 单击页面右上角“导出”。
   b. 在“导出工作流应用”页面选择工作流应用，单击“导出”。工作流将以一个jsonl格式的文件下载至本地。
4. 导入工作流。
   a. 单击页面右上角“导入”。
   b. 在“导入”页面选择工作流应用，单击“选择文件”选择需要导入的jsonl文件。
   c. 选择导入文件后，选择解析内容。

      平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。
   d. 单击“导入”，导入成功的工作流将在“我的工作台 > 应用管理 > 工作流应用”页面中展示。

18.8 管理 Agent 平台插件

18.8.1 插件管理

 版权所有 © 华为云计算技术有限公司

18.8.1.1 插件介绍

在Agent开发平台中，插件是大模型能力的重要扩展。通过模块化方式，插件能够为大模型提供更多专业技能和复杂任务处理能力，使其在多样化的实际场景中更加高效地满足用户需求。

通过插件接入，用户可以为应用赋予大模型本身不具备的能力。插件提供丰富的外部服务接口，当任务执行时，模型会根据提示词感知适用的插件，并自动调用它们，从外部服务中获取结果并返回。这样的设计使得Agent能够智能处理复杂任务，甚至跨领域解决问题，实现对复杂问题的自动化处理。

Agent开发平台支持两种类型的插件：

- 预置插件：平台为开发者和用户提供了预置插件，直接可用，无需额外开发。例如，平台提供的“Python解释器插件”能够根据用户输入的问题自动生成Python代码，并执行该代码获取结果。此插件为Agent提供了强大的计算、数据处理和分析功能，用户只需将其添加到应用中，即可扩展功能。
- 自定义插件：为了满足更个性化的需求，平台允许开发者创建自定义插件，支持将API通过配置方式快速创建为插件，并供Agent调用。这样，开发者可以根据特定需求为应用增加专属功能。

18.8.1.2 创建插件

18.8.1.2.1 基于 API 创建一个插件

创建插件

创建插件的步骤如下：

1. 登录Agent开发平台。
2. 单击左侧导航栏“我的工作台>组件库”，在“插件”页签，单击右上角“创建组件>创建插件”。
3. 在“创建插件”页面，填写插件名称、插件英文名称与插件描述，设置是否“仅我可见”（设置后仅插件创建者可见）。单击图片可上传插件图标，单击“下一步”。
4. 在“配置信息”页面，参照表18-29完成信息配置。

表 18-29 插件信息配置说明

参数名称	参数说明
插件URL	插件服务的请求URL地址。 ● URL协议只支持HTTP和HTTPS。 ● 系统会校验URL地址是否为标准的URL格式。 ● URL对应的IP默认不应为内网，否则会导致注册失败。
请求方式	插件服务的请求方式，POST或GET。

版权所有 © 华为云计算技术有限公司

参数名称	参数说明
权限校验	选择调用API时是否需要通过鉴权才可以调用。 ● 无需鉴权：API可以公开访问，不需要任何形式的身份验证或授权。 ● API Key：在调用API时提供一个唯一的API Key进行鉴权。需填写密钥位置，即密钥是从Header中读取还是Query中读取。并设置API Key的密钥鉴权参数名和密钥值。
请求头	填写API的请求头信息，例如： ● Key：Content-Type ● Value：application/json

图 18-119 API 请求信息配置示例



说明

自定义插件使用HTTP服务，或不增加鉴权方式可能存在安全风险。

5. 单击“下一步”，在“参数信息”页面，参照**表18-30**完成参数配置。

版权所有 © 华为云计算技术有限公司

表 18-30 插件参数配置说明

参数类型	参数名称	参数说明
请求参数	参数封装	开启后，会将请求参数封装为一个列表（数组）结构，可适配入参为数组格式的插件接口。 示例：原参数列表：{"a":"string", "b":1}，开启封装后的参数列表：[{"a":"string", "b":1}]
	参数名称	参数的名称，参数名称会作为大模型解析参数含义的依据。
	中文名称	该参数的中文名称。
	参数类型	该参数值的数据类型，String、Integer、Number等多种类型支持选择。
	位置	当前参数在请求信息中的位置，可选Body、Headers或Query。
	默认值	参数的默认值。
	描述	参数的描述，尽可能准确的描述参数的含义和要求，可提升Agent提取参数的准确率。
	参数校验	可设置当前参数的校验规则。
	必填	指定该参数是否为必填项。
响应参数	参数封装	开启后，会将请求参数封装为一个列表（数组）结构，可适配入参为数组格式的插件接口。 示例：原参数列表：{"a":"string", "b":1}，开启封装后的参数列表：[{"a":"string", "b":1}]
	参数名称	响应参数的名称，参数名称会作为大模型解析大模型输出结果的依据。
	参数描述	响应参数的描述，参数描述会作为大模型解析大模型输出结果的依据。
	参数类型	该参数值的数据类型，String、Integer、Number等多种类型支持选择。
	是否提取	开启后则该参数必须提取到，关闭则该参数允许为空或者使用默认值。

版权所有 © 华为云计算技术有限公司

图 18-120 填写 API 请求、响应参数



插件创建示例（以 API 调用为例）

准备工作：

- 创建一个服务，此服务需可以访问。

例如：以下服务是为了进行搜索的服务用例，准备相应的接口地址、其请求方法、输入参数、输出参数

接口地址：http://ip:host/bingsearch（此地址需要根据实际情况填写ip、端口和服务路径）

请求方法：POST

输入参数（示例）：

{

"query": "今天天气如何"

}

输出参数（示例）：

```
{
   "webPages": {
      "totalEstimatedMatches": 1200000,
      "value": [
         {
            "dateLastCrawled": "2025-03-11T12:34:56Z",
            "displayUrl": "xxx",
            "id": "1",
            "name": "解密Manus:靠垂直整合,抢在了AI大厂前头",
            "snippet": "Manus通过多智能体协作系统（Multi-Agent Collaboration）实现任务动态拆解，采用蒙特卡洛树搜索（MCTS）算法优化任务优先级。",
            "url": "xxx"
         }
```


版权所有 © 华为云计算技术有限公司

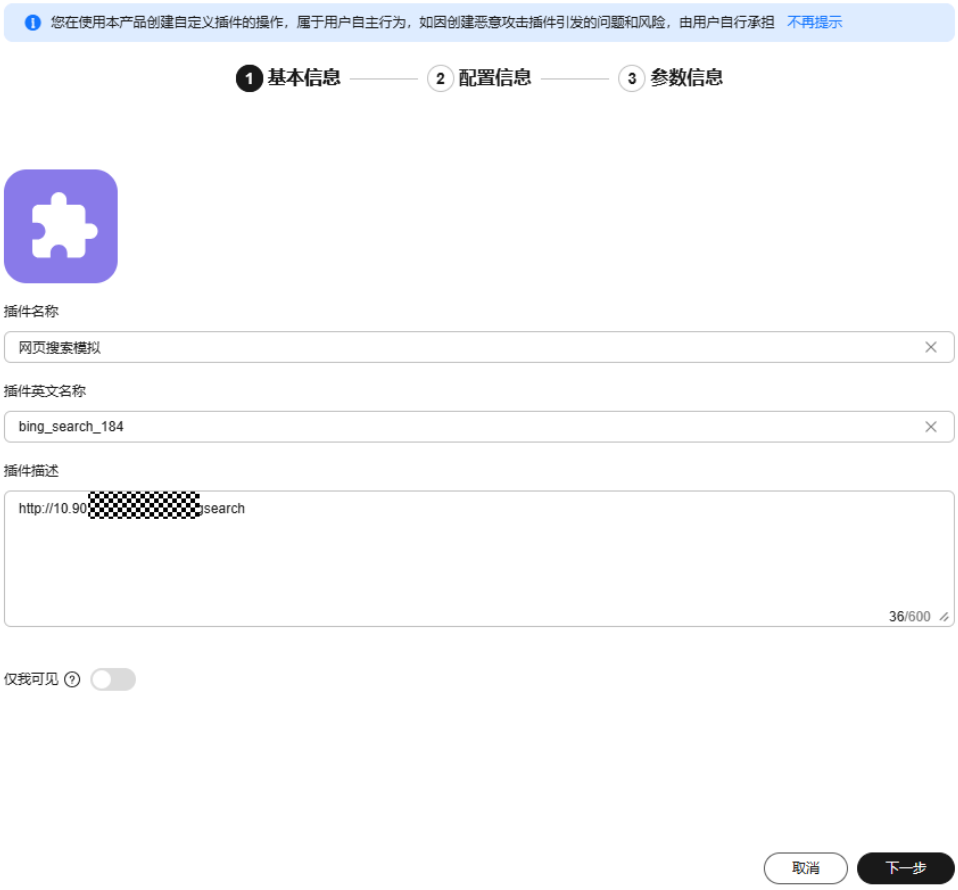
```
    ],
    "webSearchUrl": "xxx"
  }
}
```

第2步：在平台创建插件，填写基本信息

插件描述： 网页搜索模拟

插件英文名称：bing_search_184

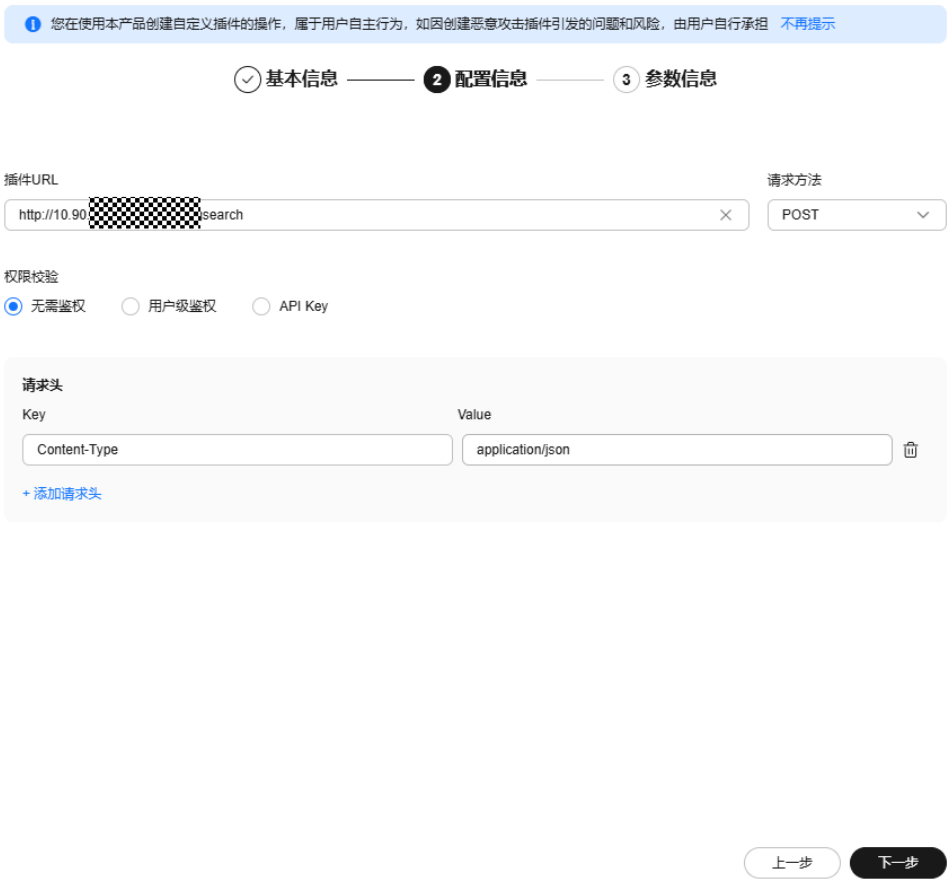
插件描述：http://ip:host/bingsearch



第3步：配置插件URL等请求信息

填入事先准备的服务信息：URL、请求方法、权限校验、请求头

 版权所有 © 华为云计算技术有限公司



第4步：根据API接口信息配置参数信息，参见准备信息。

NOTE：注意响应参数与API接口保持一致。

版权所有 © 华为云计算技术有限公司

图 18-121 API 接口配置



18.8.1.2.2 通过 JSON 文件导入插件

平台支持通过JSON文件形式导入插件。

1. 登录Agent开发平台。
2. 进入“我的工作台 >组件库> 插件”页面。
3. 导入插件。
   a. 单击页面右上角“导入”。
   b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件（jsonl文件生成参考**导出插件**由已知插件导出）。
   c. 选择导入文件后，选择解析内容。
      平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。
   d. 单击“导入”，导入成功的插件将在“我的工作台 >组件库> 插件”页面中展示。

18.8.1.3 在单智能体应用中使用插件

单智能体应用支持添加插件技能，可添加“预置插件”和“个人插件”。添加插件可以为应用配备更多技能，建议插件数量不超过5个。

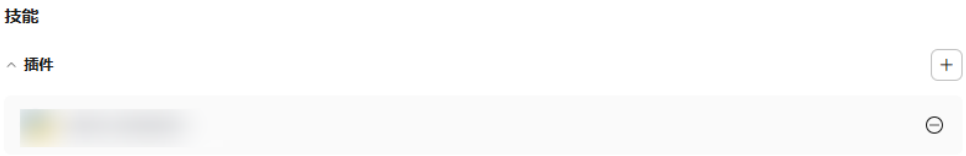
如果需要添加“个人插件”，请确保已完成**创建插件**操作。

版权所有 © 华为云计算技术有限公司

添加插件的步骤如下：

1. 在“技能 > 插件”模块，单击“ ”。
2. 在“添加插件”窗口，选择预置插件或个人插件，单击“ ”进行添加，再单击“确定”。
3. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

图 18-122 配置应用插件



18.8.1.4 在工作流中使用插件

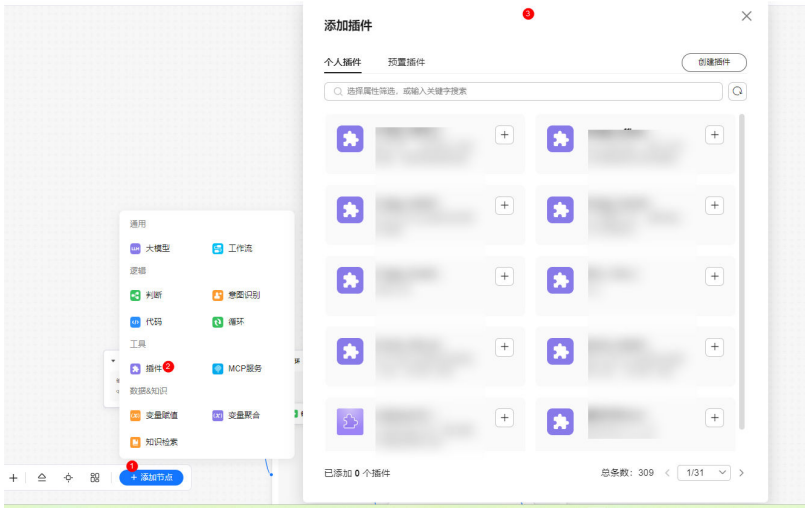
插件让工作流更强大、更智能、更自动化，用户无需开发完整功能，只需选择合适的插件即可快速实现需求。

插件节点配置步骤如下：

单击“+添加节点”，选择“插件”，在“个人插件”或“预置插件”页签单击“ ”，将插件添加至画布中。

1. 添加插件后，将插件节点与相关节点进行连接。
2. 单击画布中已添加的“插件”节点，完成插件节点的配置。

图 18-123 配置插件节点



版权所有 © 华为云计算技术有限公司



18.8.1.5 导出插件

"导出插件"功能允许用户将自定义的插件配置保存为本地文件，便于后续的迁移、备份或共享。

平台支持导出插件，步骤如下：

1. 登录Agent开发平台。
2. 进入“我的工作台 >组件库> 插件”页面。
3. 导出插件。
   a. 单击页面右上角“导出”。
   b. 在“导出”页面选择插件，单击“导出”。插件将以一个jsonl格式的文件下载至本地。

 版权所有 © 华为云计算技术有限公司

18.8.1.6 管理插件

Agent开发平台支持对插件执行获取插件ID、删除、导入、导出操作。

获取插件 ID、删除插件

1. 登录Agent开发平台。
2. 进入“我的工作台 >组件库> 插件”页面。
3. 单击“ ··· > 复制ID”，可获取当前插件ID。
4. 单击“ ··· > 删除”，可删除当前插件。



说明

删除应用属于高危操作，删除前，请确保该插件不再使用。

导出、导入插件

平台支持导出和导入插件。

1. 登录Agent开发平台。
2. 进入“我的工作台 >组件库> 插件”页面。
3. 导出插件。
   a. 单击页面右上角“导出”。
   b. 在“导出”页面选择插件，单击“导出”。插件将以一个jsonl格式的文件下载至本地。
4. 导入插件。
   a. 单击页面右上角“导入”。
   b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。
   c. 选择导入文件后，选择解析内容。
      平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。
   d. 单击“导入”，导入成功的插件将在“我的工作台 >组件库 > 插件”页面中展示。

版权所有 © 华为云计算技术有限公司

18.8.2 知识库管理

18.8.2.1 知识库介绍

平台提供了知识库功能来对文本文档等数据进行向量化存储、知识检索，支持为应用、工作流提供检索增强能力。无论是文本文档、演示文稿，还是电子表格文件，用户都可以轻松地将数据导入知识库，无需额外的转换或格式处理。

知识库支持导入以下格式的本地文档：

表 18-31 支持的文档格式

文档类型	约束说明	大小要求
知识文档	支持上传常见文本格式，包括：.doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, 格式。	• 单个文档不能超过128MB（超过60MB建议通过API上传）

EXCEL 模板格式

表 18-32 EXCEL 模板样例

答案	问题	相似问题1（非必填）	相似问题2（非必填）	相似问题3（非必须）	相似问题4（非必须）	目录（非必须）	标签（非必须，逗号分隔）
云搜索服务（Cloud Search Service，简称CSS）是一个基于Elasticsearch且完全托管的在线分布式搜索服务，为用户提供结构化、非结构化文本、以及基于AI向量的多条件检索、统计、报表。	什么是云搜索服务	什么是华为云搜索服务	云搜索服务是什么	华为云搜索服务是什么	什么是CSS	-	-

版权所有 © 华为云计算技术有限公司

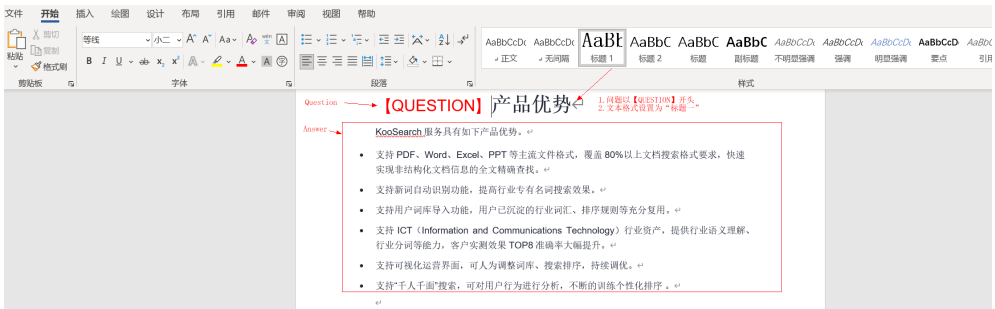
答案	问题	相似问题1（非必填）	相似问题2（非必填）	相似问题3（非必须）	相似问题4（非必须）	目录（非必须）	标签（非必须，逗号分隔）
云搜索服务可以帮助网站和APP搭建搜索框，提升用户的搜索体验；也可以用于搭建日志分析平台，助力企业实现数据驱动运维，数据驱动运营；它的向量检索能力可以帮助客户快速构建基于AI的图搜、推荐、语义搜索等丰富的应用。	云搜索服务有什么应用场景	云搜索服务有哪些应用场景	云搜索服务有什么作用	云搜索服务的作用有哪些	CSS的作用有哪些	-	-
公有云提供了Web化的服务管理平台，即管理控制台和基于HTTPS请求的API（Application programming interface）管理方式。	云搜索服务有哪些访问方式	云搜索服务提供了什么访问方式	如何访问云搜索服务	如何访问CSS	-	-	-

Word 模板格式

Word格式知识库模板将所有知识通过问答对的方式写入word文档，每一个Q&A对应一个标题对。问题以【QUESTION】开头，文本属性设置为“标题一”。答案另起一行，文本属性设置为“正文”。

样例如下：

图 18-124 Word 模板格式说明



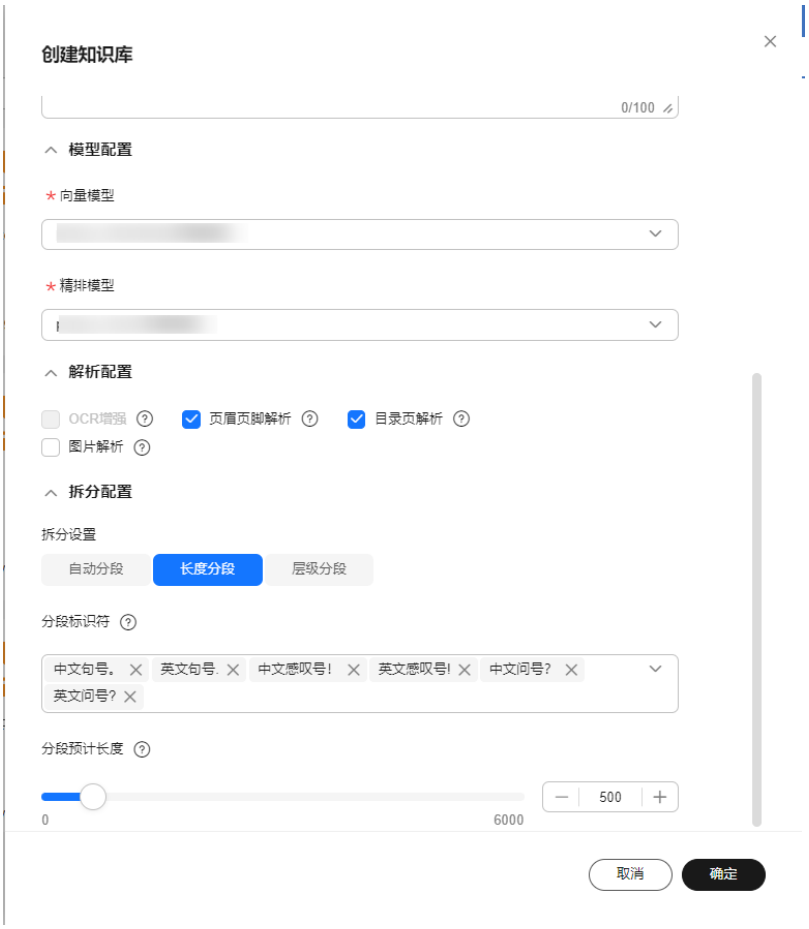
 版权所有 © 华为云计算技术有限公司

18.8.2.2 创建知识库及上传文档

创建知识库

1. 登录Agent开发平台。
2. 单击左侧导航栏“我的工作台”，在“知识库”页签，单击右上角“创建知识库”。
3. 在知识库创建界面填写基础信息、向量精排模型配置、解析切分策略配置：
   - 基本信息：配置知识库图标，填写名称、描述。
   - 模型配置：选择向量及精排模型。
   - 解析配置：配置文档解析。
   - 拆分配置：配置文档拆分。
4. 配置完成后，单击“确定”，完成知识库创建。

图 18-125 创建知识库



上传文档

1. 登录Agent开发平台。
2. 单击左侧导航栏“我的工作台”，在“知识库”页签，单击进入知识库详情。
3. 在知识库详情中选择需要上传的文档类型：

版权所有 © 华为云计算技术有限公司

- 知识文档：

a. 在知识库详情“知识文档”页签，单击“上传”进入文档上传页面；

b. 单击“点此上传”上传本地文件至知识库；

c. 支持上传.doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, 格式的文档；

d. 单个文档不能超过128MB（超过60MB建议通过API上传）；

e. 支持单次上传文件个数不超过300个。

- FAQ文档

i. 在模型详情“FAQ文档”页签，单击“上传”进入FAQ上传页面；

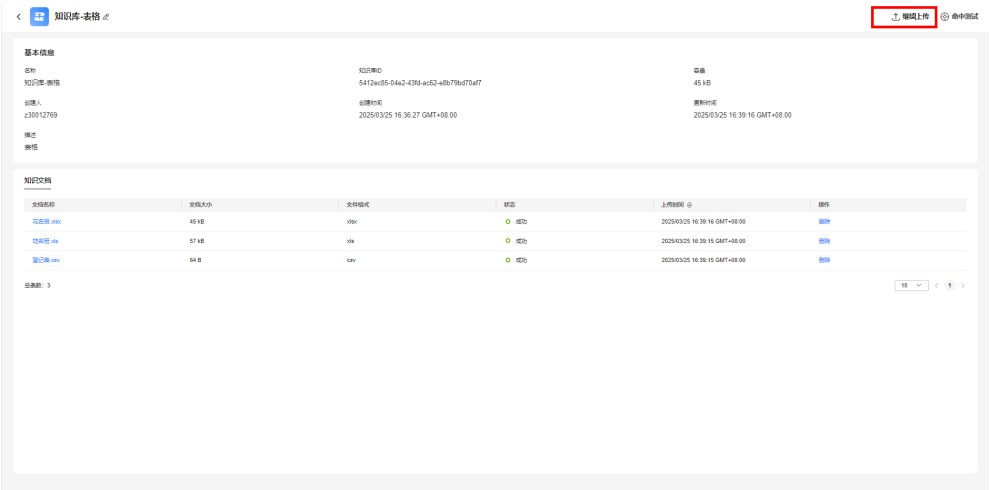
ii. 单击“点此上传”上传本地文件至知识库，要求按照模板上传文本，支持下载FAQ模板；

iii. 要求单个excel文件最大为10M，数据条数10000条，文件中不允许空行，空行后的数据将被忽略；

iv. 单个word文件最大为10M；

v. 支持单次上传文件个数不超过300个。

上传完成后，单击“确定”，完成文档的上传。



说明

知识库创建完成后，如果想在当前知识库中继续上传文件，可单击该知识库进入详情页面，选择对应的页签，继续上传文档。

配置完成的文档解析、分段策略支持修改，可进入知识库详情页，单击“高级设置”进行修改。修改后，可以对文档重新进行解析切分。

18.8.2.3 知识库命中测试

平台支持对创建的知识库进行命中测试，以评估知识库的效果和准确性。

命中测试通过将用户的查询与知识库中的内容进行匹配，最终输出与查询相关的信息，并根据匹配的程度进行排序。

知识库命中测试步骤如下：

1. 登录Agent开发平台。

版权所有 © 华为云计算技术有限公司

2. 进入“我的工作台 > 知识库”页面，单击所需知识库，进入知识库基本信息页面，单击右上角“命中测试”。
3. 在文本框中输入问题，单击“命中测试”，页面下方将根据不同的检索方式，展示多条匹配的内容，并按照匹配分值降序排列。

   用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。
4. 单击“查看历史”，可以查看用户输入的历史问题。

图 18-126 查看历史



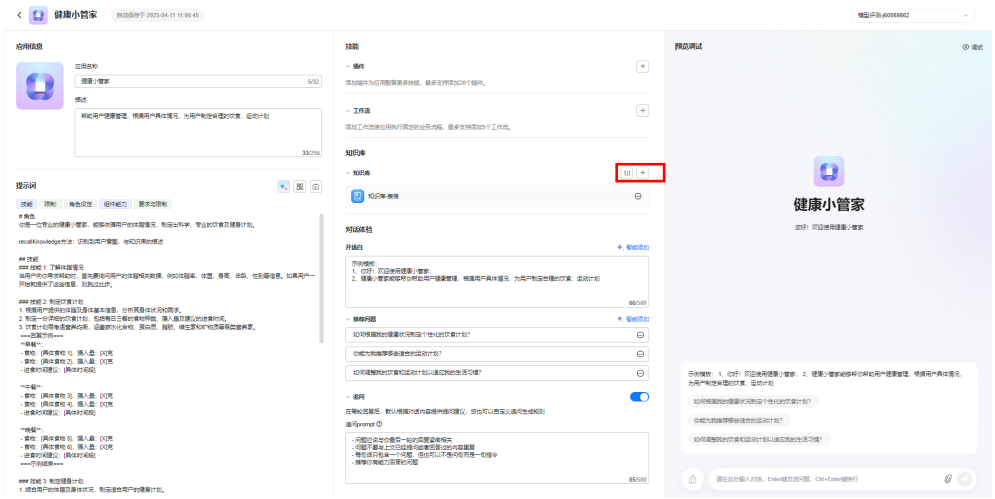
18.8.2.4 在单智能体中使用知识库

平台支持在Agent中添加引用知识库，以根据用户意图来检索召回对应的知识切片。

为平台Agent应用配置知识库的具体步骤如下：

1. 登录Agent开发平台。
2. 进入“我的工作台 > 应用管理>单智能体应用”页面，单击所需的应用，进入应用编辑页面。
3. 单击技能中的知识库模块的“添加”按钮，进入知识库添加页面。
4. 选中所需的知识库，单击“添加”，添加所需要的知识库，单击确定完成知识库添加。支持添加最多5个知识库。
5. 单击技能中的知识库模块的“高级配置”按钮，弹出知识检索配置弹窗。
6. 配置检索策略、相关度阈值、topk召回数量，单击其他位置退出弹窗，完成配置。

版权所有 © 华为云计算技术有限公司



18.8.2.5 在工作流中使用知识库

平台支持在Agent中添加引用知识库，以根据用户意图来检索召回对应的知识切片。

为平台Agent应用配置知识库的具体步骤如下：

1. 登录Agent开发平台。
2. 进入“工作台 > 工作流”页面，单击所需的工作流，进入工作流编辑页面。
3. 在左侧导航栏选中“知识检索”节点，拖拽至画布，单击“知识检索”进入参数配置页面。
4. 配置“知识检索”输入参数。
5. 单击技能中的知识库模块的“添加”按钮，进入知识库添加页面。
6. 选中所需的知识库，单击“添加”，添加所需要的知识库，单击确定完成知识库添加。支持添加最多5个知识库。
7. 在知识库模块配置检索参数，含检索策略、相关度阈值、topk召回数量。
8. 单击参数配置页面“确定”，完成知识检索节点配置。

18.8.2.6 管理知识库

新增、删除知识库中的文档

1. 登录Agent开发平台。
2. 进入“我的工作台 > 知识库”页面。
3. 单击所需知识库，进入详情页面。
   - 新增知识库中知识文档。单击右上角“继续上传”，可上传本地文档至当前知识库。
   - 删除知识库中知识文档。在“知识文档”中单击操作列“删除”可删除当前知识文档。

开启、关闭知识库

1. 登录Agent开发平台。
2. 进入“我的工作台 > 知识库”页面。单击知识库片上更多按钮，查看当前知识库状态，单击开启/停用按钮进行操作。

版权所有 © 华为云计算技术有限公司

3. 单击所需知识库，进入详情页面。单击右上角开启/停用按钮，进行操作。

说明

仅知识库为开启状态，才能在应用、工作流中引用该知识库。
停用已经被应用、工作流引用的知识库，会导致检索结果返回空值，请谨慎操作。

获取知识库 ID、删除知识库

1. 登录Agent开发平台。
2. 进入“我的工作台 > 知识库”页面。

- 单击“ ··· > 复制ID”，可获取当前知识库ID。
- 单击“ ··· > 删除”，可删除当前知识库。

说明

删除属于高危操作，删除前，请确保该知识库不再使用。

18.8.3 MCP 服务管理

18.8.3.1 MCP 服务介绍

MCP（Model Context Protocol，模型上下文协议）是由Anthropic推出的一种开放标准，旨在统一大型语言模型（LLM）与外部数据源和工具之间的通信协议。

MCP的主要目的在于解决当前AI模型因数据孤岛限制而无法充分发挥潜力的难题，MCP使得AI应用能够安全地访问和操作本地及远程数据，为AI应用提供了连接万物的接口。

18.8.3.2 创建 MCP 服务

开发者可以在“控制台”选择 “Agent开发 > MCP服务市场”使用预置的各类MCP服务，也可以选择“Agent开发 > 我的工作台>组件库”的“MCP服务”选项创建MCP服务。华为云ModelArts Studio支持在智能体应用、工作流应用开发过程中，添加MCP服务，增强Agent工具调用能力。

使用官方预置 MCP 服务

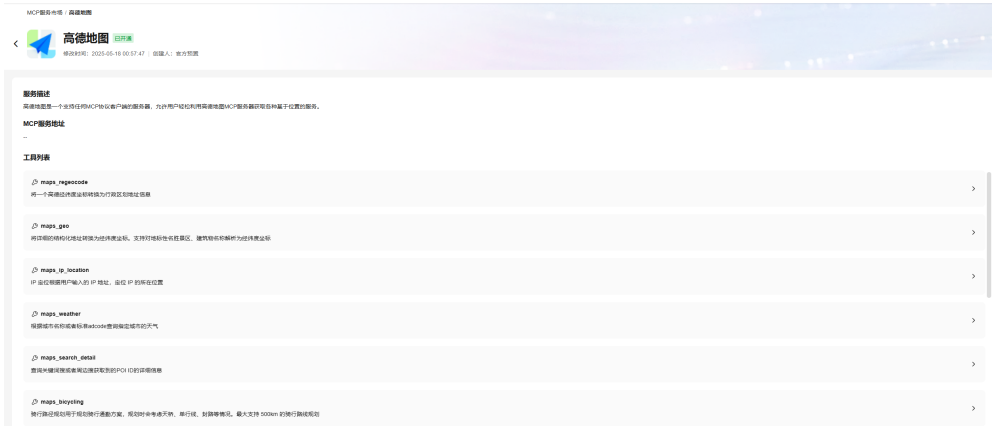
1. 选择“Agent开发”，跳转至Agent开发平台。
2. 选择“Agent开发 > MCP服务市场”。

 版权所有 © 华为云计算技术有限公司

图 18-127 MCP 服务市场



3. 通过属性类型筛选，或通过搜索MCP服务名称找到目标服务。单击目标服务，进入服务详情界面。



4. 根据MCP服务开通提示，进行相应的操作，如该MCP服务需要API Key认证，输入API Key之后，单击确认开通该MCP服务。

创建自定义 MCP 服务

如果您已有可用的工具和API资源，则可以使用MCP服务器来配置您的MCP扩展。完成MCP服务器构建后，只需要在Agent开发平台配置对应的工具即可创建自己的MCP服务。

1. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
2. 在右侧导航栏中选择“我的工作台>组件库”，在“MCP服务”页签，单击“创建MCP服务”，进入创建MCP页面。
3. 在“创建MCP服务”页面，参考**表18-33**配置“MCP服务名称”、“英文名称”、“MCP服务地址”、“描述”以及“权限校验”选项后，单击“测试并下一步”。

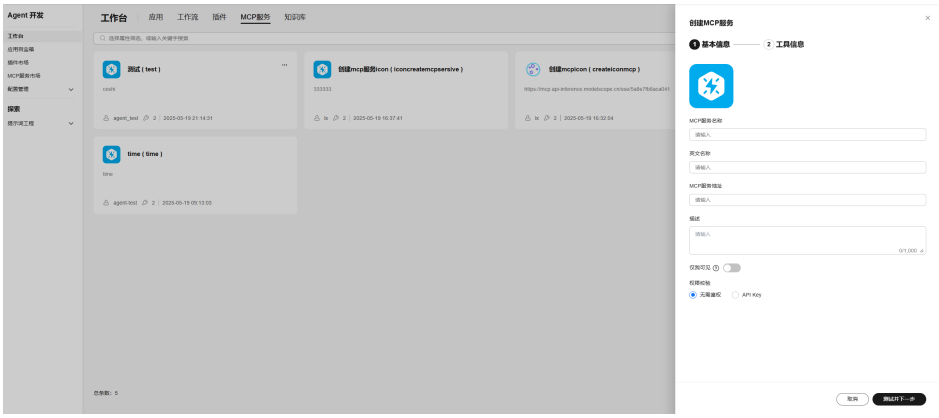
表 18-33 MCP 服务信息配置说明

参数名称	参数说明
MCP服务名称	MCP服务中文名。

版权所有 © 华为云计算技术有限公司

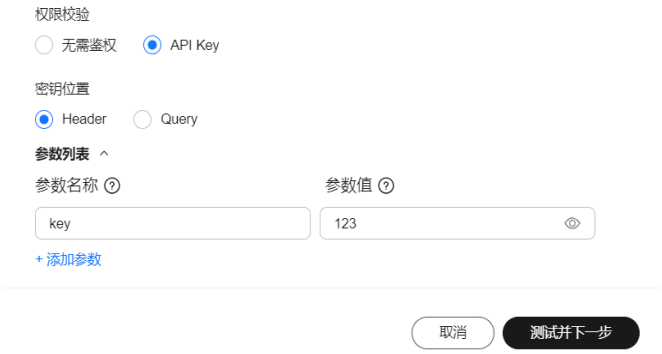
参数名称	参数说明
英文名称	MCP服务英文名，需确保同一个租户下唯一。
MCP服务地址	MCP服务SSE地址。 ● 服务URL地址只支持HTTP和HTTPS。 ● 系统会校验URL地址是否为标准的URL格式。 ● URL对应的IP默认不应为内网，否则会导致注册失败。 **说明** 支持网站公开的MCP服务，也支持**用户自行创建的MCP服务**。
描述	MCP服务描述。
仅我可见	是否仅创建者可见。
权限校验	选择调用MCP服务时是否需要通过认证才可以调用。 ● 无需鉴权：MCP服务可以公开访问，不需要任何形式的身份验证或授权。 ● API Key：在调用MCP服务时提供唯一的API Key进行鉴权。 需填写密钥位置，即密钥信息是设置在Header中还是设置在Query中。并设置API Key的密钥鉴权参数名和密钥值。 **说明** 若服务地址中包含有认证信息，建议使用平台的权限校验（API Key）功能。

图 18-128 创建 MCP 服务



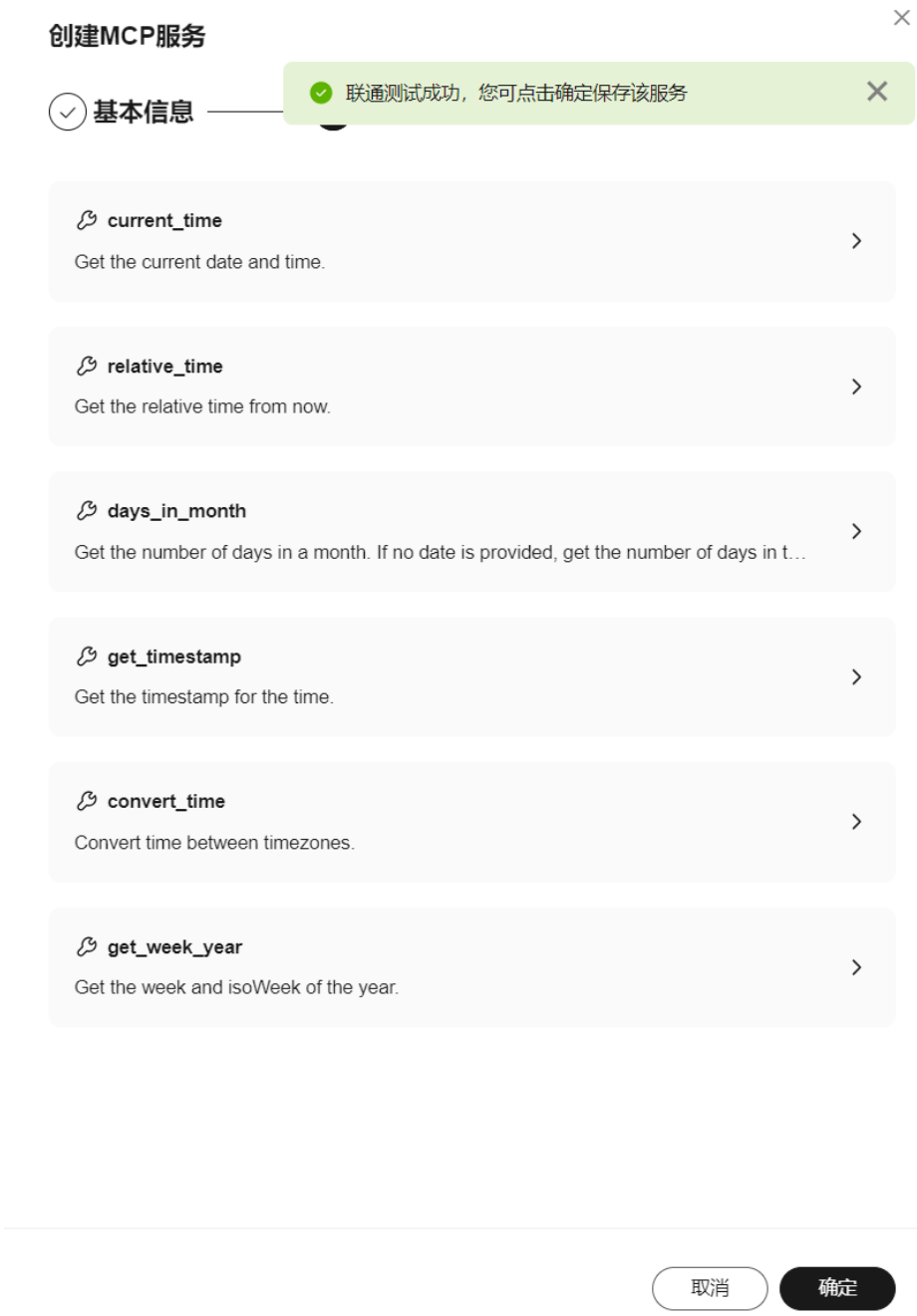
版权所有 © 华为云计算技术有限公司

图 18-129 权限校验示例



版权所有 © 华为云计算技术有限公司

图 18-130 连通性测试成功示例



4. 测试完成，单击确认完成MCP服务创建。

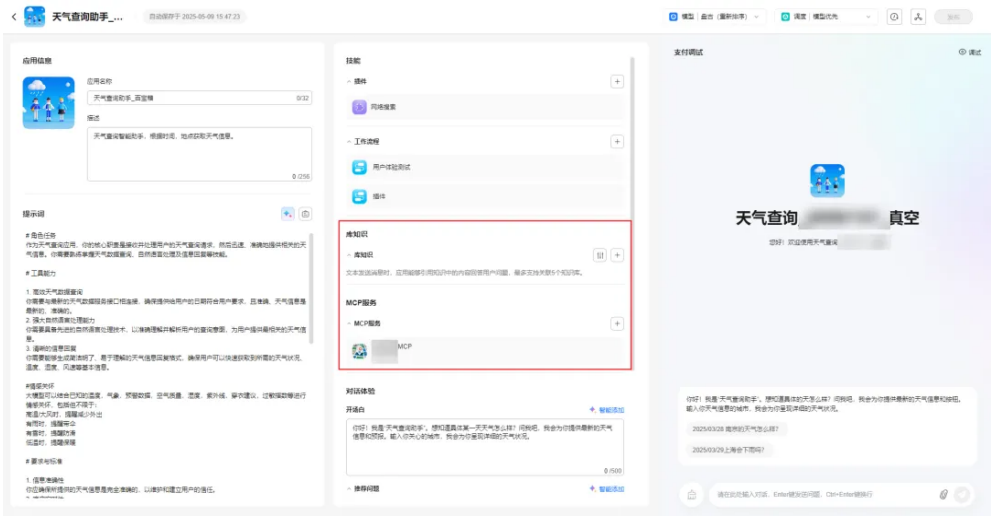
使用 MCP 服务

在Agent开发平台中，开发者可以在开发智能体应用、工作流应用时，添加预置和自定义创建的MCP服务。

- **在智能体应用开发时配置MCP服务：**
  步骤一：进入Agent开发平台后，在导航栏选择“我的工作台”。
  步骤二：在左侧导航栏中选择“应用管理>单智能体应用”，选择一个应用进入。

版权所有 © 华为云计算技术有限公司

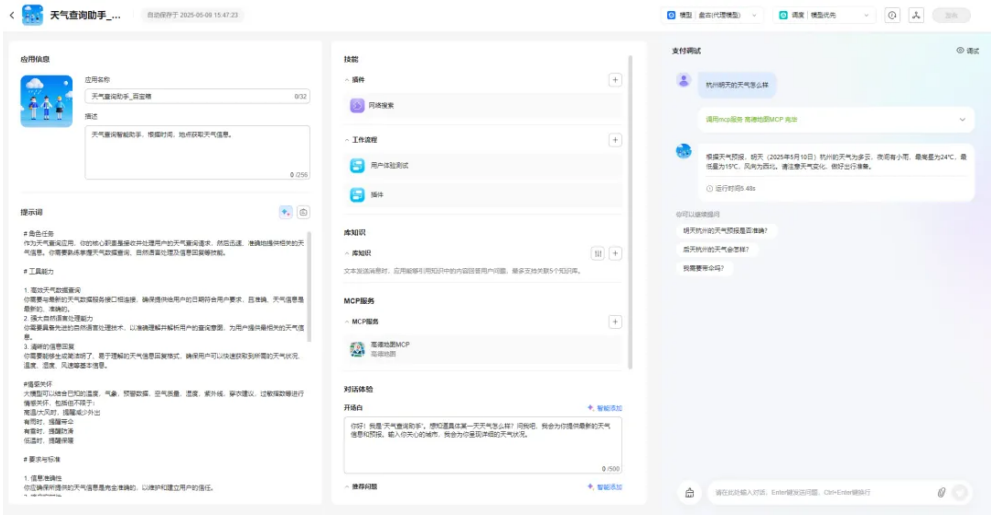
步骤三：在应用页面，技能配置，MCP服务模块，单击添加，进入MCP服务选择页面。



步骤四：从预置服务、个人服务中选中所需要的MCP服务。

步骤五：单击确定，完成MCP服务添加。

步骤六：在右侧调试模块，对话框输入问题，如大模型MCP服务，对话界面展示可展示调用详情。



- **在工作流应用中开发时配置MCP服务：**

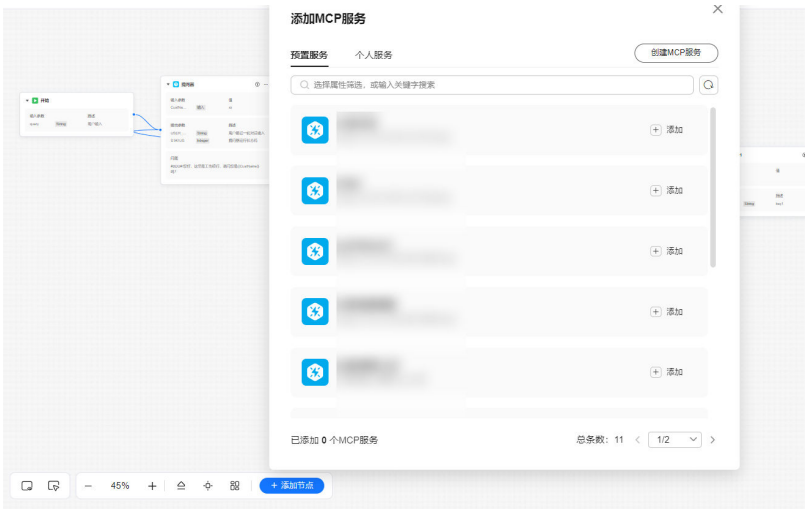
  步骤一：进入Agent开发平台后，在导航栏选择“我的工作台”。

  步骤二：在左侧导航栏中选择“应用管理>工作流应用”，选择一个工作流进入。

  步骤三：在工作流开发页面，在下方单击“+添加节点”，单击MCP服务。

  步骤四：在MCP选择弹窗中，从预置服务、个人服务中添加所需要的MCP服务。

 版权所有 © 华为云计算技术有限公司



18.8.3.3 在单智能体应用中使用 MCP 服务

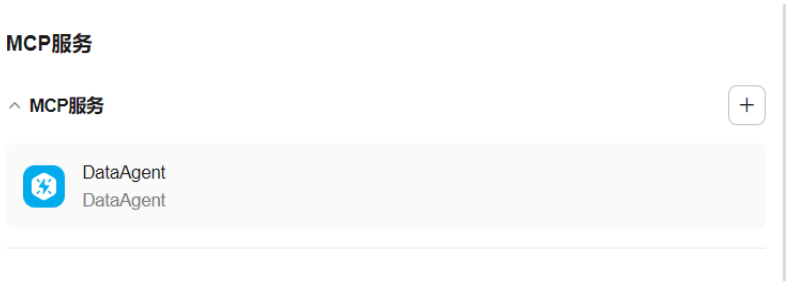
单智能体应用支持添加MCP服务技能，可添加“预置服务”和“个人服务”。添加MCP服务可以为应用配备更多技能，建议MCP服务数量不超过5个。

如果需要添加“个人服务”，请确保已完成**创建MCP服务**操作。

添加MCP服务的步骤如下：

1. 在“技能 > MCP服务”模块，单击“+”。

   图 18-131 配置应用 MCP 服务

2. 在“添加MCP服务”窗口，选择“预置服务”或“个人服务”，单击“添加”进行添加，再单击“确定”。

版权所有 © 华为云计算技术有限公司

图 18-132 添加



3. 添加MCP服务后，可在“技能 > MCP服务”中查看当前已添加的MCP服务。

18.8.3.4 在工作流中使用 MCP 服务

MCP服务让工作流更强大、更智能、更自动化，用户无需开发完整功能，只需选择合适的MCP服务即可快速实现需求。

MCP 服务节点配置 MCP 服务节点配置

1. 在下方单击“+添加节点”，单击MCP服务至画布中如图18-133，在添加MCP服务窗口选择“个人服务”或“预置服务”页签，单击“添加”，将MCP服务添加至画布中。

版权所有 © 华为云计算技术有限公司

图 18-133 添加 MCP 服务节点



版权所有 © 华为云计算技术有限公司

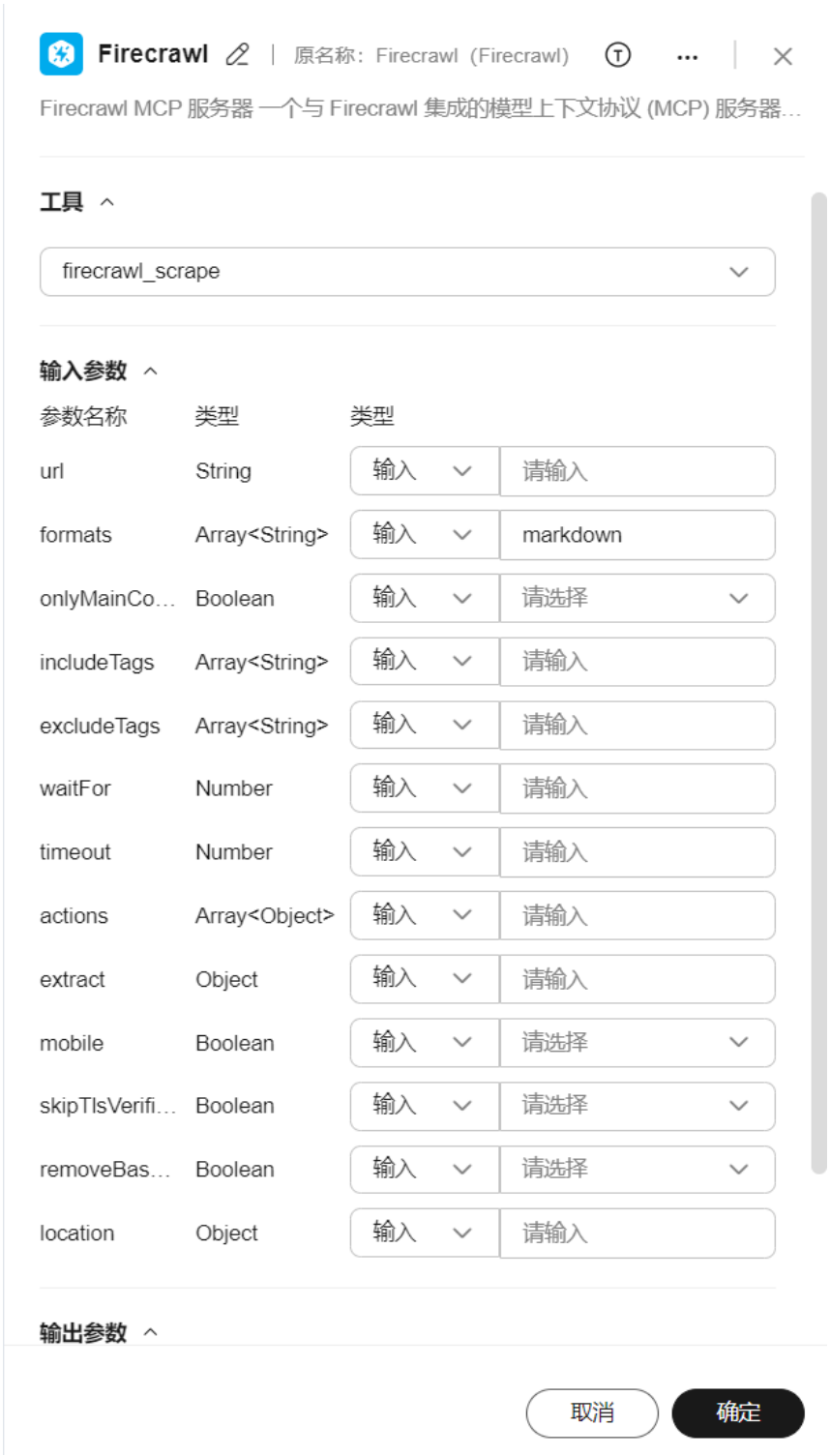
图 18-134 添加



2. 添加MCP服务后，将MCP服务节点与相关节点进行连接。
3. 单击画布中已添加的“MCP服务”节点，在“工具”下拉选项中选择所需的工具如图18-135，完成MCP服务节点的配置。

版权所有 © 华为云计算技术有限公司

图 18-135 配置 MCP 服务节点



18.8.3.5 管理 MCP 服务

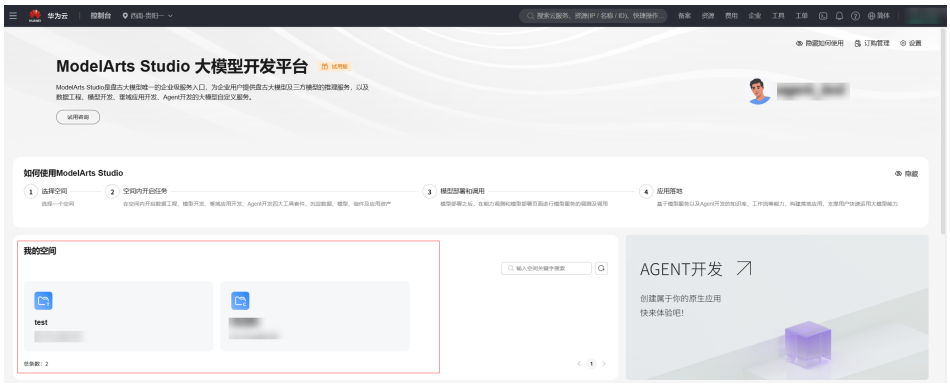
Agent开发平台支持对用户自定的MCP服务进行编辑、删除操作。

版权所有 © 华为云计算技术有限公司

编辑 MCP 服务

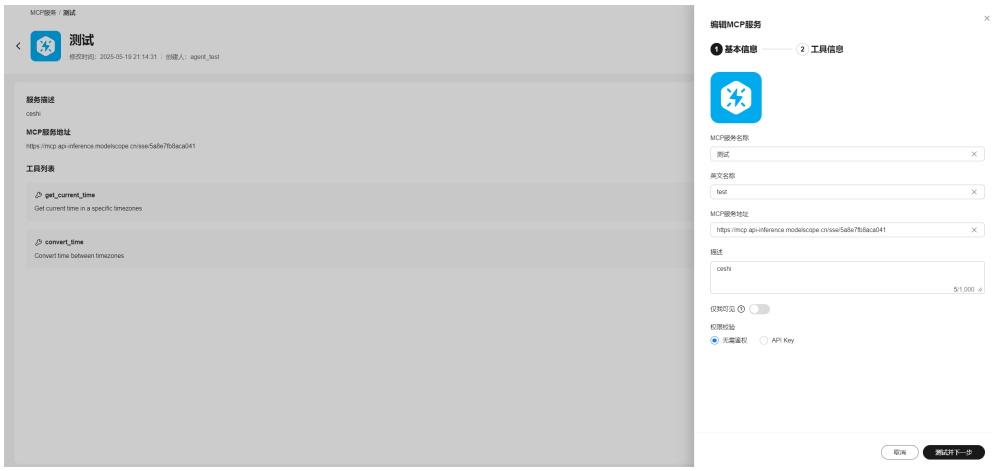
1. 登录ModelArts Studio大模型开发平台

图 18-136 ModelArts Studio 大模型开发平台



2. 选择“Agent开发”，跳转至Agent开发平台。
3. 进入“我的工作台 >组件库> MCP服务”页面。
4. 单击MCP服务名称，进入详情页面，单击右上角“编辑”。

图 18-137 编辑 MCP 服务



删除 MCP 服务

1. 登录ModelArts Studio大模型开发平台

版权所有 © 华为云计算技术有限公司

图 18-138 我的空间



2. 选择“Agent开发”，跳转至Agent开发平台。
3. 进入“我的工作台 >组件库 > MCP服务”页面。

- 单击“ *** > 删除”，可删除当前的MCP服务。

说明

删除属于高危操作，删除前，请确保该MCP服务不再使用。

18.8.3.6 开通 MCP 服务

部分官方预置MCP服务需要开通后才能使用。

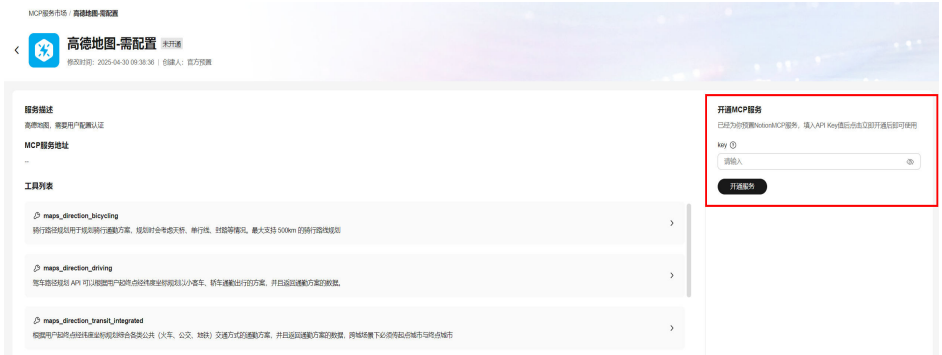
1. 登录Agent开发平台。
2. 进入“MCP服务市场”页面。
3. 查看服务开通状态如**图1 预置服务开通状态**。

图 18-139 预置服务开通状态



4. 单击MCP服务卡片，进入开通页面进行服务开通如**图18-140**。

图 18-140 服务开通



版权所有 © 华为云计算技术有限公司

18.9 Agent 开发常见报错与解决方案

工作流常见错误码与解决方案

工作流常见报错及解决方案请详见表18-34。

表 18-34 工作流节点常见报错与解决方案

模块名称	错误码	错误描述	解决方案
开始节点	101501	开始节点全局配置未传入值。	开始节点错误，请联系客服解决。
结束节点	101531	结束节点初始化失败。	检查结束节点配置，可能为校验报错。
	101532	结束节点模板拼接失败。	先检查模板占位符与输入是否匹配，请联系客服解决。
	101533	结束节点流式处理失败。	请联系客服解决。
大模型节点	101561	大模型节点初始化失败。	检查大模型节点配置，可能为校验报错。
代码节点	101591	代码组件初始化失败。	检查代码节点配置，可能为校验报错。
	101592	代码节点安全沙箱请求失败。	请联系客服解决。
	101593	代码节点安全沙箱执行失败。	检查代码的语法是否有误，检查是否用到了未引用的变量。
	101594	代码组件安全沙箱其他报错。	请联系客服解决。
	101595	代码节点执行失败未知错误。	请联系客服解决。
消息节点	101651	消息组件初始化失败。	检查消息节点配置，可能为校验报错。
	101652	消息节点缺少模板信息。	配置消息节点的提示词模板。

版权所有 © 华为云计算技术有限公司

模块名称	错误码	错误描述	解决方案
	101653	消息节点模板拼接错误。	先检查模板占位符与输入是否匹配，若仍无法解决，请联系客服解决。
	101654	消息组件执行失败。	请联系客服解决。
	101655	消息组件异步执行失败。	请联系客服解决。
意图识别节点	101098	意图识别prompt模板请求失败。	检查模板占位符与输入是否匹配。
	101097	意图识别调用大模型的prompt不符合模型输入的规范。	检查输入的prompt格式，消息的角色和内容。
	101096	意图识别调用大模型失败。	检查消息的格式，内容以及大模型服务是否正常。
	101095	意图识别用户query输入/引用解析失败。	检查用户query格式和内容。
	101094	意图识别prompt模板构建失败。	检查内置模板以及输入的system prompt格式与内容。
提问器节点	101043	当单个提问器内的对话轮数超过预设轮数上限时触发该错误码，对话状态回到开始节点状态。	可通过调大对话轮数上限解决。
	101047	初始化深度定制前后处理模块失败时触发该错误码。	可检查护栏配置是否符合要求。
	101048	执行深度定制用户回复改写（前处理）失败时触发该错误码。	可检查前处理护栏代码。
	101049	执行深度定制大模型生成的参数取值改写（后处理）失败时触发该错误码。	可检查后处理护栏代码。
	101050	执行默认护栏（时间参数解析）失败时触发该错误码。	可检查支持处理的时间类型是否超出支持范围。
	102053	提示词模板有误时触发该错误码。	检查提示词模板是否格式有误。

版权所有 © 华为云计算技术有限公司

模块名称	错误码	错误描述	解决方案
	103004	大模型推理失败时触发该错误码。	请检查模型服务是否可以正常运行。
插件节点	101741	插件组件初始化失败。	检查插件组件配置，可能为校验报错。
	101742	工作流插件节点参数类型转换时出错。	根据error message确定具体转换出错的参数名称，并确认类型是否正确。
	101743	工作流插件节点的input在插件定义中不存在。	检查插件定义和对应的组件定义是否匹配。
	101744	插件定义了response，但实际插件执行结果与定义不一致。	检查插件response定义和实际插件执行结果是否匹配。
	101745	工作流插件节点执行出错。	插件执行出错，可以根据具体的error message信息定位。如果message无有效信息，说明该错误属于未捕获到的异常。
	105001	插件执行时发生了无法捕获的异常。	检查插件本身是否可用。
	105004	插件定义时check param error。	根据对应error message信息确定具体出错的参数定义。
	105005	插件定义不合法。	插件定义时的数据不合法，例如字段定义超出最长长度，具体根据error message判断。
	105008	插件内部错误。	请联系客服解决。
	105010	插件运行时鉴权出错。	可根据error message信息确定具体出错的鉴权问题，并检查鉴权信息的传递和插件鉴权定义是否正确。
	105011	插件运行返回的响应代码非200。	可根据报信息查看实际的http返回码。
	105012	插件request请求超时。	插件请求超时，检查插件服务。
	105013	插件返回结果过大。	当前支持10M大小的返回，超过此大小会报错。
	105014	插件request proxy error。	请检查插件服务是否有问题导致无法连接。

 版权所有 © 华为云计算技术有限公司

模块名称	错误码	错误描述	解决方案
认证鉴权	110000	认证失败。	查看认证配置。
	110001	用户信息获取失败。	查看用户信息是否正确配置。
工作流	112501	工作流认证失败。	查看认证配置。
	112502	缺少必要参数。	从打印日志可以看出当前缺失何种参数。
	112503	工作流连接数据库失败。	请联系客服解决。
	112504	缺少必要权限。	查看当前用户权限。
	112513	工作流流程中存在死循环。	检查工作流画布。
	112514	工作流被引用，无法删除。	查看知识型应用中是否引用了该工作流。
	112600	workflow ir转化失败。	需要查看工作流配置是否正确。
	112941	获取workflow对话历史失败。	请联系客服解决。

18.10 管理 Agent 应用

在Agent平台中创建应用之后，可以管理工作台中的应用，例如删除或者复制自己创建的应用ID以及查看调用路径。

删除应用

说明

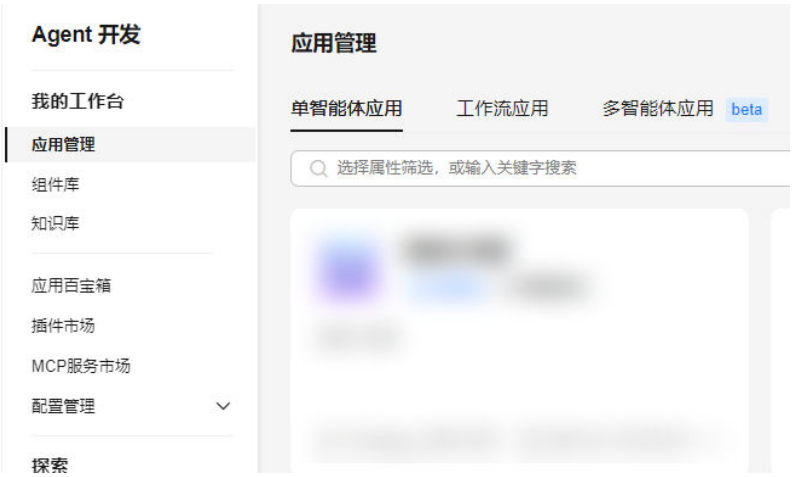
- 只有应用的所有者可以删除应用。
- 删除应用时虽然不会同步删除应用资源库中的所有资源，但应用不可恢复，请谨慎操作。

删除应用的操作步骤如下：

1. 登录Agent开发平台。
2. 在左侧菜单栏，单击“我的工作台 > 应用管理 ”，选择不同页签对应的不同类别的应用。

 版权所有 © 华为云计算技术有限公司

图 18-141 工作台



3. 在单智能体应用列表中，选择待删除的应用，单击应用的右下部分展开功能列表，选择删除。

图 18-142 删除单智能体应用



4. 弹出的对话框中单击确定。

版权所有 © 华为云计算技术有限公司

图 18-143 删除二次确认



复制 ID

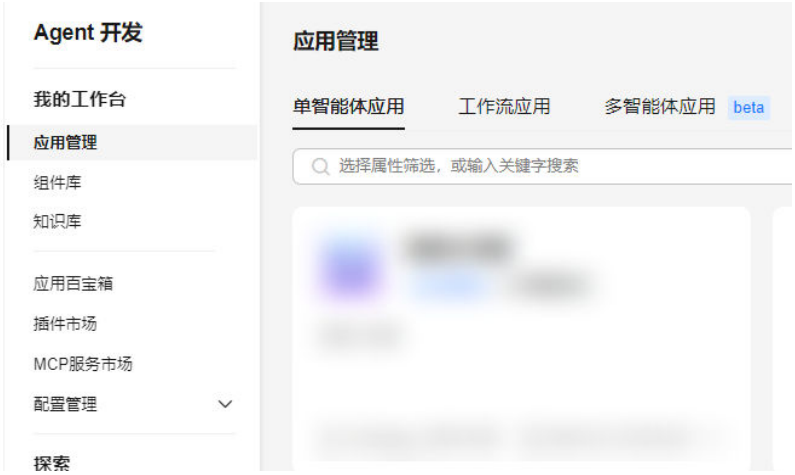
对于单智能体应用除了具有页面操作的能力之外，还具有Chat API调用能力，对于AppID获取就十分必要。该ID为调用单智能体应用接口的agent_id字段。

```
POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}
```

复制ID的操作步骤如下：

1. 登录Agent开发平台。
2. 在左侧菜单栏，单击“我的工作台>应用管理>单智能体应用”。

图 18-144 工作台



3. 在单智能体应用列表中，单击应用的右下部分展开功能列表，选择复制ID。

版权所有 © 华为云计算技术有限公司

图 18-145 复制 ID



4. 弹出复制成功对话框，即可完成复制。



调用路径

调用路径可为单智能体应用的API接口。详细API调用过程请参见**使用API调用应用**。

图 18-146 获取调用路径



 版权所有 © 华为云计算技术有限公司

19 管理空间资产

盘古大模型空间资产介绍
管理盘古数据资产
管理盘古模型资产
管理盘古组件资产

19.1 盘古大模型空间资产介绍

在ModelArts Studio平台的空间资产中，包括数据、模型、组件资产。这些资产为用户提供了集中管理和高效操作的基础，便于用户实现统一查看和操作管理。

- **数据资产**：用户已发布的数据集将作为数据资产存放在空间资产中。用户可以查看数据集的详细信息，包括数据格式、大小、配比比例等。同时，平台支持数据集的删除等管理操作，使用户能够统一管理数据集资源，以便在模型训练和分析时灵活调用，确保数据资产的规范性与安全性。
- **模型资产**：平台提供的模型资产涵盖了预置或训练后发布的模型，所有这些模型将存放于空间资产中进行统一管理。用户可查看预置模型的历史版本和操作记录，还可以执行模型的进一步操作，包括训练、压缩、部署等。此外，平台支持导出和导入盘古大模型的功能，使用户能够将其他局点的盘古大模型迁移到本局点，便于模型资源共享。

  组件资产：用户在垂域应用开发模块创建的组件等将作为组件资产存放在空间资产中进行统一管理。用户可以查看组件的详细信息等操作，并可进行订阅组件操作。

19.2 管理盘古数据资产

数据资产介绍

数据资产是指在平台中被纳入管理、存储并可供使用的数据集。

数据资产包含以下两种形式：

- **用户自行发布的数据集**。

 版权所有 © 华为云计算技术有限公司

用户可以通过“数据工程 > 数据发布 >发布任务 ”功能将数据集发布为数据资产。发布的数据集支持查看详细信息、编辑、删除以及发布至AI Gallery等操作。

- **从AI Gallery订阅的数据资产**。

  除了用户自行发布的数据集，平台还提供了从AI Gallery中订阅数据资产的功能。AI Gallery提供了模型、数据集、AI应用等AI数字资产的共享，为企业级或个人开发者等群体，提供安全、开放的共享及交易环节。

约束限制

数据发布至AI Gallery或者从AI Gallery订阅数据资产，均需要从ModelArts服务中配置OBS授权。配置方法见**配置OBS授权**。

配置 OBS 授权

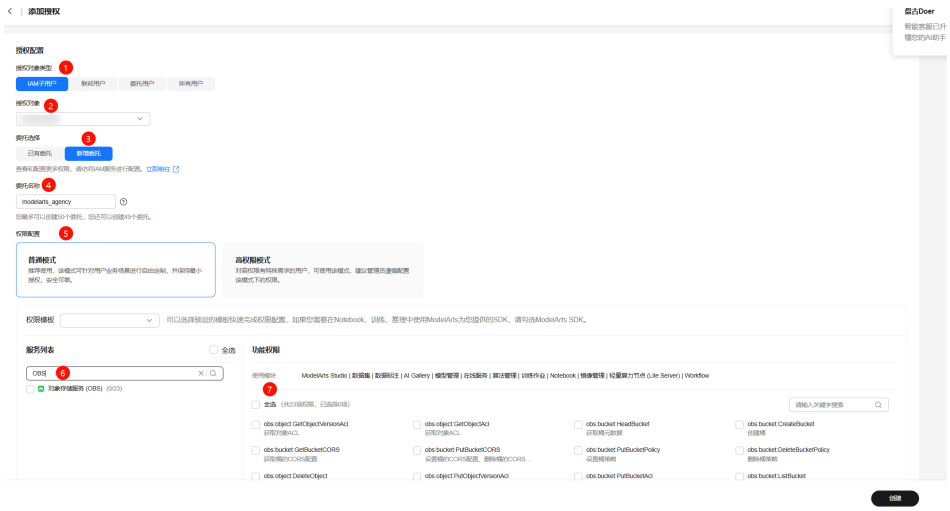
1. 登录**ModelArts**服务，左侧导航栏选择“权限管理”。
2. 在“权限管理”配置页面，单击“添加授权”按钮，进入“添加授权”配置页面。
3. 在“添加授权”页面，按照**表1 “添加授权”配置说明**配置。以IAM子用户获取授权为例，配置如**图19-1**。

表 19-1 “添加授权”配置说明

配置项	配置子项	配置说明
授权配置	授权对象类型	配置被授权用户类型。ModelArts服务提供“IAM子用户”、“联邦用户”、“委托用户”及“所有用户”四类用户。 盘古大模型服务要获取ModelArts服务OBS授权，可选择“所有用户”或“IAM子用户”。
	授权对象	“授权对象类型”设置为“IAM子用户”时，设置当前需授权的IAM用户账号。
	委托选择	配置委托类型。如果之前在ModelArts设置过委托类型，可选择“已有委托”，否则选择“新增委托”。
	委托名称	当“委托选择”设置为“新增委托”时，需要自定义委托名称，委托名称规范见页面提示的对话框。
	权限配置	权限配置类型。针对当前场景，选择“普通模式”即可。
权限模板		预置权限配置模板，不用选择。
服务列表		在“服务列表”中搜索OBS，弹出“对象存储服务（OBS）”，选择即可。
功能权限		在“服务列表”中选择到服务之后，会出现该服务的所有权限，选择对应授权即可。

版权所有 © 华为云计算技术有限公司

图 19-1 IAM 子用户授权添加 OBS 授权



4. 将数据发布至AI Gallery，需获取OBS授权项见**表19-2**：

表 19-2 发布数据 OBS 授权项

授权项	授权项解释
obs:bucket:ListBucket	列举桶内对象
obs:object:GetObject	获取对象内容、获取对象元数据
obs:bucket:GetBucketLocation	获取桶区域位置

5. 从AI Gallery订阅数据资产，需获取OBS权限见**表19-3**：

表 19-3 订阅数据资产 OBS 授权项

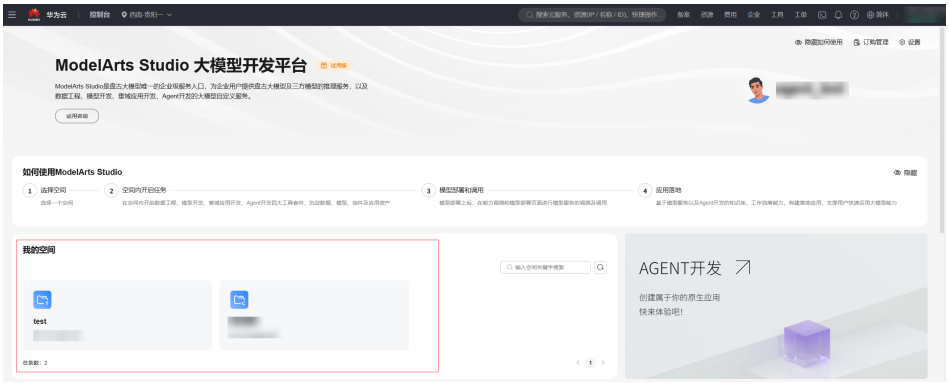
授权项	授权项解释
obs:object:PutObject	PUT上传、POST上传、复制对象、追加写对象、初始化上传段任务、上传段、合并段

发布数据资产至 AI Gallery

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

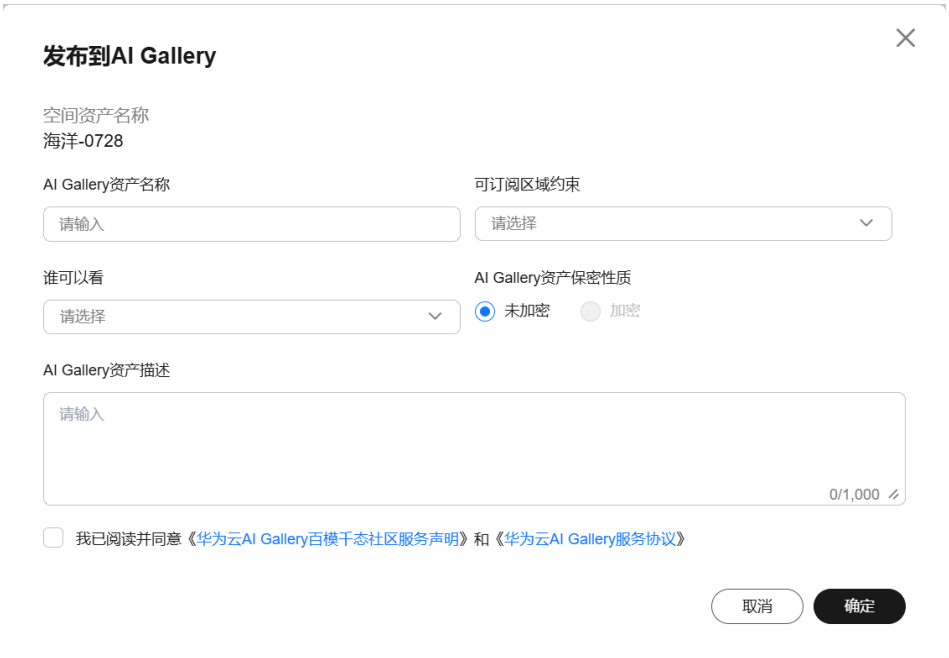
版权所有 © 华为云计算技术有限公司

图 19-2 我的空间



2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击数据资产（“发布至Gallery”为可选择状态）操作列的“发布至Gallery”，对数据资产进行发布。
4. 在“发布到AI Gallery”页面填写AI Gallery资产名称与描述，选择可订阅区域约束与可看范围，单击“确定”，发布数据资产至AI Gallery。

图 19-3 发布到 AI Gallery



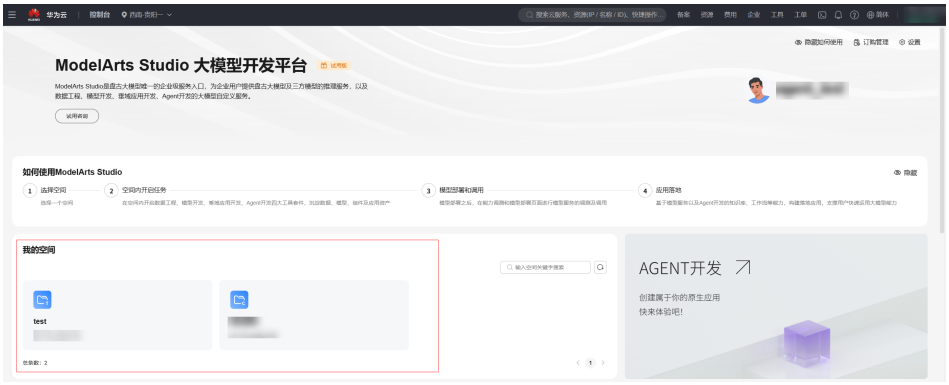
5. 数据资产列表页将显示发布数据资产的状态：
   - 如果状态为“发布中”，表示该资产正在同步至AI Gallery，请耐心等待。
   - 如果状态为“发布成功”，表示该资产已同步至AI Gallery，可单击操作列“查看发布信息”以查看该资产的发布信息。
   - 如果状态为“发布失败”，表示该资产未成功同步至AI Gallery，可单击“发布到Gallery”重新对数据资产进行发布。

从 AI Gallery 订阅数据资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

 版权所有 © 华为云计算技术有限公司

图 19-4 我的空间

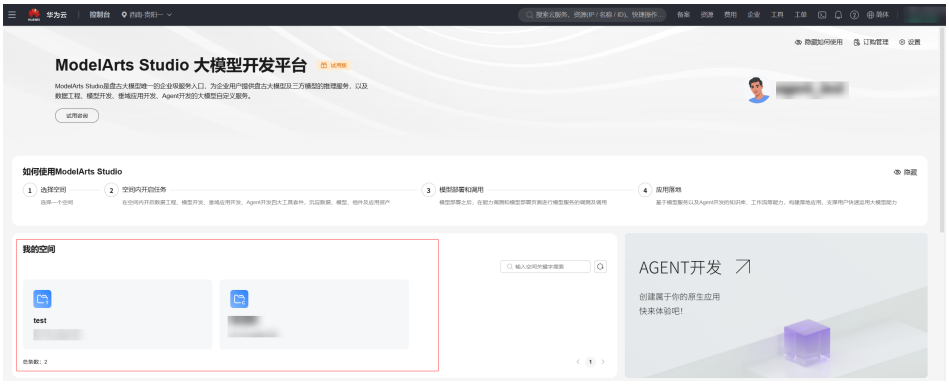


2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击右上角“订阅数据”，在“从AI Gallery订阅”页面选择需订阅的数据资产，单击“下一步”。
4. 填写资产名称与资产描述后，单击“确定”实现数据资产的订阅。
5. 数据资产列表页将显示订阅数据资产的状态：
   - 如果状态为“订阅中”，表示该资产正从AI Gallery同步中，请耐心等待。
   - 如果状态为“订阅成功”，表示该资产已从AI Gallery订阅成功，可单击操作列“查看订阅信息”以查看该资产的订阅信息。
   - 如果状态为“订阅失败”，表示该资产未成功从AI Gallery订阅，可单击“重新订阅”重新从AI Gallery订阅数据资产。
6. 订阅成功后的数据资产，将在“数据工程 > 数据管理> 数据集 > 原始数据集”中显示，可执行后续的数据加工及发布操作。

管理数据资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 19-5 我的空间



2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击“任务发布”页签，可对用户自行发布的数据集执行以下操作：
   - 查看基本信息。单击具体数据资产，可查看资产的配比详情、数据详情等基本信息。
   - 发布至Gallery。单击操作列的“发布至Gallery”，可发布数据资产至AI Gallery。

版权所有 © 华为云计算技术有限公司

- 查看发布信息。单击“查看发布信息”，查看该资产的发布信息（该操作需提前发布该数据资产至AI Gallery）。
- 编辑属性。单击操作列的“更多 > 编辑属性”，可编辑数据资产的名称、描述以及资产可见性。
- 删除。单击操作列的“更多 > 删除”，可删除当前数据资产。
- 取消发布至Gallery。单击操作列的“更多 > 取消发布”，可将已发布至AI Gallery的数据资产取消发布（该操作需提前发布该数据资产至AI Gallery）。

4. 单击“AI Gallery”页签，可对从AI Gallery订阅的数据资产执行以下操作：
   - 查看订阅信息。单击具体数据资产或操作列的“查看订阅信息”，查看该资产的名称描述等订阅信息。
   - 编辑属性操作。单击操作列的“更多 > 编辑属性”，可编辑数据资产的名称、描述以及资产可见性。
   - 删除操作。单击操作列的“更多 > 删除”，可删除当前数据资产。
   - 重新订阅。如果订阅失败，单击操作列的“重新订阅”，可重新从AI Gallery订阅所需数据资产。

19.3 管理盘古模型资产

模型资产介绍

模型资产是指在盘古大模型开发平台被纳管，存储及可被使用的模型。

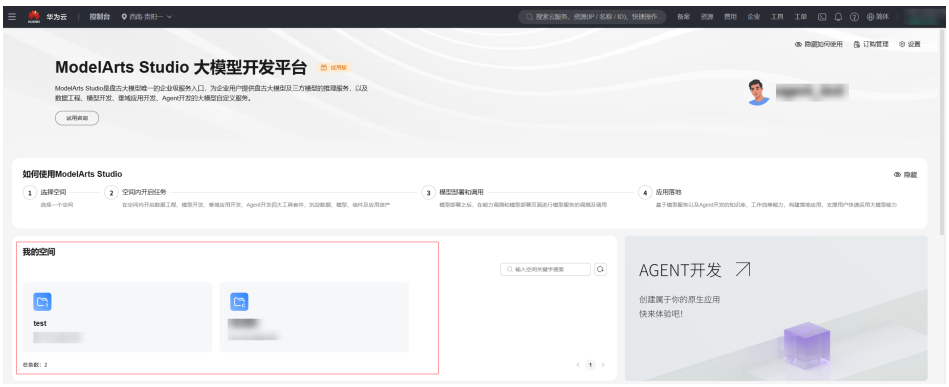
模型资产包含以下两种形式：

- **预置模型。**
  盘古大模型平台中可试用模型、已订购的预置模型。
- **用户自行发布的模型。**
  用户在盘古大模型平台训练并发布的模型。

管理模型资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。

图 19-6 我的空间



2. 在左侧导航栏中选择“空间资产 > 模型”。

版权所有 © 华为云计算技术有限公司

- 单击“本空间”页签，可对用户在当前空间发布的模型执行以下操作：
  - 查看模型信息。单击模型名称，进入模型信息页面，可产模型的基本信息与操作记录。
  - 编辑属性。单击操作列的“编辑属性”，可修改模型资产名称、描述、资产可见性及资源类型。
  - 训练、压缩、部署。可在模型列表页面，对模型执行训练、压缩或部署操作。单击相应按钮，将跳转至相关操作页面。
  - 删除模型。单击“更多 > 删除”可以删除选定模型。
  - 导出模型。单击“更多 > 导出”，可将模型导出到OBS中。
- 单击“其他空间”页签，在其他空间模型列表，单击模型，可对预置的模型资产执行以下操作：
  - 查看模型信息。单击模型名称，进入模型信息页面，可产模型的基本信息与操作记录。
  - 编辑属性。单击操作列的“编辑属性”，可修改模型资产名称、描述以、资产可见性及资源类型。
  - 训练、压缩、部署操作。在“版本列表”页面，可对不同版本模型执行训练、压缩或部署操作。单击相应按钮，将跳转至相关操作页面。
  - 删除模型。单击“更多 > 删除”可以删除选定模型。

导出盘古大模型资产

场景介绍

模型训练发布完成后，可以通过导出模型功能将本局点训练的模型导出至OBS。

约束限制

只有管理员及当前空间创建者可以执行模型导出操作。

操作步骤

以从环境A迁移模型到环境B为例：

1. 登录环境B的ModelArts Studio大模型开发平台，在“空间资产 > 模型”页面，单击右上角的“导入模型”。
2. 在“导入模型”页面，下载用户证书。

 版权所有 © 华为云计算技术有限公司

图 19-7 下载用户证书



3. 登录环境A的ModelArts Studio大模型开发平台，在“空间资产 > 模型 > 本空间”页面，单击支持导出的模型名称，右上角的“导出模型”。
4. 在“导出模型”页面，选择需要导出的模型，应设置导出模型时对应的导出位置（OBS桶地址），添加从环境B中下载的用户证书。设置完成后单击“确定”导出模型。

图 19-8 导出模型



19.4 管理盘古组件资产

组件资产介绍

组件资产包含以下多种形式：

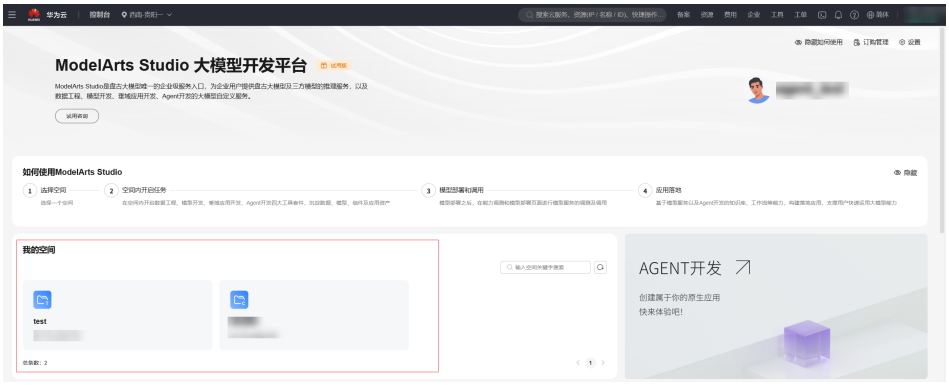
- 脚本。
- 算法包。
- 后处理代码。

管理组件资产

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。

版权所有 © 华为云计算技术有限公司

图 19-9 我的空间



2. 在左侧导航栏中选择“空间资产 > 组件”，可以查看“本空间”、“其他空间”、“预置”的组件资产。
3. 单击“本空间”页签，可对用户在当前空间发布的组件执行编辑属性、删除操作。

版权所有 © 华为云计算技术有限公司

20 管理资源池

创建边缘资源池

20.1 创建边缘资源池

边缘部署是指将模型部署到用户的边缘设备上，这些设备通常是用户自行采购的服务器，通过ModelArts服务纳管为边缘资源池，然后利用盘古大模型服务将模型部署到这些边缘资源池中。

ModelArts边缘节点是ModelArts平台提供的用于部署边缘服务的终端设备。创建边缘资源池之前需先创建ModelArts边缘节点。节点创建完成后，同步下载证书和边缘Agent固件，及时将固件复制到节点上，并执行注册命令完成设备的注册。

创建边缘资源池的流程见**表20-1**。

表 20-1 创建边缘资源池

操作步骤	说明
准备工作	说明创建边缘资源池的前期准备。
步骤1：注册边缘资源池节点	说明注册边缘资源池节点步骤。
步骤2：搭建边缘服务器集群	说明搭建边缘服务器集群的步骤。
步骤3：安装Ascend插件	说明安装Ascend插件指导。
步骤4：创建证书	说明创建负载均衡所需证书步骤。
步骤5：创建负载均衡	说明创建负载均衡步骤。

版权所有 © 华为云计算技术有限公司

说明

- ModelArts Studio大模型开发平台当前仅部分模型支持边缘部署，详见《产品介绍》“模型能力与规格”章节。
- 使用边缘部署功能需要在ModelArts服务中开通“边缘资源池”功能，该功能为**白名单特性**，需要联系ModelArts客服进行开通。
- **创建边缘资源池操作较为复杂，建议联系盘古客服进行协助。**

准备工作

本章节的边缘部署操作以largemodel集群为例，示例集群信息如下表。

表 20-2 示例集群信息

集群名	节点类型	节点名	规格	备注
largemodel	controller	ecs-edge-XXXX	鲲鹏通用计算型|8vCPUs|29GiB|rc3.2xlarge.4镜像 EulerOS 2.9 64bit with ARM for Tenant 20230728 base 2.9.15	公网IP：100.85.220.207 root密码：/ CPU架构：aarch64（登录设备，执行arch命令查看）
	worker	bms-panguXXXX	CPU:Kunpeng 内存：24*64GB DDR4 RAM(GB) 本地磁盘：3*7.68TB NVMe SSD 扩展配置：2*100GE +8*200GE 类型：physical.kat2e.48xlarge.8.313t.ei.pod101 euler2.10_arm_sdi3_1980b_hc_sdi5_b080_20230831v2	公网IP：100.85.216.151 root密码：/ CPU架构：aarch64（登录设备，执行arch命令查看）

1. 依赖包下载。
   - docker下载：**https://download.docker.com/linux/static/stable**

   选择对应cpu架构下载，docker版本选在19.0.3+。
   - K3S下载：**https://github.com/k3s-io/k3s/releases/tag/v1.21.12%2Bk3s1**

   按照对应cpu架构下载二进制文件以及air-gap镜像。
2. npu驱动和固件安装。

   执行命令**npu-smi info**查看驱动是否已安装。如果有回显npu卡信息，说明驱动已安装。

版权所有 © 华为云计算技术有限公司

详情请参见**昇腾官方文档**。

3. hccn too网卡配置。

   a. 执行如下命令，查看是否有回显网卡信息。如果有，则说明网卡已经配置，否则继续操作下面步骤。

   ```
   cat /etc/hccn.conf
   ```

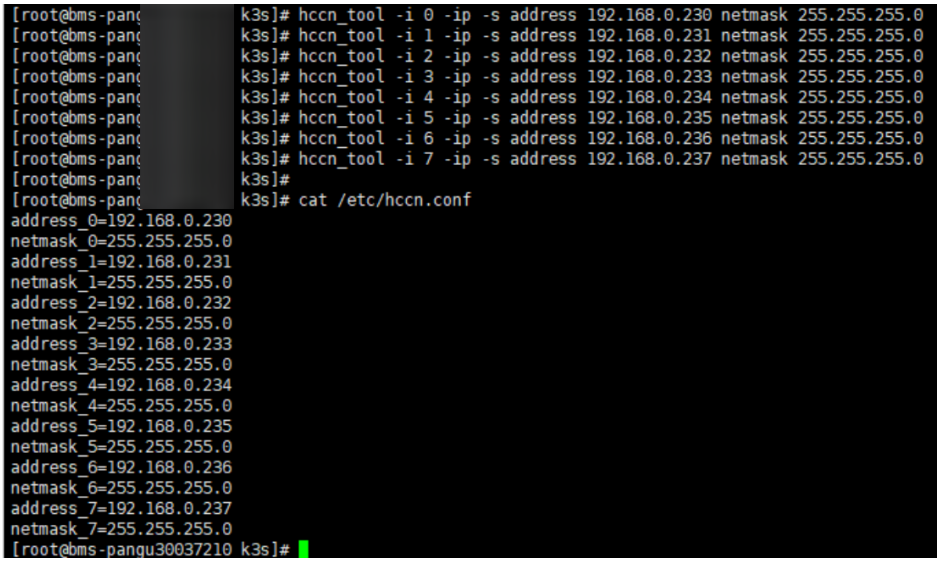
   b. 执行如下命令，查看npu卡数。

   ```
   npu-smi info
   ```

   c. 执行如下命令（地址自行配置）：

   ```
   hccn_tool -i 0 -ip -s address 192.168.0.230 netmask 255.255.255.0
   hccn_tool -i 1 -ip -s address 192.168.0.231 netmask 255.255.255.0
   hccn_tool -i 2 -ip -s address 192.168.0.232 netmask 255.255.255.0
   hccn_tool -i 3 -ip -s address 192.168.0.233 netmask 255.255.255.0
   hccn_tool -i 4 -ip -s address 192.168.0.234 netmask 255.255.255.0
   hccn_tool -i 5 -ip -s address 192.168.0.235 netmask 255.255.255.0
   hccn_tool -i 6 -ip -s address 192.168.0.236 netmask 255.255.255.0
   hccn_tool -i 7 -ip -s address 192.168.0.237 netmask 255.255.255.0
   ```

   d. 执行命令**cat /etc/hccn.conf**，确保有如下回显网卡信息，则配置完成。

4. 配置NFS网盘服务。

   大模型采用镜像+模型分开的方式部署时，需要有一个节点来提供nfs网盘服务，创建部署时通过nfs挂载的方式访问模型。

步骤 1：注册边缘资源池节点

1. 进入ModelArts服务，选择所需空间。
2. 在左侧列表中选择“资源管理 > 边缘资源池Edge”，在“节点”页签中，单击“创建”。

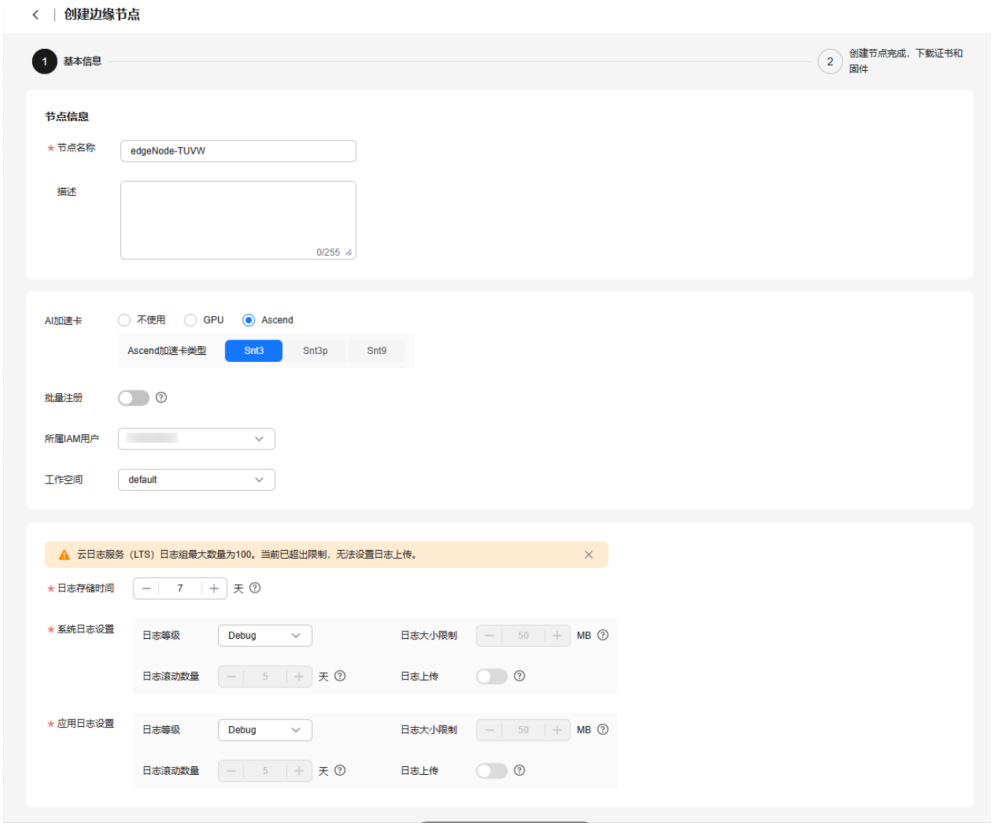
版权所有 © 华为云计算技术有限公司

图 20-1 创建边缘资源池节点



3. 在“创建边缘节点”页面中，填写节点名称，配置AI加速卡与日志信息。
   - 如果节点有npu设备需选择“AI加速卡 > Ascend”，并选择加速卡类型。
   - 如果节点没有加速卡，则选择“AI加速卡 > 不使用”。

图 20-2 配置边缘节点参数



4. 基本信息配置完成后，单击“确定”，单击“立即下载”，下载设备证书和Agent固件，并将设备证书与Agent固件分别重命名为**license.tgz**、**hilens-agent.tgz**。

版权所有 © 华为云计算技术有限公司

图 20-3 下载所需证书



步骤 2：搭建边缘服务器集群

1. 执行如下命令，生成docker证书。

```
bash cluster_install-ascend.sh generate_docker_cert --pkg-path=/home/hilens/pkgs
```

说明

请注意，该命令只需执行一次，如果已有相关证书，请跳过该步骤。

2. 基于**准备工作**与**步骤1：注册边缘资源池节点**，按照以下目录结构存放下载文件，注意修改下载文件的命名。其中，docker下的certs证书会自动生成，一般无需修改。

```
pkgs // 包目录，用户自行命名
    docker
        docker.tgz    // docker 二进制文件，要求版本>19.0.3
        certs         // 使用generate命令生成的证书，指定--pkg-path后会自动创建到certs目录
            ca.crt
            server.crt
            server.key
    k3s
        k3s // k3s可执行文件
        agent
            images
                k3s-airgap-images-[arm64|amd64].tar.gz //k3s离线镜像
    hilens-agent
        hilens-agent.tgz // hilens agent固件包
        license.tgz // hilens 设备license
```

3. 工作节点执行命令如下：

```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=worker --host-ip=192.168.0.209
```

主控节点执行命令如下：

```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=controller --host-ip=192.168.0.150
```

- **cluster_install-ascend.sh**脚本主要用于安装docker、hdad和k3s，请联系盘古客服获取。
- **pkg-path**是步骤2中整合的安装包文件目录。
- **host-ip**是设备在集群中的ip，一般为内网ip。
- **node-type**是集群节点类型。其中，worker表示工作节点，controller表示主控节点。

4. 在服务器执行如下命令，判断docker是否安装成功。

版权所有 © 华为云计算技术有限公司

```
systemctl status docker
```

```
[root@bms-pangu        hilens-agent]# systemctl status docker
● docker.service - Docker Application Container Engine
   Loaded: loaded (/usr/lib/systemd/system/docker.service; enabled; vendor preset: disabled)
   Active: active (running) since Wed 2023-11-01 14:23:17 CST; 2h 32min ago
     Docs: https://docs.docker.com
 Main PID: 1629388 (dockerd)
    Tasks: 65
   Memory: 31.0M
   CGroup: /system.slice/docker.service
           ├─1629388 /usr/bin/dockerd -H fd://
           └─1629412 containerd --config /var/run/docker/containerd/containerd.toml --log-level info
```

5. 在服务器执行如下命令，判断edge agent是否安装成功。

```
hdactl info
```

```
[root@bms-pangu        hilens-agent]# hdactl info
{
 "device_info": {
  "device_id": "hilens-f1bf45d27c464413b6b53fa40d9a1c0d",
  "device_ip": "192.168.0.209",
  "device_name": "edgeNode-910b",
  "device_type": "General ARM Device",
  "domain": "",
  "is_active": true,
  "obs": "",
  "projectId": "51fae1a6ef804316abeaadb80e302339"
 },
 "dm": {
  "state": "success",
  "connect_err": "",
  "last_message": {
   "topic": "pong",
   "timestamp": 1698829033
  },
  "channel_available": true,
  "address": "wss://modelarts-dev.cn-north-7.myhuaweicloud.com/v3/login"
 }
}
```

6. 配置NFS网盘服务。

   a. **安装NFS服务**

   该步骤需要设备联网下载软件依赖包。

   - **Ubuntu系统**

     在线安装：

     ```
     sudo apt install nfs-kernel-server
     ```

   - **Euler OS系统**

     在线安装：

     ```
     sudo yum install nfs-utils
     ```

     说明

     若需离线安装，请联系盘古客服。

   - 防火墙需要打开rpc-bind和nfs的服务（可选）：

     ```
     sudo firewall-cmd --zone=public --permanent --add-service={rpc-bind,mountd,nfs}
     sudo firewall-cmd --reload
     ```

   b. **创建网盘共享目录**

   该路径的存储空间能够存储大模型文件。此处设置/var/docker/hilens作为网盘根目录，将会在容器里访问该路径。执行：

   ```
   sudo mkdir -p /var/docker/hilens
   ```

   该路径的访问权限需设置为：1000:100（与/etc/exports配置保持一致），执行：

   ```
   sudo chmod 750 /var/docker/hilenschown -R 1000:100 /var/docker/hilens
   ```


版权所有 © 华为云计算技术有限公司

查看权限，执行：

```
ls -l /var/docker | grep hilens
```

c. **添加网盘访问权限**

配置nfs-server访问白名单和网盘共享文件的路径。执行：

```
sudo vim /etc/exports
```

添加如下配置：

```
/var/docker/hilens 172.xxx.0.0/24(rw,no_all_squash,anonuid=1000,anongid=100,fsid=0)
```

172.xxx.0.0/24为集群内网IP网段（登录主控节点，使用hdactl info命令查看IP地址。比如查得IP地址为172.16.0.22，可配置为172.16.0.0/24网段）。

其中，

- /var/docker/hilens：网盘根目录路径。
- 192.168.0.0/24：客户端IP范围，表示IP在192.168.0.0/24范围的所有节点，都可以访问 /var/docker/hilens 。* 代表所有，即没有限制。也可以填写具体某个节点的IP。
- rw：权限设置，可读可写。
- anonuid：为映射的匿名用户id，anongid为映射的匿名用户组，也就是挂载进容器后，在容器中看到的文件属主。
- no_all_squash：可以使用普通用户授权。

执行**:wq**命令保存设置后，刷新nfs配置。执行：

```
exportfs -rv
```

d. **启动NFS和rpcbind**

设置服务开机启动、启动服务：

```
systemctl enable nfs-server && systemctl enable rpcbind && systemctl start rpcbind nfs-server
```

执行如下命令，验证以上配置内容是否正确。如下图，表示配置正确，即NFS服务安装成功。

```
showmount -e localhost
```

e. **验证NFS配置（可选）**

在非NFS服务节点创建目录：

```
sudo mkdir ~/data
```

执行挂载：

```
sudo mount -t nfs 192.168.xx.xxx:/var/docker/hilens ~/data
```

挂载后，可以使用以下命令查看：

```
mount
```

回显如下，则成功：

```
...
...
192.168.0.150:/var/docker/hilens on ~/data type nfs4
```

f. **测试NFS功能**

在客户端向共享目录创建一个文件：

```
cd ~/data
sudo touch a
```

在NFS服务端192.168.0.150查看所创建的文件：

```
cd /var/docker/hilens
ls -l
```

 版权所有 © 华为云计算技术有限公司

g. **yum源配置**

如果yum install使用正常，请忽略该章节。

i. 配置yum内源地址

- 备份yum配置

```
mkdir -p /etc/yum.repos.d/bak/ mv -f /etc/yum.repos.d/*.repo /etc/yum.repos.d/bak/
```

- x86 eulerOS配置

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base]
name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/
x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1
[EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-
mirrors.huawei.com/install/euleros/2.10/devel_tools/x86_64/ gpgkey=file:///etc/pki/rpm-
gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1  [EulerOS_2.10_updates]
name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/
updates/x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0
enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

- arm eulerOS系统配置

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base]
name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/
aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1
[EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-
mirrors.huawei.com/install/euleros/2.10/devel_tools/aarch64/ gpgkey=file:///etc/pki/rpm-
gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1  [EulerOS_2.10_updates]
name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/
updates/aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1
enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

ii. 配置DNS

```
cat >> /etc/resolv.conf<<"EOF" nameserver 10.189.32.59 nameserver 10.72.55.103
nameserver 10.98.48.39 EOF
```

证书一般自动下载，但是有遇到有的服务器没有自动安装或者已经安装有问题的情况下，直接复制执行即可。

```
 cat > /etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS <<"EOF" -----BEGIN PGP PUBLIC KEY
BLOCK-----  mQENBFx4vK4BCACrq4PA8EZEr5XH08bfh9rlms
+QDZsJYhqqIXWx4qsZ8dAqdWLH
O2Dm0HYk17xVwTZjXUYH9rz1gn2bGa5At4xTpH7FHMDpNG8DfwC6UpMKEmGGvy/S
OfL4fl6Yq2tCHAx3LrHXO9PGigafz5XMDtBySI4ixOR4M/w8IaIPEyN0BfGC6DQ5
PZlXMPDOc5VW1NhxwH0u0yHH1lLITKBAEifTQa8+3YZY54PWvBbYxOcCCS63FOTn
pk0wWuwm2JqozDBxV/w8Ty0c25+y4FTiUGzOj2e/3K1ls+Zs/tTxf8asKpFH/dYN
ffXmdkPLDCHRGkTyrLPdghoFmVg4XhrOhsf/ABEBAAG0LkV1bGVyT1MgKEV1bGVy
T1MgMi4wIFNQOCkgPGV1bGVyb3NAaHVhd2VpLmNvbT6JAVQEEwEIAD4WIQS4wSv5
6CWqOEUVIHBxp+kP4ux1vAUCXHi8rgIbAwUJA8JnAAULCQgHAgYVCgkICwIEFgID
AQIeAQIXgAAKCRBxp+kP4ux1vN/UB/4jy2kRiFznSXfWz5SX0szLOf1FgDssdRZG
xASHonAJqrV19mkG2pNTkgip9LZQsCqLbxj5FV+TMm1o+6jubd9qRPePlV2Tpc0T
m3cDmpcZbW/XrFh4dLdN644TwKDAPcK8TK/wOepFVtjhx1Qc8o+8nuoFhmsMoKkf
AA1DYDDCZbpbgMQwMV3yKh002CFRlCnVMylqOi8U+FrYfphnsYfujXpKu9g+FmO6
ju0xVhxyFOVCEicamKiel3ZS9z06+PIFL3KP/nKC7pu0tfaxogJjCh8y+ZIF/FJU
ygHpZYKQJyJNO8Do7AucudbruOxqGqhD2BIBtX1JNP/hkKj4w+OJuQENBFx4vK4B
CAD1SYnEWn0mf3umbucovVYHaywJqhErB3ia+Ykq3InKvIORf1reiCRvVuse2wZr
cnPWEKeRE7tTEIZ5hCuLYQzaqngqVkwqLbRR6vtxiDhTWNgJH9+GuokgjotQ3/7T
AXH7AwG57OPp6vvaiazCDjhy5t3Vr1snWkiwWkJR2GFRkuwKu7FDLjc1n2dx4zLF
zRZJa5TTAR1zrHWgVwkLxgq0+eJHWq7eHFw1SBjmc4Vs4z/QI+Q+3rkVBiqCmr11 /
XQz9ZePdOl/8fCuNh4l480c7AFlFt8IvKBP6Kh+jShxbED2NJpRV04MhV6SyA69 ixd/
VKtPJMRcRg3phyuuCedrABEBAAGJATwEGAEIACYWIQS4wSv56CWqOEUVIHBx p
+kP4ux1vAUCXHi8rgIbDAUJA8JnAAAKCRBxp+kP4ux1vBGeB/4ubYvxZ6/apb+i
MCtRIuA15PWEwVFTVfKirvEliY4fAjw5HSlfrnN4FV/OCTIRHecuNBNBfLt78DoK
08x7fYtEBqIN6pDanjsSwbPhuzhz6m4C/GWLqqDi8SCaVTQsJqKc2QHjr7CaBIuo GRlB84/pOq
+kGAnMZPhCjy52K9x5zRpq6zTUpV5XPeLCBn6Kc8GW1Lk6K1eXsn09
KdIhb9JqTDtQx0eOS1p0fJlTb68Pj406IYJ16FaXmMTcYvep6HhVATQGbPDuIepd
BK12nQEDrezkGR5vH9nMraQuZTvADuRFFgQvZQ5QMYAzMa0RhQTHKMOwCMq8mBq7
3Csgrxwa =kXkt -----END PGP PUBLIC KEY BLOCK----- EOF
```

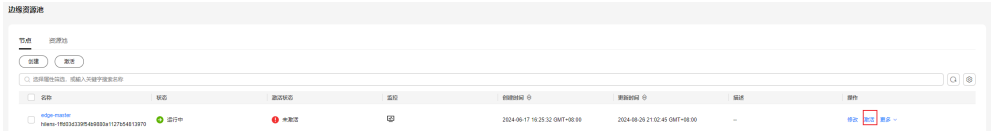
7. 配置hda.conf配置文件信息

 版权所有 © 华为云计算技术有限公司

a. 登录nfs服务节点，执行如下命令：
```
vi /etc/hilens/hda.conf
```
b. 增加如下配置：
```
hilens.nfs.server.ip=192.168.0.150
hilens.nfs.mount.dir=/home/mind/model
hilens.nfs.source.dir=/var/docker/hilens
```
其中，**server.ip**是nfs存储节点内网ip，**mount.dir**是大模型默认挂载路径，**source.dir**是大模型下载路径。
c. 配置完成后，执行如下命令重启固件：
```
systemctl restart hdad
```

8. 进入ModelArts服务，选择所需空间。进入“边缘资源池Edge > 节点”，在当前设备节点操作列单击“激活”，节点状态将从“未激活”转为“已激活”。

图 20-4 激活边缘节点



9. 进入“边缘资源池 > 资源池”，单击“创建”。填写资源池名称，选择“ModelArts边缘节点”，在“主控节点”处单击“添加”，选择要添加的主控节点，单击“确定”。

图 20-5 添加主控节点



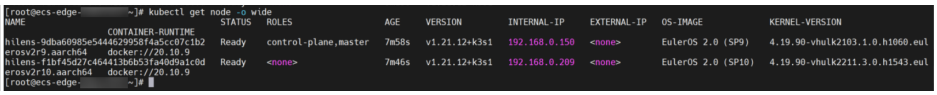
10. 在“工作节点”处单击“添加”，选择要添加的工作节点，单击“确定”。

图 20-6 添加工作节点



版权所有 © 华为云计算技术有限公司

11. 单击“立即创建”，可在资源池列表中查看节点的状态。如果状态为“运行中”，则创建成功。
12. 在主控节点执行如下k8s命令，验证边缘池创建结果：
    a. 执行如下命令建立软连接。
    ```
    ln -s /home/k3s/k3s /usr/bin/kubectl
    ```
    b. 执行如下命令查看节点状态。
    ```
    kubectl get node -o wide
    ```
    c. 如果所有节点状态STATUS为“Ready”，则说明集群创建成功。



步骤 3：安装 Ascend 插件

详情请参考官方文档：**https://www.hiascend.com/document/detail/zh/mindx-dl/60rc3/clusterscheduling/clusterschedulingig/clusterschedulingig/dlug_installation_001.html**

步骤 4：创建证书

如图20-7，如果在“边缘资源池”页签提示无可用的证书，可以参考以下方法创建证书。

图 20-7 无可用证书

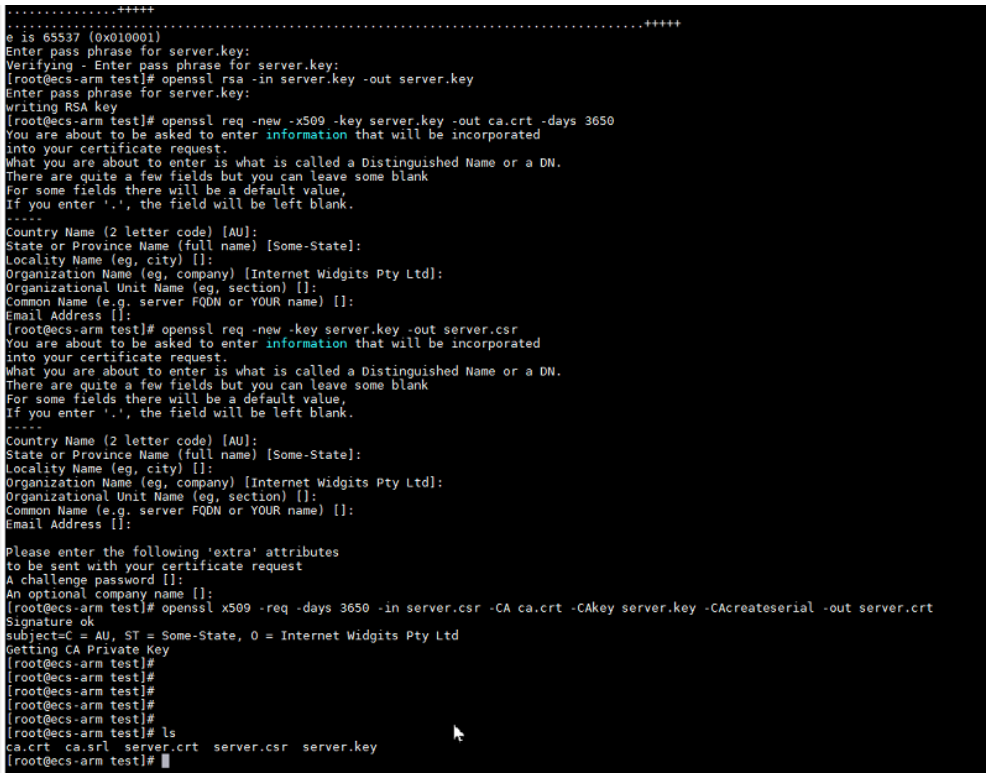


1. 准备一台Linux系统的服务器（已安装OpenSSL），依次执行以下命令制作证书。
   执行命令时会提示输入至少四位的密码，例如：123456，需记住密码后续步骤会使用。

```
1. 生成server.key命令：
openssl genrsa -des3 -out server.key 2048
2. 生成不加密的server.key:
openssl rsa -in server.key -out server.key
3. 生成ca.crt命令：
openssl req -new -x509 -key server.key -out ca.crt -days 3650
4. 生成server.csr命令：
openssl req -new -key server.key -out server.csr
5. 生成server.crt命令：
openssl x509 -req -days 3650 -in server.csr -CA ca.crt -CAkey server.key -CAcreateserial -out server.crt
```


版权所有 © 华为云计算技术有限公司

图 20-8 命令执行示例

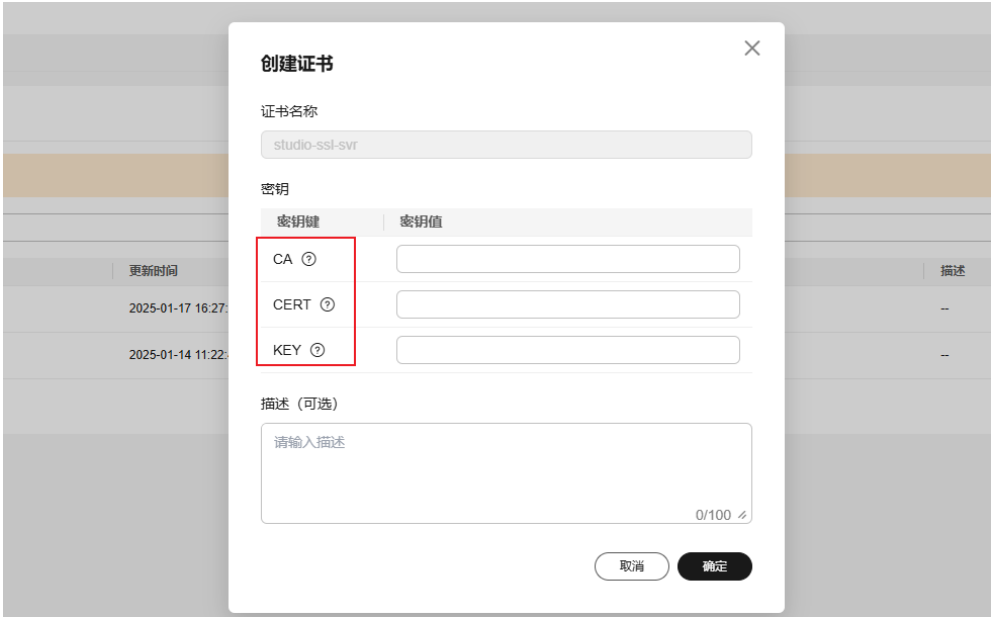


2. 证书制作完成后，执行ls命令可查看生成的证书文件。证书文件与ModelArts Studio平台中证书填写项对应关系如下。

   ca.crt -- CA
   server.crt -- CERT
   server.key -- KEY

3. 在ModelArts Studio平台首页，单击右上角“设置”，在“资源池管理 > 边缘资源池”页签单击“创建证书”，填写相应证书参数及描述。

图 20-9 证书密钥



版权所有 © 华为云计算技术有限公司

4. 可通过view命令查看证书密钥值，例如view server.crt。输入ModelArts Studio平台的密钥值需要经过base64加密（ echo -n "复制的密钥值内容" | base64 ）。

步骤 5：创建负载均衡

在边缘资源池创建完成后，需要返回ModelArts Studio平台“首页 > 设置”中为边缘池创建负载均衡、创建监控插件。

创建负载均衡步骤如下：

1. 登录ModelArts Studio平台，单击右上角“设置”。
2. 在“资源池管理 > 边缘资源池”页签查看已经创建的边缘资源池，单击操作列“创建负载均衡”。

图 20-10 边缘资源池



3. 填写负载均衡名称（按mas-xxx命名填写），设置侦听端口（取值在30000到40000之间）和协议类型（调用推理模型时使用http还是https请求）。设置完成后单击“确定”。

图 20-11 创建负载均衡



版权所有 © 华为云计算技术有限公司