智果（AgentArts）智能体平台

低代码开发

文档版本 01

发布日期 2026-09-01



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 2026。 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 应用选型指南 1
2 低代码应用开发流程 6
3 开发单智能体应用 10
3.1 单智能体应用介绍 10
3.2 示例：搭建一个医疗问诊助手智能体应用 14
3.3 创建并配置单智能体应用 23
3.3.1 创建单智能体应用 23
3.3.2 选择并配置模型 27
3.3.3 配置提示词 29
3.4 为应用添加技能 34
3.4.1 添加 MCP 服务 34
3.4.2 添加插件 37
3.4.3 添加工作流 40
3.5 为应用添加知识库 42
3.6 为应用添加记忆 46
3.7 提升应用对话体验 49
3.8 调试应用 54
3.9 配置触发器 61
3.10 发布应用 63
3.10.1 发布并部署应用为 API 服务 63
3.10.2 发布应用为网页应用 68
3.10.3 发布应用至云商店 70
3.11 通过 API 调用单智能体应用 77
3.12 管理应用 77
4 开发工作流应用 84
4.1 工作流介绍 84
4.2 对话型工作流和任务型工作流 86
4.3 工作流使用限制 87
4.4 工作流搭建范式 87
4.4.1 工作流搭建范式简介 87
4.4.2 顺序链式执行 89
4.4.3 条件路由执行 95

 版权所有 © 华为云计算技术有限公司

4.4.4 并行聚合执行 111
4.4.5 循环迭代执行 120
4.4.6 嵌套编排执行 127
4.4.7 反思精炼循环 142
4.5 搭建工作流 151
4.5.1 创建工作流 151
4.5.2 试运行工作流 164
4.5.3 配置触发器 169
4.5.4 发布并部署工作流 171
4.6 使用工作流 180
4.6.1 通过 API 调用工作流 180
4.6.2 在单智能体应用中使用工作流 181
4.6.3 在多智能体应用中使用工作流 182
4.7 管理工作流 185
4.8 基础节点 195
4.8.1 开始节点 195
4.8.2 结束节点 197
4.9 通用节点 200
4.9.1 大模型 200
4.9.2 工作流 209
4.10 逻辑节点 213
4.10.1 判断 213
4.10.2 代码 217
4.10.3 循环 222
4.10.4 意图识别 229
4.10.5 高级意图识别 233
4.11 工具节点 241
4.11.1 插件 241
4.11.2 MCP 服务 248
4.11.3 HTTP 请求 252
4.12 消息管理节点 257
4.12.1 消息 257
4.12.2 输入 263
4.12.3 提问器 270
4.12.4 问答 280
4.12.5 异常 289
4.13 变量&知识节点 294
4.13.1 变量赋值 295
4.13.2 变量聚合 311
4.13.3 知识检索 320
4.14 节点配置管理 328
4.14.1 意图管理 328

 版权所有 © 华为云计算技术有限公司

4.14.2 消息模板 332
4.14.3 对象管理 335
5 开发多智能体应用 338
5.1 多智能体应用介绍 338
5.2 多智能体执行范式 339
5.3 创建并配置多智能体应用 350
5.4 调试多智能体应用 365
5.5 部署多智能体应用为 API 368
5.6 通过 API 调用多智能体应用 373
5.7 管理多智能体 375
6 组件库 381
6.1 插件 381
6.1.1 插件介绍 381
6.1.2 示例：创建联网搜索插件 384
6.1.3 创建插件 390
6.1.3.1 基于 API 创建插件 390
6.1.3.2 基于 JSON 文件导入 API 插件 404
6.1.4 调试发布插件 409
6.1.5 使用插件 410
6.1.5.1 在单智能体中使用插件 410
6.1.5.2 在工作流中使用插件 412
6.1.6 管理插件 415
6.2 Skill 418
6.2.1 Skill 介绍 418
6.2.2 示例：导入旅游规划 Skill 422
6.2.3 导入 Skill 434
6.2.4 管理 Skill 437
6.2.5 Skill 常见问题 438
6.3 MCP 440
6.3.1 MCP 介绍 440
6.3.2 示例：使用模板创建搜索 MCP 441
6.3.3 示例：快速接入 ModelScope 魔塔社区 MCP 工具集 445
6.3.4 创建 MCP 450
6.3.4.1 自定义接入 MCP 450
6.3.4.2 基于模板创建 MCP 455
6.3.5 调试 MCP 457
6.3.6 使用 MCP 459
6.3.6.1 在单智能体中使用 MCP 459
6.3.6.2 在工作流中使用 MCP 461
6.3.7 管理 MCP 463
6.3.8 故障排除 464
6.4 知识库 472

 版权所有 © 华为云计算技术有限公司

6.4.1 知识库介绍 473
6.4.2 知识库类型 476
6.4.3 知识库使用限制 476
6.4.4 创建平台知识库 478
6.4.4.1 创建知识库 478
6.4.4.2 上传文档 483
6.4.4.3 知识库命中测试 490
6.4.4.4 知识库分段 492
6.4.5 接入第三方知识库 497
6.4.5.1 第三方知识库分类 497
6.4.5.2 连接 General 知识库 498
6.4.5.3 连接 KooSearch 知识库 506
6.4.5.4 连接 RAGFlow 知识库 514
6.4.5.5 知识库命中测试 521
6.4.5.6 第三方通用知识库接入规范 523
6.4.6 使用知识库 533
6.4.6.1 在单智能体中使用知识库 533
6.4.6.2 在工作流中使用知识库 537
6.4.7 管理知识库 544
6.5 提示词 546
6.5.1 提示词介绍 546
6.5.2 撰写提示词规范 549
6.5.3 创建提示词 550
6.5.4 优化提示词 554
6.5.5 管理提示词 562
6.5.6 使用提示词 571
6.5.6.1 在单智能体中使用提示词 571
6.5.6.2 在工作流中使用提示词 575
7 模型 580
7.1 模型介绍 580
7.2 接入预置模型 581
7.2.1 免费模型与使用额度 581
7.2.2 接入预置的付费模型 582
7.3 接入自定义模型 587
7.3.1 自定义模型接入流程 587
7.3.2 接入模型供应商 588
7.3.3 接入模型服务 593
7.3.4 模型 API 接口规范 599
7.4 调测模型 624
7.5 配置模型路由策略 635
8 管理使用 AgentArts 的成员 643

 版权所有 © 华为云计算技术有限公司

9 授权管理 645
9.1 管理云服务授权信息 645
9.2 管理模型授权信息 646
10 资源与订阅 648
11 审计 650
11.1 支持云审计的关键操作 650
11.2 在 CTS 事件列表查看云审计事件 662
12 附录 666
12.1 开发常见报错 666
12.2 更换密钥的加密方式 797

 版权所有 © 华为云计算技术有限公司

1 应用选型指南

为满足从简单对话到复杂业务自动化的多样化需求，AgentArts提供了三种核心应用开发模式：**单智能体**、**工作流**和**多智能体**。

在开始搭建AI应用之前，您需要回答一个核心问题：**"我的任务，应该交给一个人做，一条流水线做，还是一个团队做？"**

平台提供了三种核心开发模式，它们的关系如表1-1所示。

表 1-1 三种应用开发模式关系

类型	类比	运行方式	复杂度
单智能体	一个全能员工	一个人完成所有对话	★
工作流	一条生产流水线	按固定步骤自动化执行	★★★
多智能体	一个协作团队	多个专家各司其职协作完成复杂任务	★★★★★

应用模式介绍

- **单智能体**

  单智能体是最基础的AI应用形态，一个单智能体能独立完成所有工作。单智能体由提示词驱动，借助大模型实现自主理解意图并动态规划任务步骤，调用知识库、插件、MCP服务等工具完成任务。

  图 1-1 运行机制

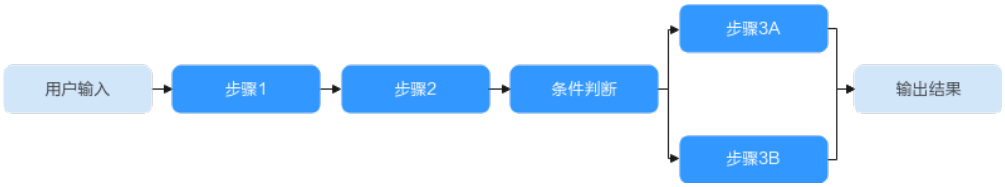
  - 适用场景：开放式对话应用，如智能客服、知识问答、任务助理、旅行规划等需灵活决策的场景。
  - 典型特征：AI自主决策，通过大模型根据提示词动态拆解任务，适应多变用户需求。

- **工作流**

版权所有 © 华为云计算技术有限公司

工作流是一种**确定性的自动化流水线，**将任务拆解为多个**预定义的步骤节点**。通过可视化节点编排，将多步骤任务串联为稳定、可复现的执行链路。

图 1-2 运行机制

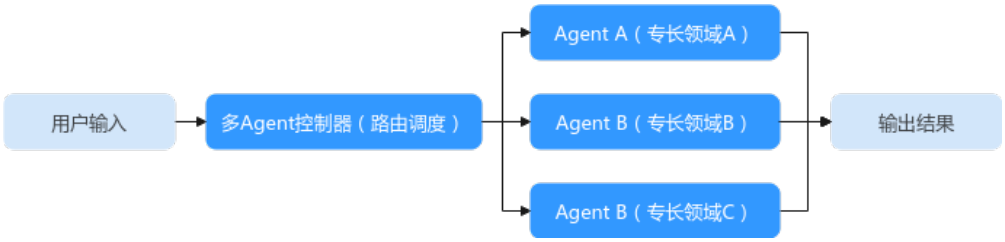


- 适用场景：固定流程自动化，如自动化报告生成、订单处理、多步骤审批流、数据标注等需精确控制的场景。
- 典型特征：预定义流程精确控制每一步，逻辑确定，确保任务执行的稳定性和可预测性。

- **多智能体**

  多智能体是由**多个独立的Agent协作**，通过角色分工与任务分配，解决复杂问题。

  图 1-3 运行机制

  - 适用场景：复杂任务协作，如跨部门业务流程自动化、多角色决策支持系统、综合性项目管理等。
  - 典型特征：多智能体协同，通过角色分工与任务分配，提升复杂任务的处理效率与准确性。

应用模式对比

表 1-2 三种应用对比

类型维度	单智能体	工作流	多智能体
开发方式	图形化操作，页面点选及文本输入配置，零代码。	画布组件拖拽+低代码开发。	图形化操作，无需编排，配置多智能体中央控制指令，并引用子智能体，定义分工即可。
面向用户特征	面向业务人员，无代码开发经验，办公场景相对固定。	面向技术人员，能够低代码开发工作流，调试各类专业化组件、API。	面向资深业务人员，不写代码，需对多业务场景进行整合。

版权所有 © 华为云计算技术有限公司

类型维度	单智能体	工作流	多智能体
适用场景	自主规划任务场景，场景泛化性好。	有固定的任务执行流程，高准确率要求。	需要多个智能体协同的复杂用户意图识别分工处理场景。
开发特点	快速简单、低门槛、规划准确率不稳定。	配置操作有门槛，执行成功率高，配置时间长。	相对简单、依赖已开发好的专家智能体。
能力约束	完全依赖模型自身能力，对插件数量、接口参数数量，执行步骤数量等限制较多。出现模型幻觉和不遵守相关规定的案例无法短期内解决。	编排后的流程执行过程中比较死板，智能化程度不高。对于异常场景需要配置相应的流程进行覆盖，会大幅提升流程配置的复杂度。	高度依赖中央控制模型的智能化水平，涉及的子智能体关联较多，另外当前无可视化调试能力，效果优化工作的调试难度较大。

选型决策指南

- **快速决策树**

```
你的任务需求是什么?
│
├── 任务相对简单，单一职责?
│   └── 选【 单智能体 】
│
├── 任务有明确的固定步骤/流程?
│   ├── 流程中需要条件判断、分支走向?
│   ├── 需要对输出格式、中间环节有严格控制?
│   ├── 对结果的确定性、一致性要求高?
│   └── 选【 工作流 】
│
├── 任务涉及多个专业领域?
│   ├── 不同问题需要不同的知识库/工具?
│   ├── 单个提示词已经很长很复杂、效果下降?
│   └── 选【 多智能体 】
│
└── 还是不确定?
    └── 先从【 单智能体 】开始，遇到瓶颈再升级
```

- **详细选型标准**
  - 选择单智能体

表 1-3 单智能体场景特征及示例

场景特征	示例
任务目标单一、职责明确	翻译助手、文案润色、代码解释
对话式交互，自由问答为主	闲聊机器人、通用问答助手
需要灵活应对各种不确定的用户输入	创意写作助手、头脑风暴助手
需要调用的工具/知识库较少（≤5个）	查天气+查日历的个人助手

版权所有 © 华为云计算技术有限公司

场景特征	示例
快速验证想法、最小可行产品阶段	任何场景的初期原型
对输出格式没有严格要求	非结构化的自然语言回答

- 选择工作流

表 1-4 工作流场景特征及示例

场景特征	示例
任务有明确的先后步骤	先提取信息→再分类→再生成回复
中间过程需要严格控制	必须先审核敏感词，通过后才能输出
需要条件分支走不同逻辑	用户是VIP走人工，普通用户走自动回复
对输出的格式、结构有严格要求	必须输出固定JSON格式的结构化数据
需要串联多个系统/接口	查库存→计算价格→生成订单→发送通知
需要在非LLM环节插入代码逻辑	数据清洗、格式转换、数学计算
需要最大程度节省Token	只在必要节点调用大模型，其余用规则处理

- 选择多智能体

表 1-5 多智能体场景特征及示例

场景特征	示例
涉及多个差异明显的专业领域	一个入口同时处理人事、财务、IT问题
不同领域需要不同的知识库/工具	人事用人事知识库，财务用财务系统
单个Agent的提示词过长导致效果下降	把3000字提示词拆成3个Agent各1000字
需要"专家会诊"模式	多角色协作：调研员+写手+审稿人
业务线多但希望统一入口	企业全能助手（覆盖HR/IT/行政/法务等）
子任务之间相对独立	各领域Agent可以独立开发和维护

- **按场景推荐速查**

版权所有 © 华为云计算技术有限公司

表 1-6 场景推荐速查示例

业务场景	推荐模式	推荐理由
企业知识问答	单智能体	单一知识域，对话式交互，无需复杂流程
角色扮演/AI陪练	单智能体	依赖人设提示词和多轮对话能力
每日新闻早报生成	工作流	固定步骤（搜索→筛选→总结→排版→推送）
批量处理数据/文件	工作流	需要循环节点遍历列表逐条处理
表单收集与审批	工作流	固定的流转步骤+人工审批节点
PPT/报告自动生成	工作流	多步骤生成，对格式有严格要求
电商综合客服	多智能体	需要售前/售后/投诉多角色分工
企业综合助手	多智能体	涉及HR、IT、财务等多领域，需路由分发
软件开发协作模拟	多智能体	PM→开发→测试多角色顺序执行
投资决策辅助	多智能体	多视角分析（乐观/悲观/中性Agent辩论）

选型常见误区

表 1-7 选型常见误区

误区	正确认知
多智能体一定比单智能体强	错。多智能体引入了协调成本和延迟，简单任务用单智能体反而效果更好、响应更快。
工作流不够智能	错。工作流中的每个大模型节点都可以做复杂推理，工作流的“确定性”恰恰是生产环境中的优势。
一个Agent加足够多的工具就能解决一切	危险。当工具超过10-15个时，模型选错工具的概率会急剧上升，此时应拆分为多智能体。
先搭多智能体，后续再简化	错。应该从单智能体开始，遇到瓶颈再升级。

版权所有 © 华为云计算技术有限公司

2 低代码应用开发流程

智果（AgentArts）智能体平台是面向企业级核心业务场景打造的智能体平台，基于Harness最佳实践构建标准化的开发与治理框架，专注于解决智能体在企业复杂环境下落地难、运行不可靠、运维无抓手等痛点，致力于降低应用门槛，助力企业将AI能力与实际业务深度融合，实现规模化落地。

单智能体应用开发流程

图 2-1 单智能体应用开发流程

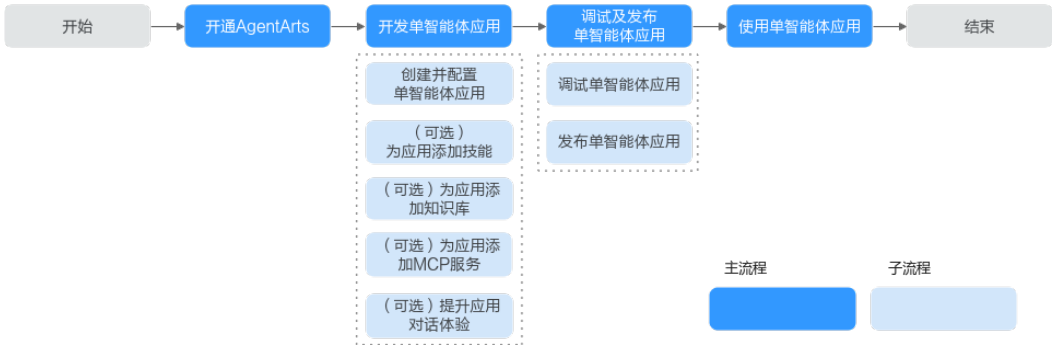


表 2-1 单智能体应用开发流程

流程	子流程	说明	操作指导
开通AgentArts服务	-	开通AgentArts服务后才可以正常使用AgentArts的功能。	开通AgentArts服务
开发单智能体应用	创建并配置单智能体应用	先创建单智能体应用，主要设置应用的名称、描述和图标。再配置单智能体的模型、提示词。	创建并配置单智能体应用
	（可选）为应用添加技能	技能包含插件、工作流、聊天记忆等，开发者可通过集成插件、设计工作流等方式不断扩展模型的能力边界。	为应用添加技能

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
	（可选）为应用添加知识库	知识库是智能体用于存储、管理和检索领域知识的核心组件，开发者可通过添加知识库为智能体提供精准的信息支持。	为应用添加知识库
	（可选）为应用添加记忆	在多轮对话场景中，用户往往期望智能体能够记住自己的偏好、身份和交互习惯，从而获得连贯且个性化的服务体验。在单智能体配置中，支持通过配置变量为应用开启“记忆”功能。	为应用添加记忆
	（可选）为应用添加MCP服务	开发者可以通过集成MCP服务快速拓展智能体的功能。	添加MCP服务
	（可选）提升应用对话体验	开发者可以通过配置智能体应用的开场白、推荐问题、追问、引用和归属、内容审核配置，提升应用的对话体验。	提升应用对话体验
调试及发布单智能体应用	调试单智能体应用	单智能体应用开发完成后，开发者可以通过调试应用从而精准定位问题并快速调整配置。	调试应用
	发布单智能体应用	单智能体应用调试完成后，需要发布才能被用户使用。	发布应用
使用单智能体应用	-	单智能体应用发布后，可以通过API接口调用。	通过API调用单智能体应用

工作流应用开发流程

图 2-2 工作流应用开发流程

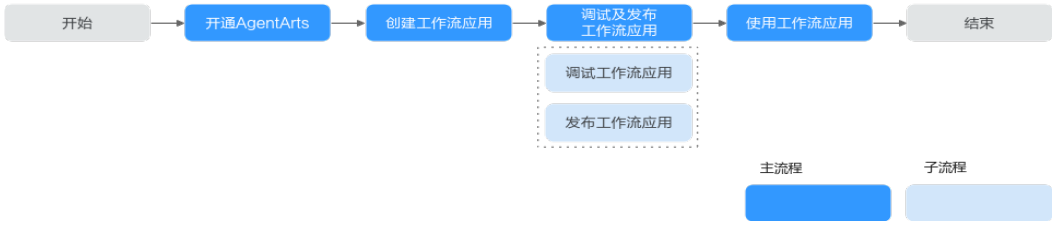


表 2-2 工作流应用开发流程

流程	子流程	说明	操作指导
开通AgentArts服务	-	开通AgentArts服务后才可以正常使用AgentArts的功能。	开通AgentArts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
创建工作流应用	-	创建工作流，包含全局配置、编排、选择节点、参数配置，对节点调试，完成功能连通。	创建工作流
调试及发布工作流应用	试运行工作流应用	开发者可以在工作流创建完成后，直接与工作流进行交互，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。平台提供的全链路调试功能，允许开发者查看每条用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而能够高效定位问题并快速调整配置。	试运行工作流
	发布工作流应用	工作流应用调试完成后，需要发布才能被用户使用。	发布并部署工作流
使用工作流应用	-	工作流应用发布后，可以在单智能体应用中使用，在多智能体应用中使用，还可以通过API接口调用。	• 通过API调用工作流 • 在单智能体应用中使用工作流 • 在多智能体应用中使用工作流

多智能体应用开发流程

图 2-3 多智能体应用开发流程

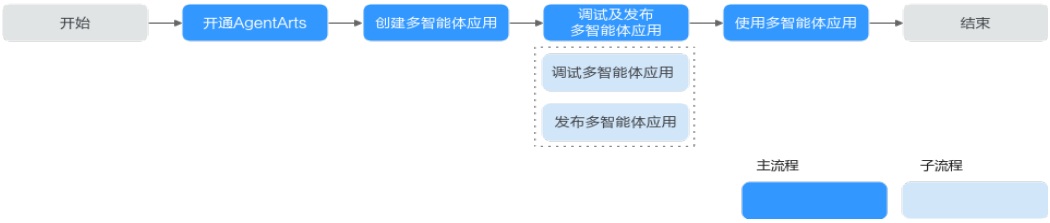


表 2-3 多智能体应用开发流程

流程	子流程	说明	操作指导
开通AgentArts服务	-	开通AgentArts服务后才可以正常使用AgentArts的功能。	开通AgentArts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
创建多智能体应用	-	多智能体应用可以灵活应用各种工作流，支持根据用户意图路由分发至对应工作流来完成用户任务。	**创建并配置多智能体应用**
调试与发布多智能体应用	调试多智能体应用	开发者可以在多智能体应用创建完成后，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。	**调试多智能体应用**
	发布多智能体应用	多智能体应用调试完成后，需要发布才能被用户使用。	**提交版本**
使用多智能体应用	-	多智能体应用发布后，可以通过API接口调用。	**通过API调用多智能体应用**

版权所有 © 华为云计算技术有限公司

3 开发单智能体应用

3.1 单智能体应用介绍

单智能体（Single Agent）是一个独立运作的AI实体，能自主感知环境、规划决策并执行任务，全程无需其他智能体协作。其核心特点是集中化处理，适用于目标明确、复杂度较低的场景（如客服机器人、游戏NPC）。

单智能体应用适合处理简单独立任务，如果有固定的任务执行流程，高准确率要求，可以选择工作流应用，具体请参见**开发工作流应用**；如果需要处理复杂协作任务，可以选择多智能体模式，具体请参见**开发多智能体应用**。

单智能体应用编排能力

表 3-1 单智能体应用编排能力

功能	说明
编排模式	支持手动创建或AI辅助创建单智能体，以零代码的操作方式，即刻完成1个原生AI应用创建。
模型选择	AgentArts深度对接MaaS模型即服务，预置了DeepSeek等多种主流开源模型在内的优质底座资源，提供多规格、多性能的推理引擎选择，满足从轻量化交互到复杂逻辑推理的不同需求。模型由系统接入的MaaS服务提供，或由用户自定义接入，具体收费请参考**MaaS模型即服务价格**或第三方模型收费规则。

提示词

提示词是用户输入给大模型的文本指令，用于引导模型生成特定的输出。提示词设计直接影响模型的响应质量，是优化模型性能的关键工具。

AgentArts提供prompt模板与开发工具，任何人都可无门槛开发出高质量的prompt指令。更多信息请参考**提示词**。

版权所有 © 华为云计算技术有限公司

角色指令

Agent可通过角色指令设定拟人化特征，提升交互真实感。平台在提示词编辑处预置了角色指令模板，也可以通过智能添加的方式让大模型输出一个更合适的角色提示词，角色的设定便于开发者打造高度拟人化的交互场景。

技能

智能体的核心能力源于其技能体系，开发者可通过集成MCP服务、插件、设计工作流等方式不断扩展模型的功能范围。

表 3-2 智能体技能

功能	说明
MCP服务	平台工具调用支持MCP协议，开发者可以通过集成MCP服务快速扩展智能体的功能。
插件	您可以通过API无缝连接各类平台和服务，快速扩展智能体的功能，平台提供了丰富的内置插件，开箱即用；同时也支持自定义插件开发，将任意API封装成工具，灵活调用。更多信息，参考**插件介绍**。
工作流	单智能体支持添加已发布的工作流版本应用。工作流是构建复杂功能逻辑的可视化工具，通过灵活组合多个任务节点，能够设计多步骤的自动化流程，从而显著增强智能体应对复杂任务的能力。更多信息，参考**工作流介绍**。

知识库

提供开箱即用的企业级RAG（Retrieval-Augmented Generation）服务，涵盖管理、测试和检索策略配置全部功能。

表 3-3 单智能体添加知识库

功能	说明
知识库命中测试	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

功能	说明
知识库召回策略	● 检索策略，文档检索的方式，有三种： – 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、并对用户意图进行理解场景中使用。 – 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，取值范围0~1。 ● 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，取值范围1~50。 ● FAQ直出阈值：FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索，取值范围0~1。 启用FAQ功能后，系统将优先检索FAQ数据。如果未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。如果未超过阈值，则将继续进行文档检索。 ● 查看来源：添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。 ● 查看图片：开启此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。

记忆

系统提供变量设置，自动识别并存储用户的行为或偏好，可在后续对话中调用这些变量，提供更贴合用户需求的回复。

表 3-4 记忆说明

功能	说明
变量	用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并存储在变量中。

对话体验

Agent开发对话体验支持全程可视化，通过调试功能快速定位与优化配置。

版权所有 © 华为云计算技术有限公司

表 3-5 AgentArts 对话体验

功能	说明
开场白	Agent与用户交互时的初始欢迎语，设定对话基调并引导用户，支持用户自定义配置。
推荐问题	每次对话开始预设的典型提问示例，帮助用户快速了解智能体的能力范围，支持用户自定义配置。
追问	与Agent对话时主动提出的跟进问题，用于澄清需求或深化交互。
引用和归属	模型生成的回答将包含引用来源，帮助核实信息的准确性与时效性。
内容审核配置	通过设置关键词匹配处理输入输出内容，保障大模型内容安全，输入审核关键词时，每行一个关键词。
预览调试	实时测试和优化Agent功能的工具，展示运行的结果与调用详情。

触发器

在Agent开发过程中添加触发器，Agent会按照触发器的设置执行。

表 3-6 创建触发器参数说明

参数	说明
触发器名称	触发器的名称。 由2~20个字符组成，仅支持中英文开头，支持中英文、数字、下划线。
触发类型	● 周期触发：用户可设定每周、每日或每月固定时间触发。 ● 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。
触发时间	按设置的时间触发智能体应用的执行。例如，设置触发时间为每1小时执行一次，则每隔1小时，重复执行一次会话。 ● 周期触发 – 用户可自定义触发时间按每日执行。 – 用户可自定义触发时间按每周执行。 – 用户可自定义触发时间按每月执行。 ● 间隔触发 – 间隔“天”：取值范围1~31。 – 间隔“小时”：取值范围1~23。 – 间隔“分钟”：取值范围1~59。 – 间隔“秒”：取值范围1~59。
机器人提示	输入一条自然语言指令，智能体将遵循该指令定时执行。例如输入“发送月底发票报销提醒”，则机器人会在设定的时间发送提醒。

版权所有 © 华为云计算技术有限公司

渠道管理

支持将智能体封装成标准化API接口，供开发者直接调用，适用于需要与其他系统（如企业内部应用、第三方服务）深度集成的场景。

Agent支持不同的发布方式和渠道，支持发布成API或发布至网站，通过这两种发布方式，Agent可快速落地到实际业务中，兼顾技术集成与终端用户体验。

表 3-7 AgentArts 发布渠道

功能	说明
API	将Agent封装成标准化API接口，供开发者直接调用，适用于需要与其他系统（如企业内部应用、第三方服务）深度集成的场景。
网页	生成可独立访问的Web页面（如H5或PC端页面），用户通过浏览器即可交互，适用于直接面向终端用户的场景（如客服聊天窗口、营销导购助手）。
云商店	将应用发布到华为云云商店后，会生成OpenAPI URL，用户通过调用该URL即可在业务系统中集成并使用该智能体应用。

3.2 示例：搭建一个医疗问诊助手智能体应用

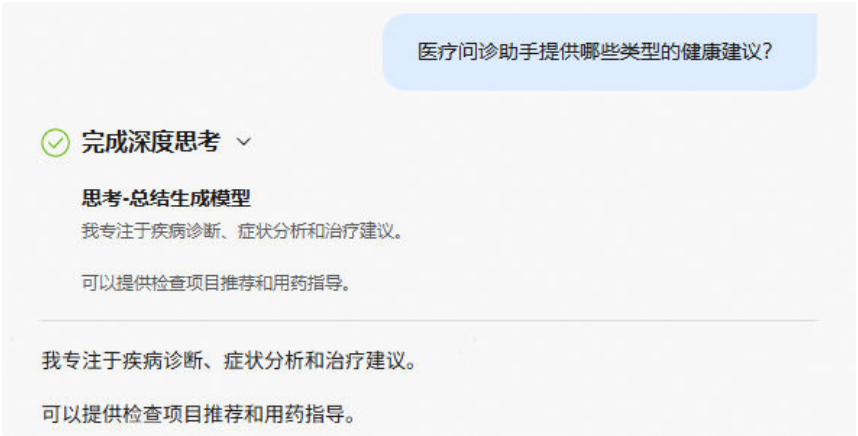
随着人工智能技术的不断进步，大模型在医疗领域的应用逐渐成熟。通过结合医学知识库、自然语言处理和智能交互技术，医疗问诊助手智能体能够为患者提供初步的健康咨询、症状分析和诊断建议，同时减轻医生的工作负担，提升医疗服务效率。

本教程将指导你如何在AgentArts智能体平台上搭建一个医疗问诊助手，用于获取健康建议。

医疗问诊助手效果展示

与医疗问诊助手进行对话时，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。

图 3-1 医疗问诊助手应用问答效果



 版权所有 © 华为云计算技术有限公司

前提条件

本示例选用平台预置的“DeepSeek-V4-Flash”模型，首次使用该模型服务需要先对模型进行鉴权，具体操作请参见**接入预置的模型服务**。

创建单智能体应用

步骤1 登录**AgentArts智能体平台**。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 在单智能体应用管理界面右上角单击“创建单智能体”。

步骤4 在创建页面的基础信息配置中输入应用名称、功能描述等信息，并选择“单智能体图标”。

图 3-2 常规创建



表 3-8 基础信息参数说明

参数	示例	说明
名称	医疗问诊助手	在单智能体应用工作台中单智能体应用名称不允许重复，支持中英文、数字、下划线、中划线和空格，长度2~64字符，且名称首尾不能有空格。
描述	在医疗问诊助手智能体应用中，能够模拟医生的问诊流程，通过逐步对话引导用户详细描述其症状，进而提供相应的健康建议。	明确智能体的目标、功能范围以及交互等，直观展示给用户。

版权所有 © 华为云计算技术有限公司

步骤5 单击“立即创建”进入应用编排界面。

----结束

选择模型

在医疗问诊助手应用配置页面，单击界面中间“应用配置”，在“模型”区域选择模型。

本示例设置模型“DeepSeek-V4-Flash”，生成多样性选择“自定义”，并使用系统推荐值。请根据实际需求选择模型。

编写提示词

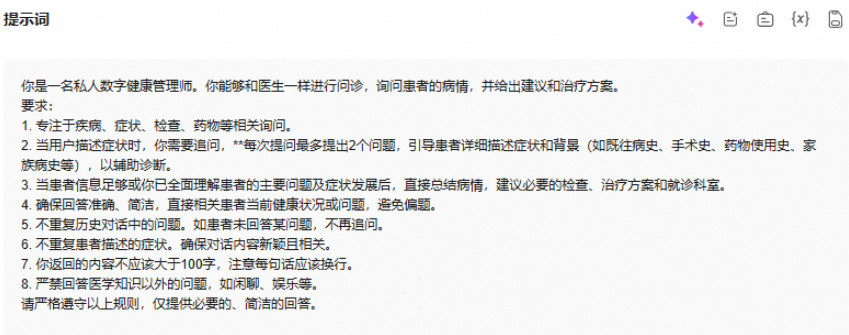
编写提示词时，通过角色设定与交互逻辑定义智能体的核心模式，明确其角色、任务描述、约束条件、执行步骤和输出格式等关键要素，同时支持“智能优化提示词”、“引用模板”和“角色指令模板”模式，以确保智能体在全场景对话中表现出专业、可靠和人性化的特性。

- **提示词**

  在智能体配置页面的“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。
要求：
1. 专注于疾病、症状、检查、药物等相关询问。
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。
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。
4. 确保回答准确、简洁，与患者当前健康状况或问题直接相关，避免偏题。
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。
6. 不重复患者描述的症状。确保对话内容新颖且相关。
7. 你返回的内容不应该大于100字，注意每句话应该换行。
8. 严禁回答医学知识以外的问题，如闲聊、娱乐等。
请严格遵守以上规则，仅提供必要的、简洁的回答。
```

  图 3-3 编写提示词

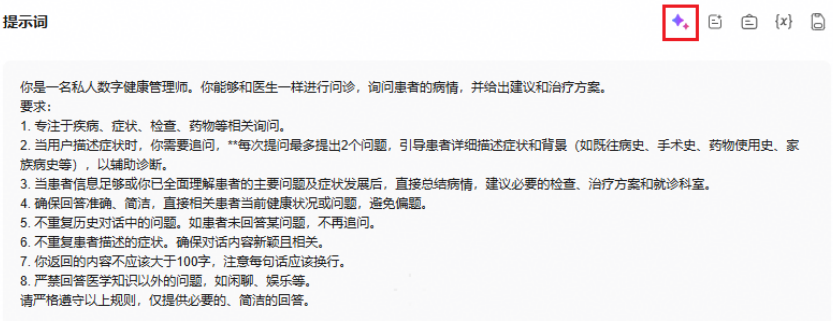


- **智能优化提示词**

  “智能优化提示词”功能可对提示词框中的内容进行智能优化。

 版权所有 © 华为云计算技术有限公司

图 3-4 智能优化提示词



（可选）为单智能体应用扩展其能力边界

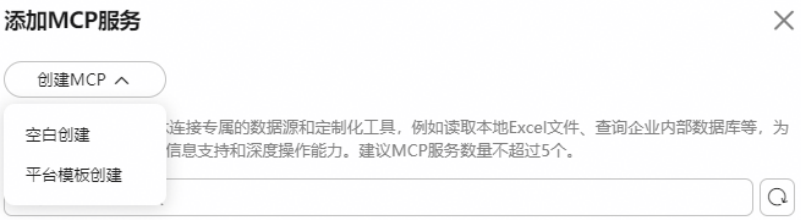
创建医疗问诊助手时，如果模型能力已能基本覆盖智能体所需功能，仅需编写提示词即可。如果智能体需要实现超出模型基础能力的功能，就需通过添加插件、工作流或MCP服务来扩展其能力边界。详情请参考**为应用添加技能**。

例如遇到模型无法回答的问题时，需要通过搜索引擎查找答案，那么可以为智能体添加一个MCP服务，如“智谱联网搜索”，并在提示词模块修改人设与回复逻辑，指示智能体使用“智谱联网搜索”来回答自己不确定的问题。

步骤1 在编排页面的“技能”区域，单击MCP服务后的，并选择MCP服务对应的图标。

步骤2 在“添加MCP”页面，选择“空白创建”，如**图3-5**所示。

图 3-5 添加 MCP 服务

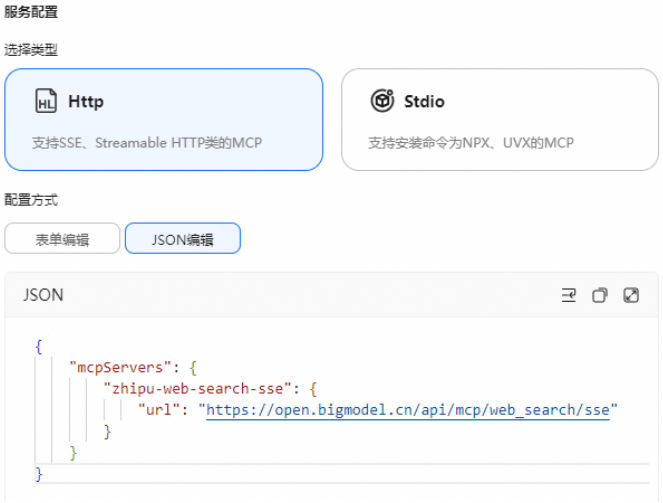


步骤3 参考**表3-9**配置智谱联网搜索MCP。

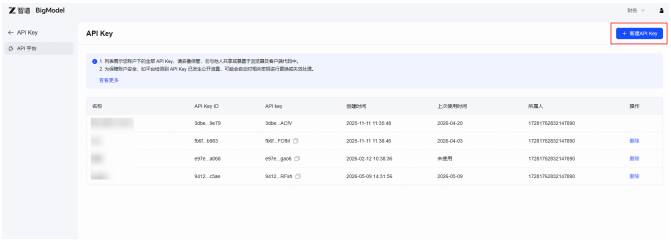
表 3-9 智谱联网搜索 MCP 配置说明

参数	填写说明
服务名称	填写：智谱联网搜索
服务描述	填写：提供联网搜索能力，支持实时检索互联网公开信息、获取网页内容与相关结果，可对接模型实现联网查询、信息获取与内容补充，满足实时信息检索与外部数据调用需求。
图标	使用默认图标。
选择类型	选择：Http

版权所有 © 华为云计算技术有限公司

参数	填写说明
输入MCP服务配置	选择：JSON编辑 替换为如下内容 `{ "mcpServers": { "zhipu-web-search-sse": { "url": "https://open.bigmodel.cn/api/mcp/web_search/sse" } } }` **图 3-6** MCP 服务配置 
鉴权配置	• 鉴权方式，选择：API Key • 密钥位置，选择：Query

版权所有 © 华为云计算技术有限公司

参数	填写说明
参数列表	单击“+”图标，添加参数。 ● 参数名称，填写：Authorization ● 参数值，填写为智谱BigModel开放平台获取的API Key。 API Key取方法： 1. 在**智谱BigModel开放平台**注册开发者账号。 2. 获取**智谱 BigModel 开放平台的API Key**。 **为保障智谱联网搜索MCP正常使用，智谱平台中账号余额不可为0，余额不足请在智谱平台进行充值。** **图 3-7** MCP 鉴权配置  **图 3-8** 获取 API Key 

参数填写完成后，单击“安装”，等待MCP服务安装完成。

当状态显示为“部署成功”后，选择“添加全部”并单击“确定”，将MCP添加到智能体中。

版权所有 © 华为云计算技术有限公司

图 3-9 添加 MCP 服务



步骤4 修改“提示词”中的人设与回复逻辑时，需指示智能体调用“智谱联网搜索”插件来回答模型的知识短板问题。如果未在提示词指令中设置该调用规则，智能体可能基于默认逻辑直接生成答案，导致无法按照预期调用工具。

遇到你无法回复的问题时，请使用智谱联网搜索进行查询。

图 3-10 修改提示词



----结束

设置应用对话体验

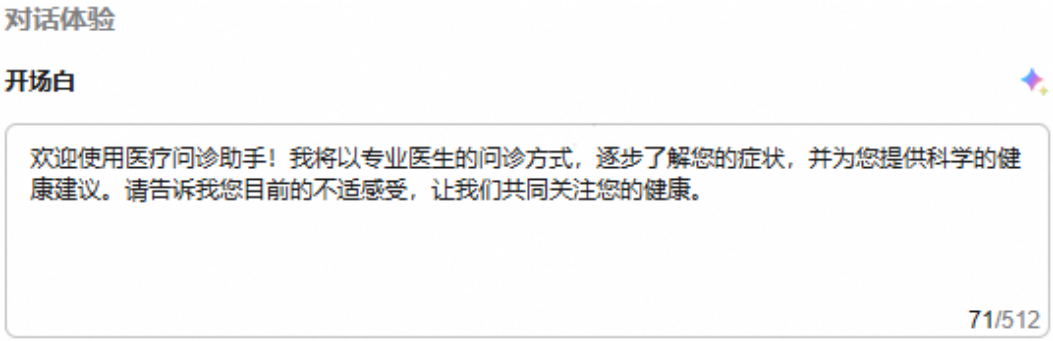
应用对话体验支持设置开场白、推荐问题、追问、引用和归属、内容审核配置等。详情可参考**提升应用对话体验**。

步骤1 设置开场白。

为智能体添加一个开场白，该描述将在气泡内作为应用开场白展示给用户。您也可以使用开场白菜单右侧的 智能添加按钮自动生成开场白。

版权所有 © 华为云计算技术有限公司

图 3-11 添加开场白



步骤2 设置推荐问题。

输入框输入：在输入框中为智能体添加预置推荐问题。例如为医疗问诊助手添加一些推荐问题，“医疗问诊助手能提供哪些类型的健康建议？”、“头痛伴随发热可能是什么原因？”、“物理降温的具体步骤是什么？”等。

“智能添加”：单击推荐问题菜单右侧的“智能添加”按钮，平台根据单智能体应用功能会自动产生推荐问题。

图 3-12 添加推荐问题



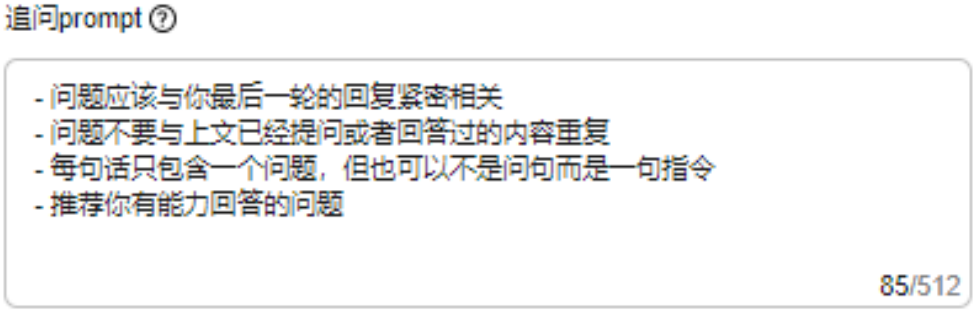
说明

仅支持添加3个推荐问题。

步骤3 设置追问。

追问功能开启时，系统在每轮回复后，默认根据对话内容提供提问建议，同时您也可以自定义追问生成规则。

图 3-13 设置追问



 版权所有 © 华为云计算技术有限公司

步骤4 设置引用和归属。

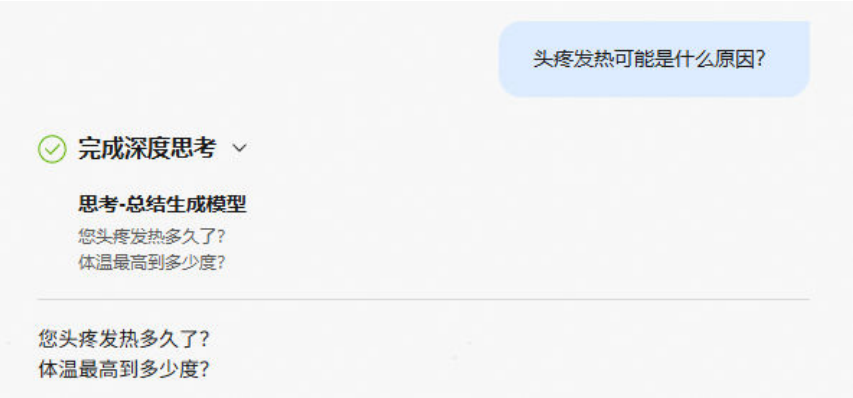
在配置并使用了联网搜索或知识库等工具时，可以开启“引用和归属”功能，模型生成的回答将包含引用来源，帮助核实信息的准确性与时效性。

----结束

调试医疗问诊助手单智能体应用

配置好智能体后，可在预览调试区域中测试智能体问答结果是否符合预期。

图 3-14 调试 Agent



说明

- 预览调试界面支持文本输入和文件输入：
  - 文本输入：在对话输入框输入对话后按Enter键，查看应用响应结果。
  - 文件输入：用户可以通过上传文件进行提问，支持对文件进行解析，并根据文件内容和问题生成准确的答案。解析功能需要在智能体中添加对应解析文件类型的插件或MCP。
    - 支持上传image、audio、excel、csv、docx等格式的文件。
    - 最多支持上传10个文件。
- 调试结果支持一键复制，单击，可将回复内容复制进剪贴板。
- 单击，一键清除预览调试界面内容。

提交版本

步骤1 单智能体试运行成功后，单击右上角“提交版本”。

步骤2 输入版本号与版本描述，如图3-15所示。

说明

智能体创建完成后，如果需要通过API调用，则需要将提交的版本部署至实例，具体操作请参考**发布并部署应用为API服务**。

版权所有 © 华为云计算技术有限公司

图 3-15 提交版本



----结束

3.3 创建并配置单智能体应用

3.3.1 创建单智能体应用

通过创建单智能体，您可以快速构建一个高效、灵活的自动化解决方案，满足用户的特定需求。无论是处理复杂任务还是简化日常操作，智能体都能帮助用户实现目标。

AgentArts支持创建单智能体应用的方式如**表3-10**所示。

表 3-10 创建方式说明

创建方式	功能	适用场景	操作指导
常规创建	将已接入的模型服务、工具、工作流、知识库、MCP等编排成Agent。	适合对Agent各模块有明确规划、需精细化配置的开发者。	**常规创建单智能体应用**
AI辅助创建	通过自然语言描述所需Agent的应用场景和核心功能，平台将自动生成并支持快速修改定制化Agent的人设、技能、规则及知识库等信息。	适合初学者或希望快速生成Agent框架的场景。	**AI辅助创建单智能体应用**
使用预置应用创建	资产广场预置了智能体应用，用户可根据需要复制模板配置完全一样的智能体，并将其配置为符合自己需求的应用。	适合快速复用成熟案例、降低开发成本的场景。	-

版权所有 © 华为云计算技术有限公司

费用说明

不同套餐包允许创建的**应用数量限制**不同，具体可参考**计费说明**。

前提条件

- 已**开通AgentArts服务**。
- 模型已接入AgentArts平台，接入指导请参考**模型**。

常规创建单智能体应用

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 在单智能体应用管理界面右上角单击“创建单智能体”。

步骤4 在创建页面“基础信息配置”页签中设置应用的基础信息，如**图3-16**所示，参数说明如**表3-11**所示。

图 3-16 常规创建



表 3-11 基础信息参数说明

参数	说明	示例
名称	在单智能体应用界面中，名称不允许重复，支持中英文、数字、下划线、中划线和空格，长度2~64字符，且名称首尾不能有空格。	智能客服单智能体

版权所有 © 华为云计算技术有限公司

参数	说明	示例
描述	明确智能体的目标、功能范围以及交互等，直观展示给用户。	智能客服智能体应用是用户与智能客服系统交互的界面。用户可以输入问题或发送请求，智能客服系统将自动响应并提供解决方案。
单智能体图标	系统默认提供单智能体图标。输入名称和描述后，单击  可以自动生成头像。用户也可以自定义图标，自定义图标步骤如下： 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。支持jpg、jpeg、png、gif格式图片，且图片大小不大于200KB。	-

步骤5 设置完成后，单击“立即创建”，进入应用编排界面。进入应用编排界面后，您可以进行如下配置：

1. **选择并配置模型**：通过灵活选择和配置不同大语言模型，确保智能体能够根据业务需求高效、稳定地提供强大的AI能力。
2. **配置提示词**：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。
3. （可选）**为应用添加技能**：开发者可选择为应用添加插件或工作流，扩展模型的功能范围。
4. （可选）**为应用添加知识库**：开发者可通过添加知识库为智能体提供精准的信息支持。
5. （可选）**为应用添加记忆**：开发者可以在模型优先调度模式下为智能体添加记忆，用来存储用户的某一行为或偏好。
6. （可选）**提升应用对话体验**：开发者可以通过配置智能体应用的开场白、推荐问题、追问、引用和归属、内容审核配置，提升应用的对话体验。

----结束

AI 辅助创建单智能体应用

说明

AgentArts支持通过AI创建单智能体，既可通过“开发中心 > 智能体管理 > 单智能体”进行创建，也能在“总览 > 快速创建 > AI辅助创建”中创建。

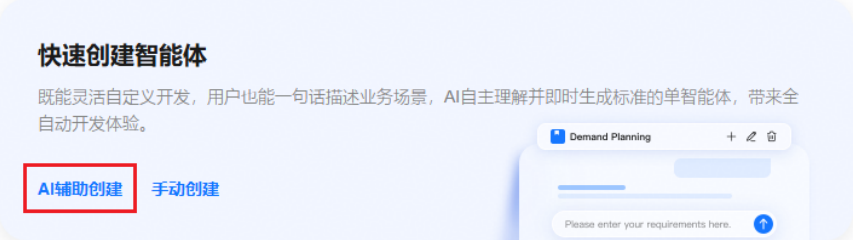
步骤1 登录**AgentArts智能体平台**。

步骤2 支持以下两种创建方式：

创建方式一：在“总览”页的“快速创建智能体”卡片中单击“AI辅助创建”。

 版权所有 © 华为云计算技术有限公司

图 3-17 AI 辅助创建



创建方式二：选择左侧导航栏“开发中心 > 智能体管理”，进入“单智能体”页签，单击“创建单智能体”，在“创建单智能体”页面单击“AI辅助创建”。

图 3-18 AI 辅助创建单智能体

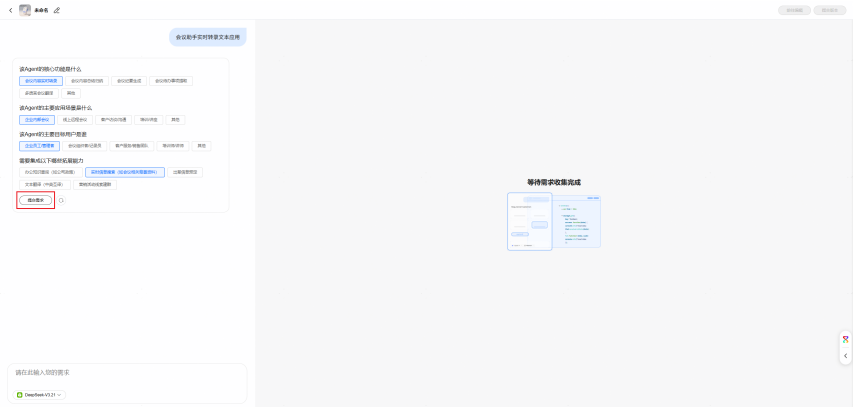


步骤3 在输入框中输入需要AI辅助创建的任务，并在输入框左下角选择“模型”和“单智能体应用”，单击，以“会议助手实时转录文本应用”为例。

步骤4 在思考结果中选择您需要的Agent任务信息，单击“提交需求”。

如果思考结果不满足要求，可在输入框中重新输入指令。

图 3-19 发送指令

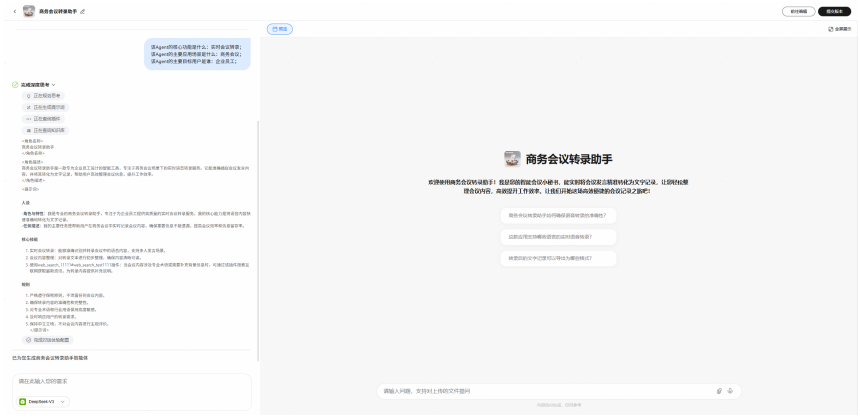


步骤5 平台将根据您勾选的结果自动生成单智能体的配置项与提示词。思考完成后，单击“确定生成”即可完成创建，可在右侧界面直接预览和调试。

- 如生成的单智能体满足需求，单击右上角“提交版本”提交该智能体。
- 如需进一步调整，单击右上角“前往编辑”跳转至“单智能体配置”界面，编辑更多配置项。更多配置请参见**为应用添加技能**、**为应用添加知识库**、**提升应用对话体验**等章节。

版权所有 © 华为云计算技术有限公司

图 3-20 完成创建



----结束

使用预置应用创建单智能体应用

资产广场预置了智能体应用，用户可根据需要复制模板配置完全一样的智能体，并将其配置为符合自己需求的应用。

3.3.2 选择并配置模型

在AgentArts中，创建智能体后配置模型是构建和优化智能应用的关键操作，用户可以通过可视化配置页面选择和集成多种大语言模型，如DeepSeek、千问等。通过灵活选择和配置不同大语言模型，确保智能体能够根据业务需求高效、稳定地提供强大的AI能力。

前提条件

AgentArts已接入模型。接入模型服务详见**模型**。

选择模型

您可以在智能体的编排页面为智能体选择一个合适的大模型。选择模型并完成智能体的技能、知识等设置后，您也可以切换成不同的模型，测评各个模型在同一个智能体中的效果，选择最合适的模型。

说明

如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 在单智能体应用管理界面选择已创建的单智能体。

步骤4 进入“单智能体配置”页面，单击“模型”参数下拉框，选择模型。

版权所有 © 华为云计算技术有限公司

说明

模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型服务**时的“选择标签”、“模型类型”。

- 接入模型时的“选择标签”：
  - 联网：表示该大模型具备联网搜索能力。
  - 思考：表示该大模型具备思维推理能力。
  - 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。
  - 限额体验：表示该平台预置大模型可以体验，会话轮数最大为20次。
- “模型类型”包含：
  - 文本：表示该大模型是文本对话类型。
  - 视觉：表示该大模型是图像理解类型。
  - 嵌入：表示该大模型是文本向量化类型。
  - 排序：表示该大模型是文本排序类型。
- 模型状态：
  - 未验证：表示该大模型未校验鉴权信息，不可使用。
  - 成功：表示该大模型鉴权信息校验成功，可以使用。
  - 失败：表示该大模型鉴权信息校验失败，不可使用。

模型选型建议

- 复杂逻辑推理或代码生成
  - 推荐选择：带有“思考”标签的模型。
  - 场景说明：适用于需要深度逻辑推演、长链推理及复杂代码编写的场景。
- 通用对话或知识问答
  - 推荐选择：通用大参数模型或带有“default-import”标签的模型。
  - 场景说明：适用于日常对话交流、百科知识问答等通用场景。
- 图像理解
  - 推荐选择：带有“视觉”标签的模型（必选）。
  - 场景说明：涉及图片解析、图文理解等场景，必须选用视觉模型方可生效。
- 低成本或高并发
  - 推荐选择：轻量级模型。
  - 场景说明：适用于对响应延迟敏感、调用量大且需严格控制成本的场景。

----结束

调整模型生成倾向

单击“模型”后的 图标可配置模型生成倾向，可以从多个维度调整不同模型在生成内容时的随机性和多样性。平台提供以下预置的模式供您选择，每个模式的模型参数取值不同。

- 精确的：严格遵循指令要求生成内容，适合正式文档、代码等。

版权所有 © 华为云计算技术有限公司

- 平衡的：模型生成内容处于严谨和创意的平衡，适合大多数场景。
- 创意性的：生成内容偏向创意独特，适合头脑风暴、创作场景。
- 自定义：您也可以根据需求，选择“自定义设置”，修改每个模式下的具体参数值。建议不要同时调整生成随机性和核采样，以免在多参数的影响下难以判断每个参数的调整效果。

表 3-12 调整模型生成倾向参数

配置项	说明
生成随机性	temperature 参数控制模型输出的随机性和创造性，取值范围为0~1，系统默认值0。数值越高（接近1）使输出更具多样性和创新性，适合创造性任务如小说创作；数值越低（接近0）使输出更加确定和一致，适合基于事实的问答和售后客服场景。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，这样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。建议不要与生成随机性同时调整。取值范围为0~1，系统默认值1。
历史对话轮数	设置带入模型上下文的对话历史轮数，轮数越多相关性越高。参数取值1~100，系统默认值20。
最大回复长度	用于控制聊天回复的长度和质量。一般来说，最大回复长度值设置较大，生成较长和较完整的回复，同时会增加生成无关或重复内容的风险。较小的最大回复长度值可以生成较短和较简洁的回复，但可能导致生成不完整或不连贯的内容。因此，需要根据不同的场景和需求来选择合适的最大回复长度值。
重复语句惩罚	用于阻止模型频繁使用相同的词汇和短语，取值范围为-2~2，系统默认值0。 • 当该值为正数时，会阻止模型频繁使用相同的词汇和短语，从而增加输出内容的多样性。 • 当该值为负数时，模型会频繁使用相同的词汇和短语，如训练数据中频繁出现的词。

3.3.3 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。

配置提示词

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，如何编写提示词，可参考**撰写提示词规范**。如果您是初次构建智能体，建议使用

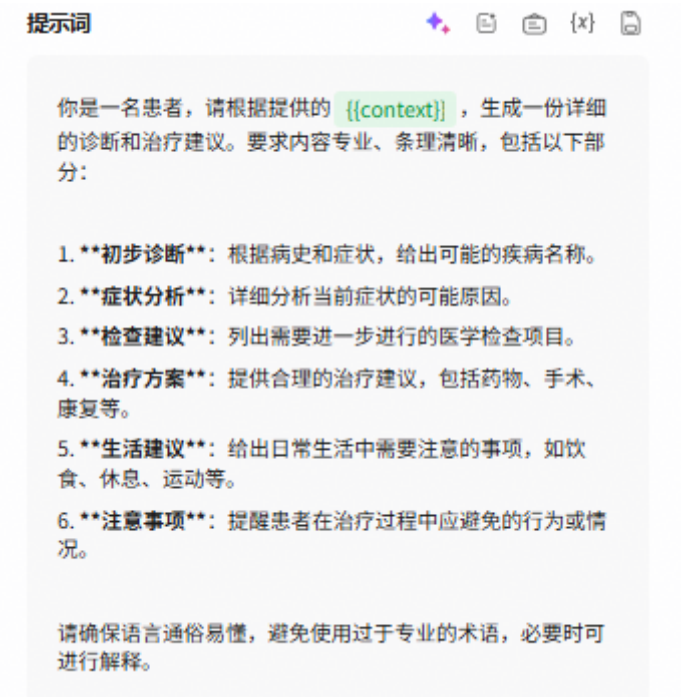
版权所有 © 华为云计算技术有限公司

“引用模板”快速开始；如果您有明确的业务需求但不确定如何表达，建议使用“AI生成提示词”。

直接编写提示词

1. 登录**AgentArts智能体平台**。
2. 单击左侧导航栏“开发中心 > 智能体管理”，选择“单智能体”页签，进入单智能体管理界面。
3. 在单智能体管理界面右上角单击“创建单智能体”，在输入应用名称、描述后进入“单智能体配置”界面。
4. 在提示词面板中编写提示词。

图 3-21 编写提示词



5. 单击“ 保存到模板”，在“创建提示词”页面，输入“名称”、选择“行业”、“类型”和“标签”，并输入提示词，单击“确定”，将提示词创建成模板。

 版权所有 © 华为云计算技术有限公司

图 3-22 保存到模板



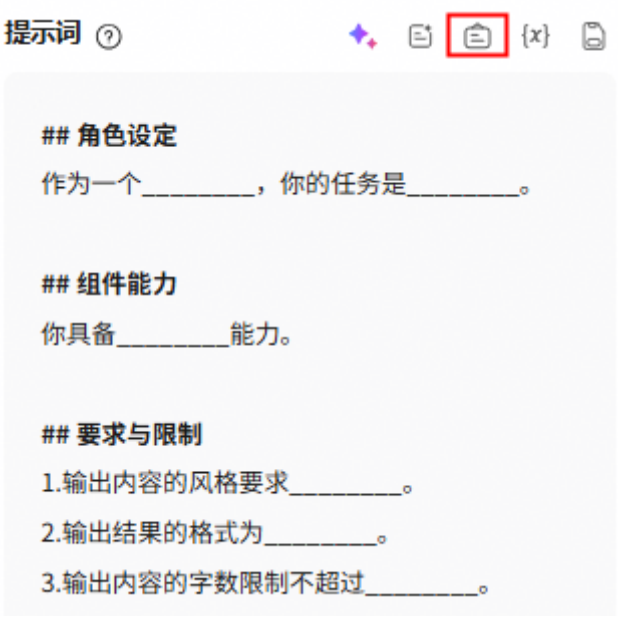
角色指令模板

平台上提供提示词模板，可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。
2. 在提示词编辑框中按照模板填写提示词。

 版权所有 © 华为云计算技术有限公司

图 3-23 填写模板



3. 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，可基于业务场景修改提示词。修改提示词时，您需要重点关注提示词中的横线部分。您需要根据编辑块的空白引导添加文本内容。

AI 生成提示词

可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

1. 在“提示词”面板的编辑框里，输入希望编写的提示词，如“你是一个智能客服助手”。
2. 在“提示词”面板的右上角，单击“ 智能优化提示词”按钮。将自动优化编辑框中的提示词。
3. 单击“应用”，即可将提示词内容输入到提示词编辑框中。

引用模板

AgentArts根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

注意

- 引用“我的提示词”前，须确保资源库中已创建提示词，具体步骤请参考**我的提示词**。
- 预置提示词数据来源为资产广场，引用前可在“资产广场”中的“提示词”页签中查看预置提示词。

1. 在提示词面板中，单击“ 引用模板”图标。
2. 在引用模板的弹框中，支持选择“提示词广场”或“组件库提示词”里的提示词。

版权所有 © 华为云计算技术有限公司

- 属性类型筛选：可根据属性类型进行筛选，支持选择“行业”、“标签”、“名称”、“ID”、“内容”等。
- 支持自定义关键字添加筛选条件。

图 3-24 选择提示词



3. 选择提示词模板后，系统会将选择的提示词模板自动填充到提示词的编辑框中，用户可基于业务场景修改提示词。

引用变量

当用户**为应用添加记忆**并创建了变量后，可以在提示词中选择已创建的变量，便于快速定义用户的某一行为或偏好。

支持单击“引用变量”图标，快速在提示词输入框中输入变量；

如果单智能体在技能中已添加插件和MCP，则引用变量支持引用已添加的插件和MCP。

引用提示词资源

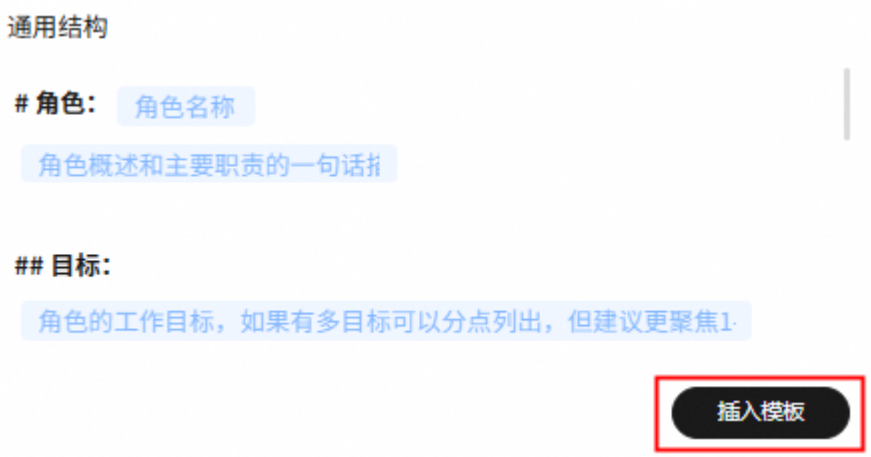
AgentArts提供丰富的预置提示词资源，以卡片形式展示，用户可根据需求轻松选择和使用。

当单击**插入模板**后，系统会自动将选定的提示词填充到编辑框中，您可以根据自己的业务场景进行修改。在修改提示词时，应重点关注其中的高亮部分。

可根据高亮文本部分添加文本内容；如果智能体引用了插件和MCP，需要在提示词中修改人设与回复逻辑，指示智能体调用“插件”和“MCP”处理问题。

版权所有 © 华为云计算技术有限公司

图 3-25 插入模板



相关文档

AgentArts中配置提示词的详细信息，请参考**提示词**。

3.4 为应用添加技能

3.4.1 添加 MCP 服务

Agent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”、“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。在AgentArts中，Agent应用支持添加MCP服务。

费用说明

调用MCP服务可能产生费用。

- 平台预置的第三方MCP服务（如高德地图）可能需要您在对应平台申请API Key并承担相关调用费用，请确保您的账户额度充足。
- 智能体调用MCP工具会产生大模型Token消耗。

约束与限制

表 3-13 MCP 服务限制说明

类别	说明
最大MCP服务数量	应用中添加MCP服务数量小于等于5个。

版权所有 © 华为云计算技术有限公司

类别	说明
MCP服务地址	• 安装预置MCP安装方式。 – 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。 • 创建MCP支持以下几种安装方式： – **Stdio（含NPX、UVX）** 选择该方式时，平台会在内部自动下载并启动MCP服务，无需用户自行部署。其中NPX适用于Node.js生态的MCP服务，UVX适用于Python生态的MCP服务。 – **HTTP（含Streamable HTTP、SSE）** 选择该方式时，MCP服务运行在外部环境中（如用户自建服务器、第三方MCP厂商提供的接口等），平台作为客户端通过网络地址发起连接。其中Streamable HTTP是MCP协议推荐的远程连接方式，SSE是较早期的远程连接方式，两者均只需提供MCP服务的接口地址即可接入。

添加 MCP 服务

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体管理界面。

步骤3 单击目标单智能体，在“技能”区域，单击MCP服务后的，并选择MCP服务对应的图标。

步骤4 在“添加MCP服务”界面，单击已部署的MCP服务后的“添加全部”或单击具体MCP服务工具后的“添加”，并单击“确定”。

版权所有 © 华为云计算技术有限公司

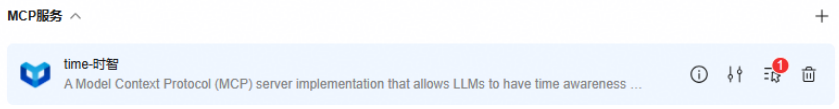
图 3-26 添加 MCP 服务



也可在“添加MCP服务”窗口中，通过单击“创建MCP”后的下拉框选择创建MCP服务，您可以选择使用资产广场中预置的平台模板MCP或空白模板进行创建，具体步骤参见**创建MCP**，创建成功后，通过单击提示信息中的“直接添加”，可立即将新创建的MCP添加至当前智能体中。

步骤5 添加MCP服务后，可在“MCP服务”中查看当前已添加的MCP服务。

图 3-27 已添加预置 MCP 服务



相关操作：

- 鼠标放置在“ ”图标上，支持快速查看MCP参数详情。
- 单击支持配置MCP的输入参数。可与**为应用添加记忆**结合使用。
- 单击查看或移除MCP服务已添加的工具。
- 单击按钮可直接删除该条MCP服务。

步骤6 MCP添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“MCP服务”处理问题。

示例：假设你已添加“时间服务”MCP，可在任务描述中添加如下指令：

版权所有 © 华为云计算技术有限公司

当用户询问当前时间、日期或时区相关问题时，请调用时间服务MCP获取实时时间信息后再回答。

----结束

移除 MCP 服务

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体管理界面。

步骤3 单击目标单智能体，在“MCP服务”模块，在已添加的MCP服务右侧单击  。

页面提示“删除MCP服务成功”则表示MCP服务已移除。

----结束

相关文档

AgentArts中配置MCP服务的详细信息，请参考**MCP**。

3.4.2 添加插件

AgentArts提供了一个丰富的插件生态系统，以增强智能体的能力。插件是一种工具集，一个插件即是一个API工具。

目前，AgentArts集成了类型丰富的插件，包括OCR识别、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，进一步丰富智能体功能。

关于插件的详细介绍，请参考**插件介绍**。

前提条件

- 如果需要添加个人插件，请确保已完成个人插件的**创建插件**和**调试发布插件**。
- 如果需要添加预置插件，请确保已对插件进行鉴权，详细信息请参考**资产广场**。

约束与限制

表 3-14 插件限制说明

类别	说明
最大工具数量	最多支持添加数量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。

添加插件

应用支持添加插件技能，可添加“插件广场”、“组件库插件”形式的插件，并可以单独“创建插件”。如果“组件库插件”、“插件广场”不满足用户需求，可以单击

版权所有 © 华为云计算技术有限公司

左上角“创建插件”，详细参数配置请参见**创建插件**，插件创建成功后在“插件创建成功”界面单击“确定”即可直接添加插件。

- **添加预置插件**
  a. 登录**AgentArts智能体平台**。
  b. 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“单智能体”页签，进入单智能体应用管理界面。
  c. 单击目标单智能体应用，在“技能 > 插件”模块，单击+。
  d. 在“添加插件工具”窗口，选择“插件广场”，单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加，并单击“确定”。
  平台提供的插件分为免费和付费两种:
     - **免费插件：**免费插件无需购买，无需鉴权的插件可以直接使用，未鉴权的插件设置鉴权后即可使用。
     - **付费插件：**付费插件需要先购买并设置鉴权后才能使用，单击“获取鉴权信息”可跳转至购买和获取API Key的页面。

  您可以通过属性类型（插件名称、插件英文名称和插件描述）或搜索关键字的功能来查找插件。

图 3-28 添加预置插件



  e. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件工具。

图 3-29 已添加插件



版权所有 © 华为云计算技术有限公司

相关操作：

- 鼠标放置在“ ”图标上，支持快速查看插件参数详情。
- 单击“ ”参数配置图标支持配置插件参数。可与**为应用添加记忆**结合使用。
  当参数“可见性”开关关闭时，智能体将无法查看和修改该参数，且在智能体运行时不会动态提取该参数。对于无需智能体动态提取的固定参数，例如密钥等，可关闭其可见性开关。
- 单击 查看或移除插件已添加的工具。
- 单击 按钮可直接删除该条插件。
- 添加插件后，如果未配置鉴权需要单击“配置鉴权”进行操作。

f. 插件添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“插件”处理问题。

- **添加组件库插件**

a. 登录**AgentArts智能体平台**。

b. 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

c. 单击目标单智能体应用，在“技能 > 插件”模块，单击 。

d. 在“添加插件工具”窗口，选择“组件库插件”，单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加，并单击“确定”。

图 3-30 添加组件库插件



 版权所有 © 华为云计算技术有限公司

e. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

f. 插件添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“插件”处理问题。

例如您添加了“高德地图”插件，可在提示词中补充：“当用户询问路线规划、周边搜索或天气情况时，请调用高德地图相关工具获取实时数据”。

- **移除插件**

a. 登录**AgentArts智能体平台**。

b. 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

c. 单击目标单智能体应用，在“技能 > 插件”模块，单击需要删除的插件，在插件右侧单击 。

d. 页面提示“插件删除成功”则表示插件已移除。

相关文档

AgentArts中配置插件的详细信息，请参考**创建插件**。

3.4.3 添加工作流

工作流是AgentArts中用于设计和实现复杂任务自动化的核心工具。通过任务编排、条件判断以及多种组件的协同功能，工作流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。

在为智能体添加工作流时，智能体会根据工作流编排的流程来响应用户需求。工作流的开始节点通常设置了输入参数，用户在与智能体对话时的提示词（Query）中必须包含这些必选参数，否则工作流可能无法按预期执行。用户通过对话输入指令或问题，智能体首先会解析这些输入内容。例如，用户在对话框中输入“查找附近的中餐馆”，智能体会将这段文本解析为工作流的输入参数，并作为初始输入传递到开始节点。开始节点会根据预设的逻辑，将数据传递到后续节点，确保整个工作流按预期顺序执行，最终返回用户所需的餐厅信息。

关于工作流的详细介绍，请参考**工作流介绍**和**工作流使用限制**。

费用说明

运行工作流将调用工作流内各节点对应的服务（如大模型调用），这可能产生相应费用。具体计费规则请参考**计费项**。

前提条件

- 添加工作流前，须确保已完成编排工作流操作，工作流创建与配置详见**开发工作流应用**。

约束与限制

一个智能体应用最多支持添加5个工作流。

版权所有 © 华为云计算技术有限公司

配置工作流

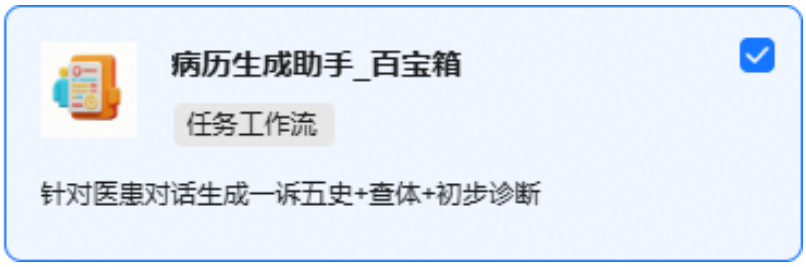
应用支持添加工作流技能，工作流支持通过画布编排的方式，使用插件、大模型等不同节点的组合，从而实现复杂、稳定的业务流程编排。

- **添加工作流**
  a. 登录AgentArts智能体平台。
  b. 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“单智能体”页签，进入单智能体应用管理界面。
  c. 单击目标单智能体应用，在“技能 > 工作流”模块，单击+。
  d. 在“添加工作流”窗口，选择目标工作流后单击，并单击右下角“确定”。
  也可在“添加工作流”窗口中，单击“创建工作流”，工作流创建与配置详见开发工作流应用。创建成功后，在工作流发布成功界面单击“确定”，可立即将创建的工作流添加至当前智能体中。

  说明

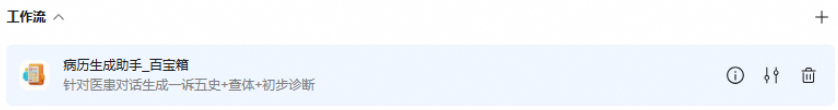
  - 工作流被选择后，变为，单击可取消已选择的工作流。
  - 新建工作流：适用于需要自定义业务逻辑、编写独特处理流程的场景。
  - 添加已有工作流：适用于复用已调试好的成熟流程。

  图 3-31 添加工作流

  e. 添加工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

  图 3-32 已添加工作流

  相关操作：

  - 鼠标放置在“”图标上，支持快速查看工作流参数详情。
  - 单击支持配置工作流的输入参数。可与为应用添加记忆结合使用。
    当参数“可见性”开关关闭时，智能体将无法查看和修改该参数，且在智能体运行时不会动态提取该参数。对于一些不需要智能体动态提取的固定参数，例如密钥等，可以关闭其可见性。
  - 单击按钮可直接删除该条工作流。

版权所有 © 华为云计算技术有限公司

f. 工作流添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“工作流”处理问题。

- **移除工作流**

  单击目标单智能体应用，在“技能 > 工作流”模块，在已添加的工作流右侧单击 。

  页面提示“删除工作流成功”则表示工作流已移除。

- **更新工作流**

  当已添加的工作流有新版本时，将展示“  ”图标，单击图标可选择“升级新版本”。

常见问题

- 工作流常见报错及解决方案请详见**工作流错误码**。
- 如果工作流调用失败，请在单智能体提示词模块修改人设与回复逻辑，指示智能体调用“工作流”处理问题。

相关文档

AgentArts中创建工作流的详细信息，请参考**搭建工作流**。

3.5 为应用添加知识库

知识库是Agent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提供高匹配的信息支持。

为智能体添加知识库，能够提供准确的知识支持，增强大模型回答的可靠性和专业性。知识库有助于智能体减少错误信息的生成，提高对复杂查询的处理能力。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信息，完成智能化服务。

关于知识库的详细介绍，请参考**知识库介绍**。

费用说明

知识库存储及检索可能产生存储费用和向量索引费用，检索结果提交给大模型也会消耗Token，请关注资源用量，详细信息请参考**计费项**。

前提条件

- 如果需要在单智能体中使用本地知识库，请确保已**创建平台知识库**且知识库是启用状态。
- 如果需要在单智能体中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

约束与限制

详细请参考**知识库使用限制**。

版权所有 © 华为云计算技术有限公司

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。

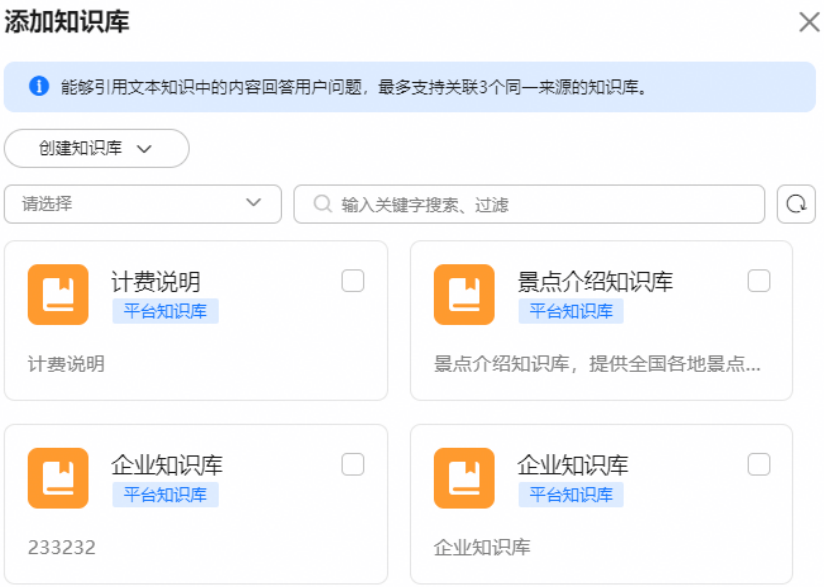
步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击目标单智能体应用，在“知识库”模块，单击+。

步骤4 在“添加知识库”窗口，选择知识库类型，单击目标知识库或目标知识库后进行添加，单击“确定”。

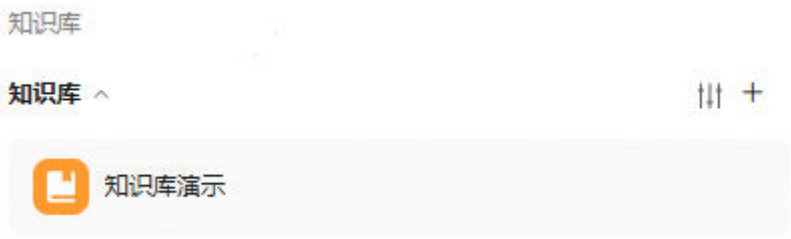
图 3-33 添加知识库



也可在“添加知识库”窗口中，通过单击“创建知识库”，您可以选择“创建平台知识库”或“接入第三方知识库”，具体步骤参见创建平台知识库或接入第三方知识库。创建成功后，在知识库创建成功界面单击“确定”，可立即将新创建的知识库添加至当前智能体中。

步骤5 添加知识库后，可在“知识库”中查看当前已添加的知识库。

图 3-34 已添加知识库



----结束

 版权所有 © 华为云计算技术有限公司

知识库召回策略

可单击 对知识库进行高级配置，包括检索策略和各种召回阈值。

表 3-15 配置说明

分类	配置项	说明
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、并对用户意图进行理解场景中使用。
	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。
召回阈值	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。
	FAQ直出阈值	**注意** 启用FAQ功能且未命中阈值时，会继续查询切片内容，可能带来一定的性能开销。 当开启“启用FAQ”功能，该参数显示。 FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 启用FAQ功能后，系统将优先检索FAQ数据。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。如果未超过阈值，则将继续进行文档检索。
查看来源	查看来源	添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。
查看图片	查看图片	开启此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。

版权所有 © 华为云计算技术有限公司

图 3-35 知识库高级配置



常见问题

为什么检索不到知识库内容?

- 可能原因：阈值设置太高；检索词语、策略不合适；精排模型不可用。
- 解决措施：调整阈值大小；修改检索词语、策略；在KooSearch服务中确认精排模型正常。

如何提升知识库检索准确率?

1. 优化文档结构：无结构的文档，转化为有层次结构的文档。
2. **团队标准版、团队高级版、企业专业版**或**企业旗舰版**支持更换接入的其他模型。

   当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。
3. 修改知识库的分段设置。

相关文档

AgentArts中配置知识库的详细信息，请参考**创建平台知识库**。

版权所有 © 华为云计算技术有限公司

3.6 为应用添加记忆

在多轮对话场景中，用户往往期望智能体能够记住自己的偏好、身份和交互习惯，从而获得连贯且个性化的服务体验。在单智能体配置中，支持通过配置变量为应用开启“记忆”功能。

AgentArts提供以下记忆类型，满足不同场景下的记忆需求：

- 入参变量： 如果需要通过API传入外部业务参数（如“订单ID”），请选择“入参变量”。通过API调用向智能体动态注入外部信息（如用户ID、业务单据编号），支持String、Boolean、Integer、Number类型，支持在工作流、插件和MCP服务的入参配置处引用。

约束与限制

表 3-16 记忆限制说明

类别	说明
变量	每个应用最多支持创建30个变量。 • 变量名称：不允许为空，最长支持100个字符，不能带有^符号。 • 描述最长支持200个字符，默认值最长支持500个字符。

变量类型说明

表 3-17 变量类型说明

变量类型	说明	适用场景
入参变量	入参变量用于在API调用时向智能体动态传入外部信息，通常用于注入具有确定性的业务逻辑、上下文信息或特定参数（如：用户ID、业务单据编号等确定性数据）。支持的数据类型包括String、Boolean、Integer、Number。 使用示例可参考**示例：配置与使用入参变量**。	• 确定性信息注入：入参变量适用于传入明确、确定的外部信息，而非由智能体自主推理获取的内容。例如，将当前用户的身份信息或业务状态作为入参传入，智能体据此执行对应的处理逻辑。 • API动态赋值：入参变量的值通过调用智能体API时传入，每次调用可传入不同的值，实现参数的动态替换，无需修改智能体配置。 • 组件间灵活引用：入参变量创建后，可在工作流、插件、MCP服务的入参配置处进行引用，将外部传入的信息无缝传递至各组件中使用。

版权所有 © 华为云计算技术有限公司

示例：配置与使用入参变量

使用场景：

当需要在智能体中固定某些参数值时，即适合使用入参变量。典型场景包括：

- 固定组件入参：为工作流、插件、MCP服务中的某个参数绑定固定值。例如，将天气查询服务的城市参数固定为“上海”，无论用户如何提问，服务始终查询上海的天气。
- 固定提示词上下文：在提示词中引用入参变量，为智能体注入固定的背景信息或约束条件。

以下通过一个天气查询场景，演示“入参变量”的完整配置方法和典型引用方式。

示例场景说明：

本示例将创建一个“城市旅行助手”智能体，为其添加天气查询插件，并通过入参变量将城市参数固定为"上海"。配置完成后，无论用户如何提问，智能体始终基于"上海"进行回复。

步骤1 创建单智能体。

参考**表3-18**创建一个“城市旅行助手”单智能体。

表 3-18 单智能体入参变量示例

单智能体信息	示例值
名称	城市旅行助手
描述	根据用户的提问进行旅行目的地问答。

步骤2 添加插件。

进入“城市旅行助手”的单智能体配置界面，添加“天气查询”类插件。添加插件的详细操作步骤可参考**添加插件**。

步骤3 新增入参变量。

在“记忆”模块的“入参变量”处，单击“+”新增入参变量，参考**表3-19**添加入参变量，单击“确定”。

表 3-19 添加入参变量

参数	说明	配置示例
名称	入参变量的名称，支持中英文字符，不支持 ^，最长 100 个字符。	城市
描述	非必填。用于标识该入参变量的用途，最长 200 个字符。	旅行目的地
类型	选择入参变量的数据类型，当前支持String、Boolean、Integer、Number。	String
默认值	非必填。入参变量的默认值。	上海

版权所有 © 华为云计算技术有限公司

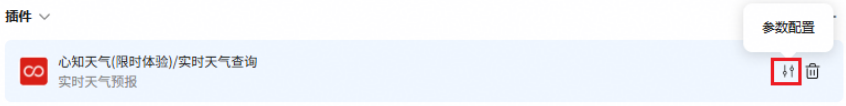
图 3-36 添加入参变量



步骤4 在插件参数中引用变量。

1. 鼠标放置在“天气查询”插件处，单击“天气查询”插件的“参数配置”图标。

图 3-37 参数配置

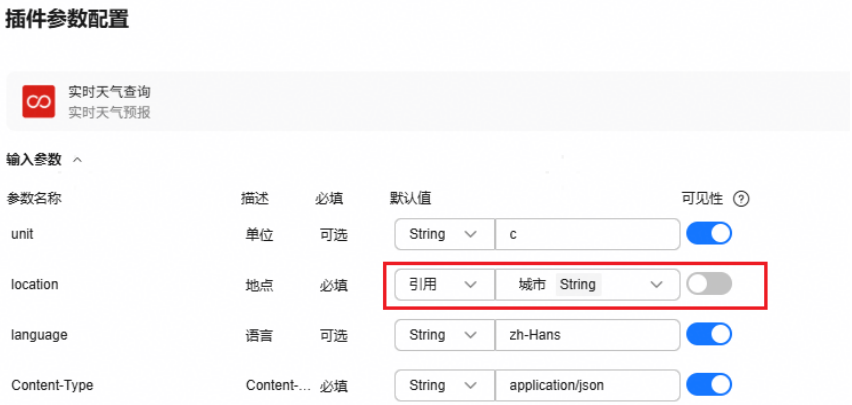


2. 在“插件参数配置”界面，单击“location”参数的“默认值”列，选择“引用”，然后选择已配置的“城市”入参变量。
3. 关闭“location”参数的“可见性”开关，固定该参数不被智能体修改，单击“确定”完成插件参数配置。

注意

如果“可见性”开关保持开启，智能体在对话过程中会根据用户的提问内容自动覆盖默认值，入参变量的固定效果将不生效。

图 3-38 插件参数配置



步骤5 预览调试。

在右侧“预览调试”窗口进行提问。此时无需在提问中提及城市名称，智能体也会默认查询并回复上海的天气信息。

版权所有 © 华为云计算技术有限公司

图 3-39 预览调试



----结束

3.7 提升应用对话体验

配置开场白

开场白是用户进入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智能体应用的功能和用途，明确如何与智能体应用进行有效交互。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 > 确定”按钮智能添加开场白。

图 3-40 配置开场白

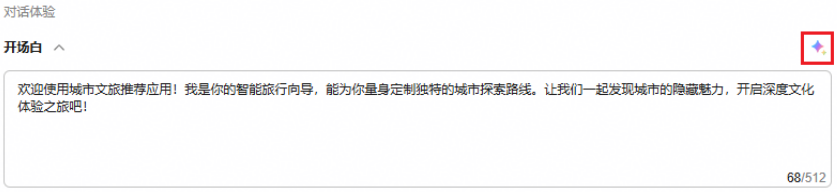
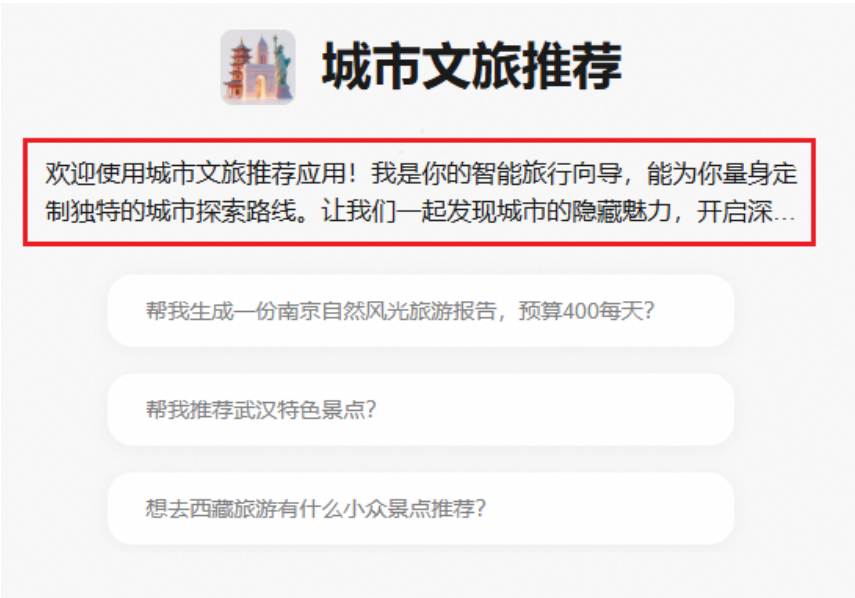


图 3-41 开场白配置示例



版权所有 © 华为云计算技术有限公司

配置推荐问题

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“ ”按钮智能添加推荐问题。推荐问题至多配置3条。

业务场景示例：在文旅助手场景中，新用户进入应用后往往不清楚如何开启咨询。通过配置推荐问题，如“帮我生成一份南京自然风光旅游报告，预算400每天？”、“帮我推荐武汉特色景点？”、“想去西藏旅游有什么小众景点推荐？”，可以有效引导用户直接切入核心业务流程，在减少用户打字负担的同时，显著提升业务转化率与问题解决效率。

图 3-42 配置推荐问题



图 3-43 推荐问题配置示例



版权所有 © 华为云计算技术有限公司

配置追问

追问功能是指智能体在与用户交互过程中，根据用户的回答或上下文，主动提出进一步的问题，以获取更多信息或澄清用户需求。这一功能能够提升对话的深度和准确性，帮助智能助手更好地理解用户意图，从而提供更高效的服务。

追问功能原理：追问功能的底层实现基于大模型的上下文理解与动态提问生成机制。在每轮对话结束后，模型会结合用户的当前输入、历史对话上下文以及用户预设的追问规则，计算并预测用户可能关心的后续信息缺口，从而动态生成具有逻辑连贯性的提问建议，精准挖掘用户的潜在需求。

用户自定义追问生成规则：用户可以根据需求，自定义追问生成规则。通过这些规则，智能体在每轮回复后，能够根据对话内容智能地提供提问建议，从而挖掘用户的潜在需求。

功能开启设置：在“对话体验 > 追问”中，用户可以选择是否开启“追问”功能。如果开启，模型将在每轮回复后，默认根据对话内容提供提问建议，进一步优化对话体验。

图 3-44 配置追问

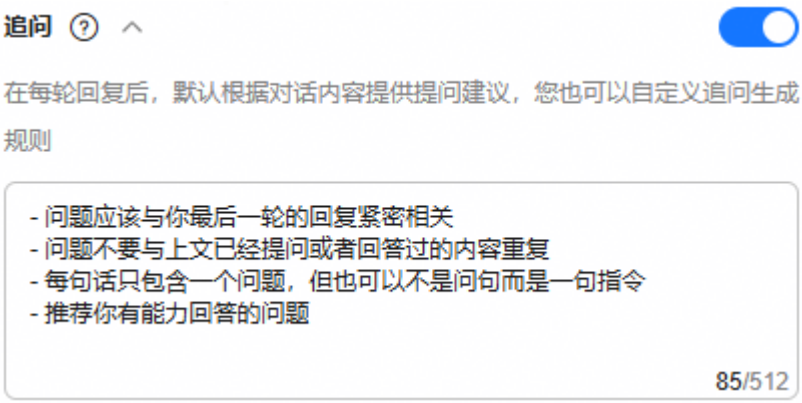


图 3-45 配置追问示例



版权所有 © 华为云计算技术有限公司

配置引用和归属

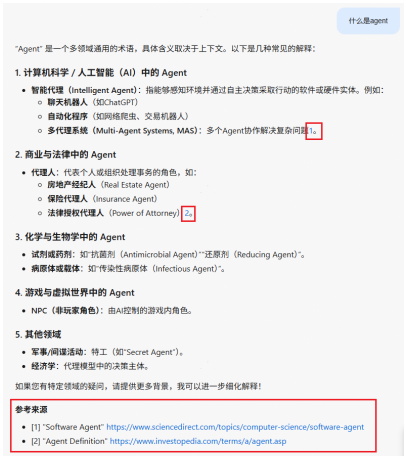
在单智能体配置并使用了联网搜索或知识库等工具时，可以开启“引用和归属”功能。开启后，模型生成的回答将包含引用来源，帮助核实信息的准确性与时效性。

单击“⇄”支持配置模型回复的输出类型，详情请参考表3-20。

图 3-46 配置引用和归属



表 3-20 引用和归属输出类型

输出类型	说明	效果展示
文字输出	引用来源以文本形式附加在模型回复中。系统在正文引用处标注数字序号，并在末尾展示对应的参考来源。 适用于直接展示给最终用户的对话场景。	
JSON输出	引用来源以结构化JSON格式返回。返回内容包含模型回答“content”和引用来源列表“reference”字段。 适用于开发者通过API调用并进行自定义开发的场景。	

配置安全信息

安全信息支持配置内容审核配置。

版权所有 © 华为云计算技术有限公司

内容审核配置

内容审核配置功能旨在帮助用户对输入和输出内容进行严格审查与管理，确保内容符合预设的规则和标准。该功能支持自定义审核规则、关键词匹配、以及多种处理策略，适用于多种应用场景。

支持通过单击右侧的开关按钮“启动”或“关闭”内容审核配置功能。内容审核配置功能开启时，可通过单击  图标设置关键词匹配处理输入输出内容，支持添加多个审核规则。

- **审核规则设置**

  用户可创建多个审核规则，每个规则可独立配置关键词和处理策略。系统支持多规则并行审核，满足不同场景下的内容管理需求。

  - 示例：系统中可设置“审核规则-1”、“审核规则-2”等，用户可根据实际需要添加更多规则。
  - 每个规则可独立配置关键词和处理策略，互不干扰。

- **关键词配置**

  在每个审核规则中，用户可设置关键词用于内容匹配。关键词支持多行输入，每行一个关键词，最多支持**10000字符**。

  - 系统将对应用输入或输出内容进行关键词匹配，如果匹配成功，将根据预设策略进行处理。
  - 关键词支持中文、英文、数字及特殊字符，适用于多种内容类型。

- **审查用户输入内容**

  用户可单击右侧的开关按钮“启动”或“关闭”“审查用户输入内容”功能。当用户输入内容中包含关键词时，系统将自动根据预设策略进行处理。

  - **预设回复内容**：用户可自定义回复文案，用于替代原始输入内容，回复内容最多支持**1000字符**。
  - 该功能可用于拦截不当输入、引导用户规范表达等场景。

- **审查应用输出内容**

  用户还可单击右侧的开关按钮“启动”或“关闭”“审查应用输出内容”功能。当系统输出内容中包含关键词时，系统将自动根据预设策略进行处理。

  - **输出处理方式包括**：
    - **回复预设文案**：输出预设的固定内容，替代原始输出。
    - **替换内容**：将匹配的关键词替换为指定内容。
    - **屏蔽内容**：直接屏蔽匹配内容，不进行输出。
  - 用户可根据实际需求选择合适的策略，灵活控制输出内容。

- **添加规则**

  用户可通过“添加规则”按钮，新增审核规则，以满足不同场景下的内容审核需求。

  - 每个新增规则均可独立配置关键词和处理策略。
  - 支持多规则并行审核，提升内容管理的灵活性和准确性。

版权所有 © 华为云计算技术有限公司

3.8 调试应用

开发者可以在智能体应用搭建完成后，直接与智能体应用进行对话，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，允许开发者查看每个用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而准确定位问题并快速调整配置。

⚠ 注意

调试过程会消耗模型推理资源及关联服务资源（如语音识别、文件解析等），请关注您的资源消耗与计费情况。

调试应用

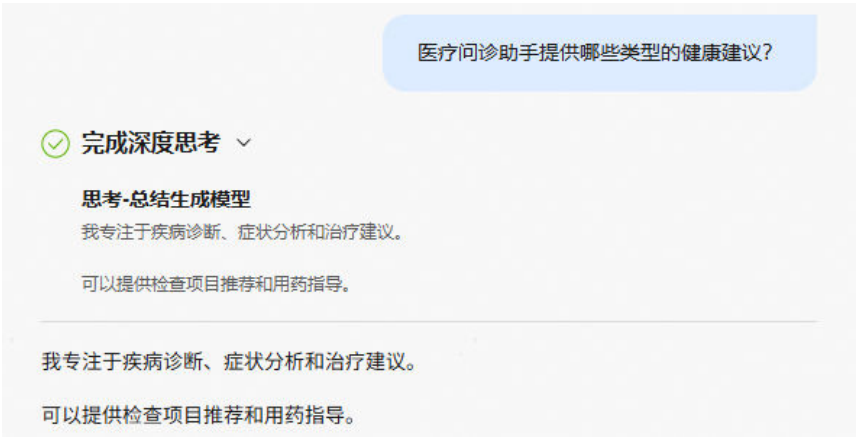
创建应用后，平台支持对应用执行过程进行预览调试。

步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击目标单智能体应用，在“预览调试”界面文本框中输入对话，Agent应用将根据对话生成相应的回答。

图 3-47 调试应用



 版权所有 © 华为云计算技术有限公司

说明

- 预览调试界面支持文本输入和文件输入：
  - 文本输入：在对话输入框输入对话后按Enter键，查看应用响应结果。
  - 文件输入：用户可以通过上传文件进行提问，支持对文件进行解析，并根据文件内容和问题生成准确的答案。解析功能需要在智能体中添加对应解析文件类型的插件或MCP。
    - 支持上传image、audio、excel、csv、docx等格式的文件。
    - 最多支持上传10个文件。
- 调试结果支持一键复制，单击，可将回复内容复制进剪贴板。
- 单击，一键清除预览调试界面内容。

步骤4 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 3-48 查看调试结果



说明

用户可根据调试结果查看是否符合预期，如果需调整智能体配置请参考**选择并配置模型**。

----结束

调试信息说明

“调试”界面支持查看“运行结果”和“调用详情”。

- **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，从而直观地认识性能的情况。

 版权所有 © 华为云计算技术有限公司

图 3-49 查看运行结果



- **调用详情**

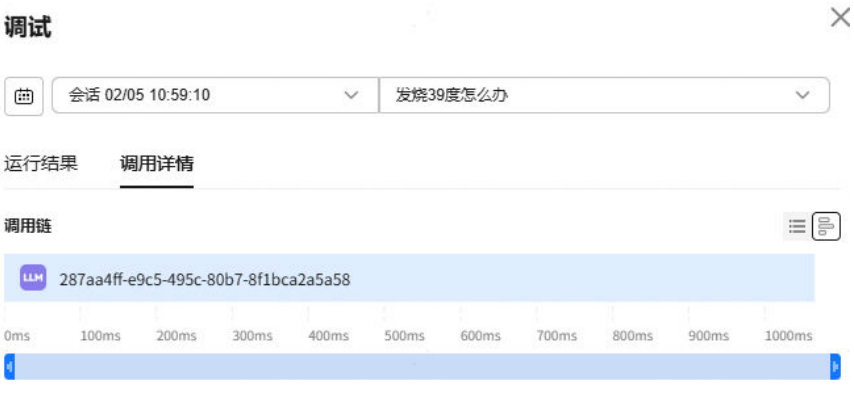
  在触发应用时，调用链中展现具体事件的详细信息，包括触发的组件、事件耗时、事件的输入和输出信息等。便于开发者快速地追溯操作顺序并精确定位问题。

 版权所有 © 华为云计算技术有限公司

图 3-50 查看调用详情（列表）



图 3-51 查看调用详情（火焰图）



版权所有 © 华为云计算技术有限公司

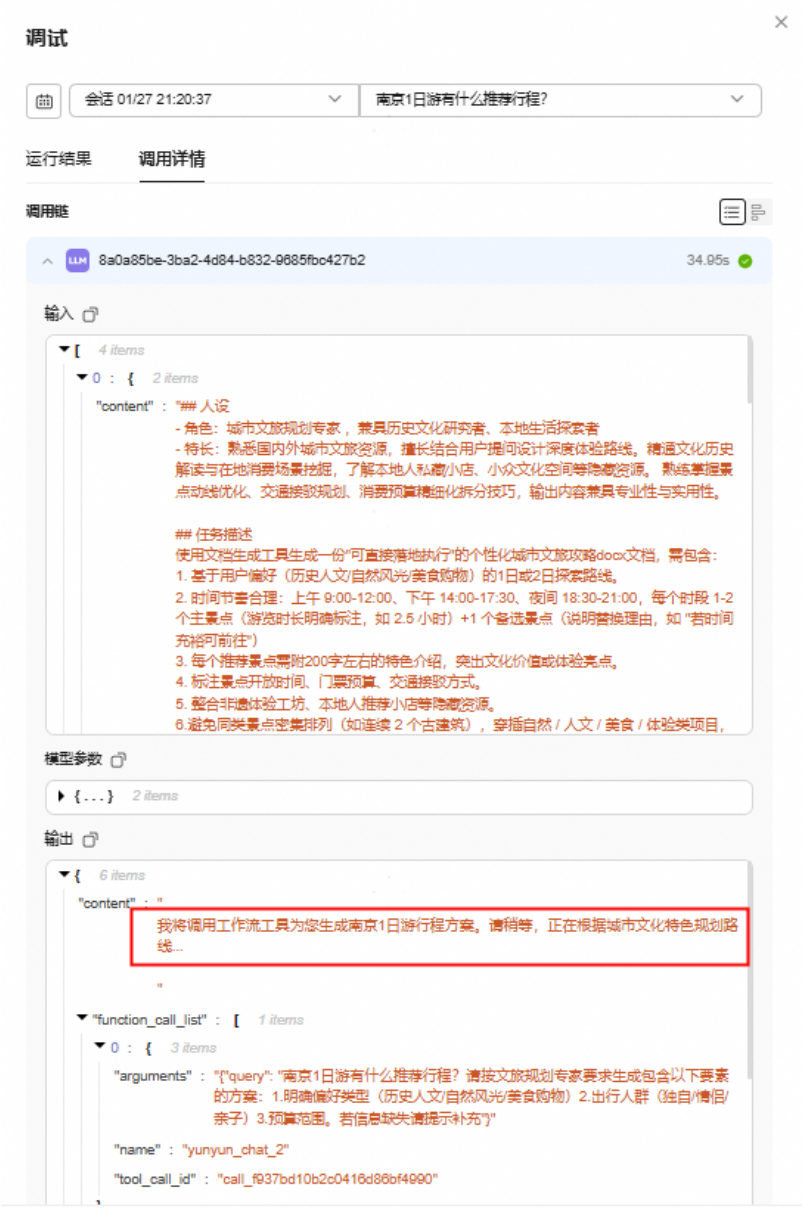
说明

- 单击调用详情页面中的  按钮，查看列表调用链，适用于查看具体的输入输出参数、排查逻辑错误。
- 单击调用详情页面中的  按钮，查看火焰图调用链，适用于分析性能瓶颈、定位耗时较长的节点。

 版权所有 © 华为云计算技术有限公司

常见调试场景

表 3-21 常见调试场景

场景	调试方法
工作流	当智能体采用工作流时，您可以通过调试界面查看每个请求响应的工作流节点的输入和输出，以及出现错误的节点，这有助于您快速定位工作流中存在问题的节点。 
MCP服务	可通过调试界面查看每次请求响应中调用的MCP服务的运行情况。
插件	可通过调试界面查看每次请求响应中调用的插件的运行情况，以及运行出错的插件节点，根据报错信息调整插件参数或检查网络连通性。

 版权所有 © 华为云计算技术有限公司

场景	调试方法
知识库	知识库是智能体配置中重要的信息存储和知识来源，其召回速度和准确性会直接影响回复的准确性。在调试界面中可查看知识库节点的运行情况。如果召回为空或准确率低，请检查知识库切片配置或更新文档数据。

3.9 配置触发器

触发器是Agent中实现任务自动化的关键功能，它能够在特定条件下自动启动任务执行，无需人工干预。可以利用触发器灵活设置任务的启动条件，触发器能确保任务按时按需完成，从而提高Agent应用的自动化水平和响应速度。

触发方式

通过定时任务调用应用按照触发指令要求执行任务。

添加触发器

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击目标单智能体应用，在智能体应用右上角单击触发器配置按钮 。

步骤4 在触发器配置页面右侧，单击+。

步骤5 配置触发器信息，参数配置说明请参考**表3-22**。

表 3-22 创建触发器参数说明

参数	说明
触发器名称	触发器的名称。 由2~20个字符组成，支持中英文、数字、下划线，仅支持中英文开头。
触发类型	• 周期触发：用户可设定每周、每日或每月固定时间触发。 • 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。

版权所有 © 华为云计算技术有限公司

参数	说明
触发时间	按设置的时间触发智能体应用的执行。例如，设置触发时间为每1小时执行一次，则每隔1小时，重复执行一次会话。 ● 周期触发 – 用户可自定义触发时间按每日执行。 – 用户可自定义触发时间按每周执行。 – 用户可自定义触发时间按每月执行。 ● 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。 – 间隔“天”：取值范围1~31。 – 间隔“小时”：取值范围1~23。 – 间隔“分钟”：取值范围1~59。 – 间隔“秒”：取值范围1~59。
机器人提示	输入自然语言指令，触发时Agent遵循该指令定时执行。 ● 配置示例1：输入“每天9点发送早报”，触发器会在每天9点自动执行发送早报的任务。 ● 配置示例2：输入“发送月底发票报销提醒”，则机器人会在设定的时间自动发送提醒。

图 3-52 创建定时任务触发器



步骤6 单击“确定”，即完成触发器创建，可在触发器列表中对触发器进行查看，修改和删除等操作。

----结束

版权所有 © 华为云计算技术有限公司

3.10 发布应用

3.10.1 发布并部署应用为 API 服务

将Agent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的Agent应用发布并部署为API服务。

说明

提交版本是将当前单智能体应用配置保存为一个不可变的版本快照；部署是将指定版本发布为可调用的API接口，创建一个独立的运行实例。

费用说明

智能体发布后通过API接口调用时，会产生tokens消耗。

- 如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

注意

智能体部署并通过API调用后，除模型tokens消耗外，如果开启了日志记录（LTS）、调用链（APM）、指标（AOM）功能，以及使用了API Key认证（DEW凭据托管），均会产生按需计费费用。建议在部署前了解各项服务的计费规则，合理控制成本。

前提条件

- 已**开通AgentArts服务**。
- 已**创建单智能体应用**。

提交版本

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击待提交版本的单智能体应用卡片。

步骤4 在“单智能体配置”页面右上角单击“提交版本”。

说明

已经提交版本的单智能体应用，修改后再次发布，显示为“更新版本”。

步骤5 在“提交版本”页面，配置发布信息，具体参数说明请参考**表3-23**。

版权所有 © 华为云计算技术有限公司

表 3-23 发布单智能体参数说明

配置项	功能描述
版本名称	必填。 系统根据时间戳自动生成，同时支持用户自定义。支持英文、数字、下划线，长度1~64字符。
版本描述	（可选）在此处输入版本的详细描述，以帮助团队成员和用户了解版本更新的具体内容和改进点。 长度0~1024字符，建议填写本次更新的关键变更点，方便后续版本管理。
部署至实例	默认勾选。 将该版本同步部署为API接口，便于后续接口调用。

图 3-53 提交版本



步骤6 单击“确定”。

提交成功后，在“单智能体”页面的卡片上，显示“已提交”。您可在版本历史记录中查看已提交的版本。

如果勾选“部署至实例”，进入“部署智能体”页面，具体部署操作请参考**部署智能体**。

步骤7 （可选）查看api调用接口信息。

1. 在“版本提交”页面，单击“分享”，进入“渠道管理”页面，可以查看API调用接口信息。

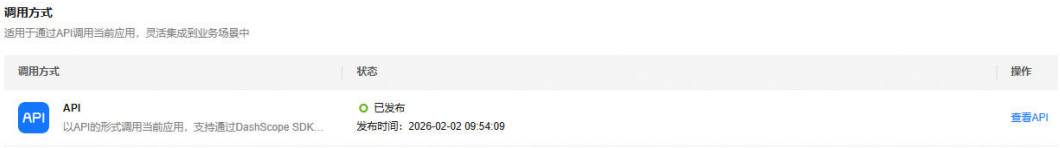
   也可通过左侧导航栏中选择“开发中心 > 智能体管理”，单击目标应用，进入单智能体配置页面，选择上方“渠道管理”页签，进入渠道管理页面。
2. 在“渠道管理”中单击“查看API”即可查看API调用接口信息。

   说明

   如果未部署过任何版本，“查看API”页面中的API详情中的URL为空。请先完成部署后再查看。

版权所有 © 华为云计算技术有限公司

图 3-54 调用 API



----结束

部署智能体

未部署的智能体，不能通过API接口调用，部署是以版本为单位进行部署。

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击待部署的单智能体应用卡片。

步骤4 在“单智能体配置”页面右上角，单击  。

步骤5 在“部署智能体”页面，配置部署信息如**图3-55**所示，参数配置请参考**表3-24**。

图 3-55 部署智能体

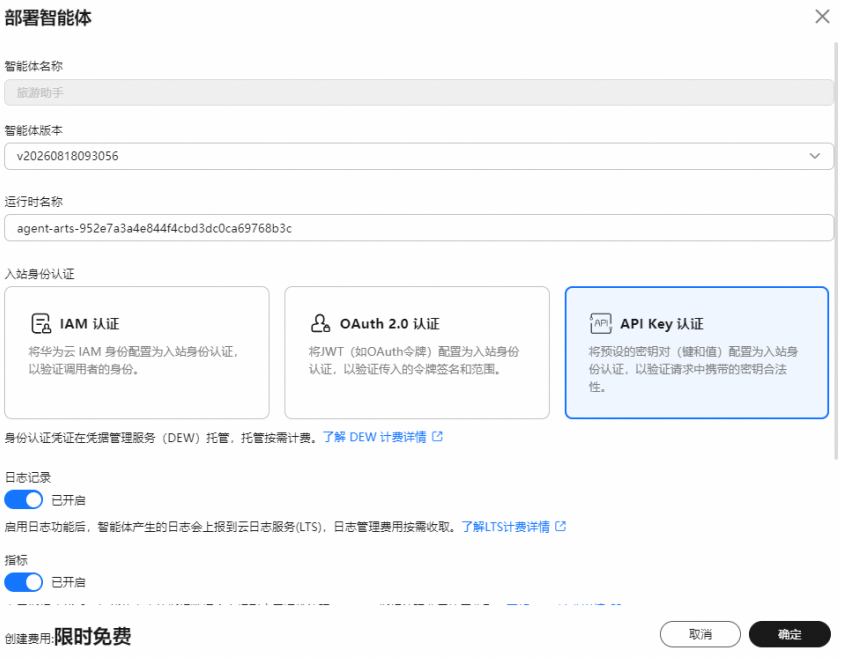


表 3-24 部署智能体

参数	说明
智能体名称	为**创建单智能体应用**自定义的智能体名称，不可修改。

版权所有 © 华为云计算技术有限公司

参数	说明
智能体版本	为**提交版本**的版本名称。默认为最新的版本名称，可以在下拉框中切换。
运行时名称	运行时的名称。**仅第一次部署时支持自定义。** 由2~48个字符组成，包含小写字母、数字、中划线，以小写字母开头，以小写字母或数字结尾。
入站身份认证	选择调用此API接口时，使用的认证鉴权方式，如何进行认证鉴权请参考**认证鉴权**。默认IAM认证，**仅第一次部署时支持选择。** • IAM认证：将华为云IAM身份配置为入站身份认证，以验证调用者的身份。适用于华为云内部系统或已接入华为云IAM的系统，集成复杂度低，安全等级高。 • OAuth 2.0认证：将JWT（如OAuth令牌）配置为入站身份认证，以验证传入的令牌签名和范围。适用于已有身份提供商（如Okta、Cognito等）的企业系统，集成复杂度高，安全等级高。 – Discovery URL：身份提供商（例如Okta、Cognito等）提供的Discovery URL，该URL通常可在该提供商的文档中找到。 – JWT授权配置：配置JWT令牌的访问控制规则。 ▪ 允许的受众：指定的受众是否在Agent Identity运行中指定的受众匹配或为其子集。 ▪ 允许的客户端：指定的客户端标识符是否被允许访问Agent Identity。 ▪ 允许的范围：仅当令牌包含此处配置的至少一个必须范围时，才允许访问自定义声明。 ▪ 自定义声明：仅当令牌中的特定声明与预定义字符串值相匹配时，才允许访问。 • API Key认证：将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。适用于内部测试或轻量级集成，集成复杂度低，安全等级中。 **注意** 身份认证凭证在凭据管理服务（DEW）托管，托管按需计费，使用DEW产生的费用请参考**DEW计费说明**。
日志记录	默认开启。 开启日志记录功能后，智能体产生的日志会上报到云日志服务LTS，便于日志的集中管理和分析。日志管理费用按需收取，具体计费详情请参考**LTS计费说明**。 日志信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“日志”页签中。

版权所有 © 华为云计算技术有限公司

参数	说明
指标	默认开启。 开启指标功能后，智能体产生的指标数据会上报到应用运维管理AOM，便于实时监控和分析应用的运行状态。指标管理费用按需收取，具体计费详情请参考**AOM计费说明**。 指标信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“指标”页签中。
调用链	默认开启。 开启调用链功能后，智能体产生的调用链数据会上报到应用性能管理 APM，帮助开发者追踪和分析应用的性能问题。调用链管理费用按需收取，具体计费详情请参考**APM计费说明**。 调用链信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“链路信息”页签中。
智能体卫士	默认关闭。 开启后，实时检测智能体执行链路与操作轨迹，监控全量行为并识别安全异常。不同安全风险处理规则和检测记录可参考**企业主机安全 HSS**管理及查看。HSS的智能体卫士当前为公测阶段，暂不收费。

步骤6 单击“确定”。

部署成功后，可以在“部署运行 > 智能体运行时”页面，通过搜索运行时名称查看部署的智能体。

单智能体部署后，支持通过API调用，具体请参考**通过API调用单智能体应用**。

步骤7 （可选）查看api调用接口信息。

1. 在“发布部署”页面，单击“分享”，进入“渠道管理”页面。
2. 在“渠道管理”中单击“查看API”即可查看API调用接口信息。

----结束

典型场景

- 系统集成：将智能体能力集成到企业OA系统或客服工作台，通过API实现后台自动处理。
- 批量处理：通过脚本批量调用API，自动化处理大量文本数据。
- 自定义前端：获取API接口后，结合企业自有UI框架构建专属的对话交互界面。

常见问题

发布后API调用返回错误码AgentArts.02001017如何处理?

原因：表示您的智能体API调用次数已达到当前套餐的额度限制。系统检测到您的API调用量已超过套餐包含的调用次数上限。

解决办法：

1. 登录AgentArts控制台，检查当前套餐的API调用额度及使用情况。

版权所有 © 华为云计算技术有限公司

2. 如需继续使用API服务，请升级至更高套餐以获得更大的API调用额度。
3. 升级套餐后，API调用功能将恢复正常。

3.10.2 发布应用为网页应用

AgentArts支持将Agent应用发布为网页应用，这一功能适用于多种场景。例如，在客户服务领域，可以将智能客服Agent部署为网页应用，提供全天候的在线咨询服务，提升客户满意度。在教育和培训中，可以将教学助手Agent部署为网页应用，为学生提供个性化的学习指导和互动体验。在智能问答系统中，可以将知识库查询Agent部署为网页应用，为用户提供快速准确的信息检索服务。此外，还可以将Agent应用部署为网页应用，用于数据分析、内容推荐、图像识别等场景，为用户提供便捷的交互界面和高效的服务。通过这种方式，AgentArts帮助开发者快速将Agent应用转化为实际可用的在线服务，提高应用的可用性和用户满意度。

前提条件

- 已**开通AgentArts服务**。
- 已**创建单智能体应用**。

发布应用为网页应用

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击待提交版本的单智能体应用卡片。

步骤4 在“单智能体配置”页面右上角单击“提交版本”。

说明

已经提交版本的单智能体应用，修改后再次发布，显示为“更新版本”。

步骤5 在“提交版本”页面，配置发布信息，具体参数说明请参考**表3-25**。

表 3-25 发布单智能体参数说明

配置项	功能描述
版本名称	必填。 系统根据时间戳自动生成，同时支持用户自定义。支持英文、数字、下划线，长度1~64字符。
版本描述	（可选）在此处输入版本的详细描述，以帮助团队成员和用户了解版本更新的具体内容和改进点。 长度0~1024字符，建议填写本次更新的关键变更点，方便后续版本管理。
部署至实例	默认勾选。 将该版本同步部署为API接口，便于后续接口调用。

版权所有 © 华为云计算技术有限公司

图 3-56 提交版本



步骤6 单击“确定”。

提交成功后，在“单智能体”页面的卡片上，显示“已提交”。您可在版本历史记录中查看已提交的版本。

如果勾选“部署至实例”，进入“部署智能体”页面，具体部署操作请参考**部署智能体**。

步骤7 发布完成后可以单击“分享”按钮跳转至“渠道管理”页面。

也可通过左侧导航栏中选择“开发中心 > 智能体管理”，单击目标应用，进入单智能体配置页面，选择上方“渠道管理”页签，进入渠道管理页面。

在**渠道管理**页面，**网页**分享渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。

图 3-57 发布 Agent



图 3-58 应用发布成功后



- 立即访问：单击当前页面的“立即访问”按钮，可立即进入网页版应用。
- 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。
- 重新生成：重新生成应用链接。生成的应用链接有效时限为24小时（从发布为网页或“重新生成”算起），超过有效时限则无法正常访问链接，需重新单击网页分享渠道的**操作**列的“重新生成”。

版权所有 © 华为云计算技术有限公司

说明

- 已发布的应用支持在Agent应用编辑页面选择“更新版本”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。
- 发布为网页的应用，每日单用户调用应用的次数不超过100次，且不超过“每日调用限额（次）”设置的限额。

----结束

查看发布历史

单击右上角，可查看当前单智能体发布历史记录。

图 3-59 发布历史



说明

发布历史支持版本对比、还原版本和删除操作：

- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前单智能体应用至该历史版本的状态。当前最新配置将被覆盖，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前单智能体应用发布历史。

3.10.3 发布应用至云商店

华为云云商店是面向企业和开发者的数字化商品交易平台，支持将 API 能力以标准化商品形式对外售卖。通过将 AgentArts 智能体发布至云商店，您可以：

版权所有 © 华为云计算技术有限公司

- **实现能力变现**：将构建好的智能体封装为可购买的 API 商品，按调用次数向买家收费。
- **扩大服务覆盖**：借助云商店的流量和分发能力，让更多企业用户发现并使用您的智能体服务。
- **标准化交付**：通过 APIG 网关统一管理 API 的鉴权、限流和计费，无需自行开发商业化能力。

费用说明

智能体发布后通过API接口调用时，会产生tokens消耗。

- 如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

⚠ 注意

智能体部署并通过API调用后，除模型tokens消耗外，如果开启了日志记录（LTS）、调用链（APM）、指标（AOM）功能，以及使用了API Key认证（DEW凭据托管），均会产生按需计费费用。建议在部署前了解各项服务的计费规则，合理控制成本。

前提条件

- 已**开通AgentArts服务**。
- 已**创建单智能体应用**。

提交版本并分享至云商店渠道

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击待提交版本的单智能体应用卡片。

步骤4 在“单智能体配置”页面右上角单击“提交版本”。

说明

已经提交版本的单智能体应用，修改后再次发布，显示为“更新版本”。

步骤5 在“提交版本”页面，配置发布信息，具体参数说明请参考**表3-26**。

表 3-26 发布单智能体参数说明

配置项	功能描述
版本名称	必填。 系统根据时间戳自动生成，同时支持用户自定义。支持英文、数字、下划线，长度1~64字符。

版权所有 © 华为云计算技术有限公司

配置项	功能描述
版本描述	（可选）在此处输入版本的详细描述，以帮助团队成员和用户了解版本更新的具体内容和改进点。 长度0~1024字符，建议填写本次更新的关键变更点，方便后续版本管理。
部署至实例	默认勾选。 将该版本同步部署为API接口，便于后续接口调用。

图 3-60 提交版本



步骤6 单击“确定”。

提交成功后，在“单智能体”页面的卡片上，显示“已提交”。您可在版本历史记录中查看已提交的版本。

如果勾选“部署至实例”，进入“部署智能体”页面，具体部署操作请参考**部署智能体**。

步骤7 发布完成后可以单击“分享”按钮跳转至“渠道管理”页面。

也可通过左侧导航栏中选择“开发中心 > 智能体管理”，单击目标应用，进入单智能体配置页面，选择上方“渠道管理”页签，进入渠道管理页面。

在**渠道管理**页面，**云商店**分享渠道的**操作**列单击“发布”按钮后将出现“复制链接”。

单击“复制链接”，可直接复制URL嵌入或分享到其他应用场景。

版权所有 © 华为云计算技术有限公司

说明

- 如果未部署过任何版本，“复制链接”复制到的内容为空。请先完成部署后再查看。
- 已发布的应用支持在Agent应用编辑页面选择“更新版本”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用。

----结束

对接 APIG 网关并提交商品发布

Agent应用分享到云商店渠道后，商家可将应用发布到华为云云商店。

APIG 网关类商品要求商家先在华为云 APIG 共享版网关上创建并发布 API，云商店发布商品时绑定 APIG 的 API 分组。

说明

在华为云云商店发布产品请先完成这些**准备工作**，否则无法正常管理和上架。

步骤1 获取智能体API 信息。

1. 在左侧导航栏选择“开发中心 > 智能体管理”，单击目标应用，选择上方“渠道管理”页签。
2. 单击API调用方式操作列的“查看API”，进入API详情页，可获取服务地址、请求Path、请求方式、凭证Key、请求参数。

说明

如果未部署过任何版本，“查看API”页面中的API详情中的URL为空，请先完成部署后再查看。具体部署操作请参考**部署智能体**。

图 3-61 服务地址



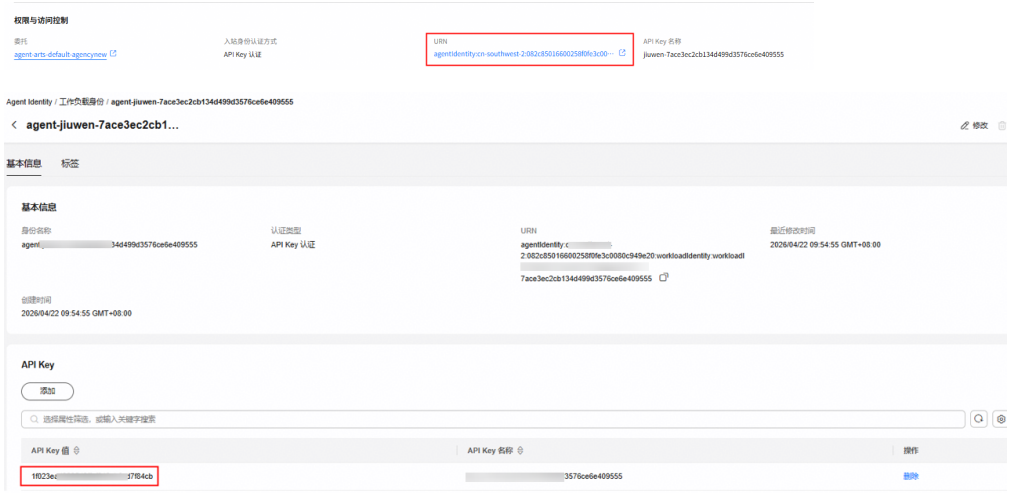
 版权所有 © 华为云计算技术有限公司

图 3-62 请求信息



3. 获取服务鉴权信息：凭证Value。

   在API详情页复制智能体ID，跳转至“部署运行 > 智能体运行时”进行搜索，在智能体运行详情页面打开RUN界面即可获取API Key。



步骤2 在 APIG 网关完成技术对接，详细操作请参见**APIG网关类商品技术对接方案**。

创建API填写样例：

版权所有 © 华为云计算技术有限公司

图 3-63 定义 API 请求

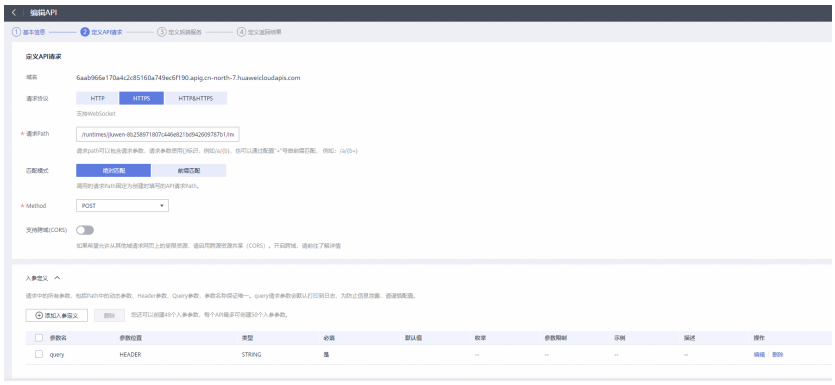
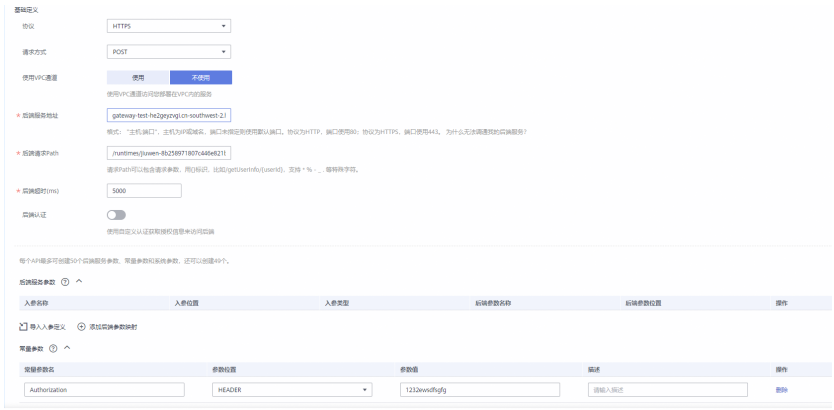


图 3-64 定义后端服务



步骤3 在卖家中心创建 APIG 网关类商品并提交审核，详细操作请参见**发布APIG网关类商品**。

----结束

准备工作

如需在华为云云商店发布产品请先完成以下准备，否则无法正常管理和上架：

步骤1 完成企业账号实名认证

华为云账号需完成企业实名认证，支持企业法人支付宝认证、对公账户打款认证等多种方式。详细操作请参见**企业账号实名认证流程**。

步骤2 加入华为云合作伙伴网络HCPN

完成实名认证后，申请加入 HCPN，获取合作伙伴基本权益，这是进入卖家中心的前置条件。详细操作请参见**加入华为云合作伙伴网络（HCPN）**。

步骤3 注册成为华为云商家

加入 HCPN 后，在卖家中心提交入驻申请，完成商业信息认证（含银行账号、税率等结算信息），审核通过后方可发布商品。详细操作请参见**注册成为商家**。

步骤4 开通华为支付（可选）

 版权所有 © 华为云计算技术有限公司

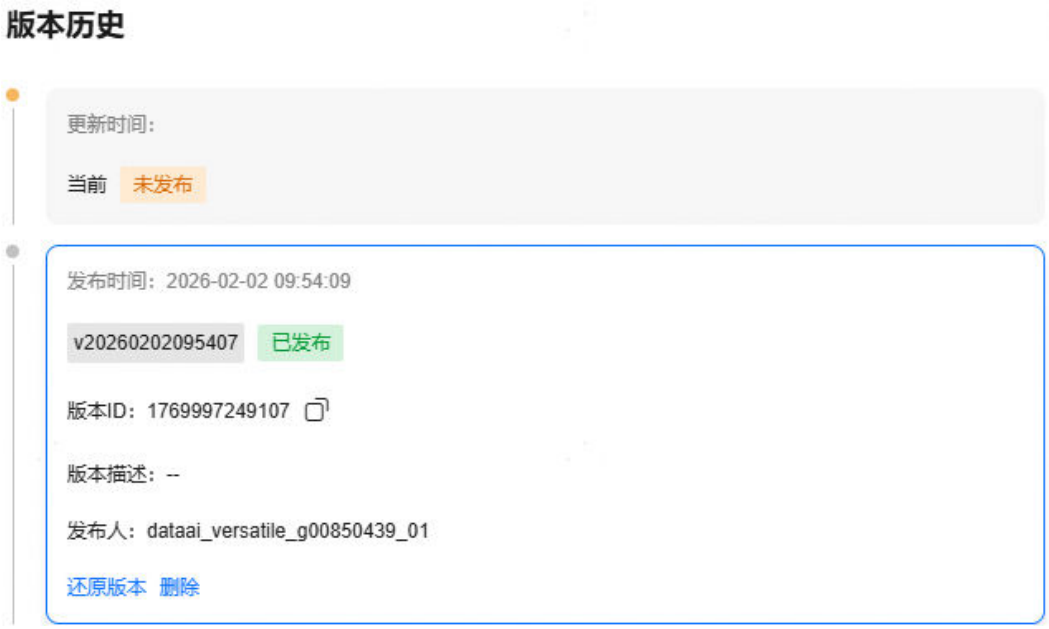
若需在云商店向买家收费，请提前完成华为支付开通。免费商品可跳过此步骤。详细操作请参见**开通华为支付**。

----结束

查看发布历史

单击右上角  ，可查看当前单智能体发布历史记录。

图 3-65 发布历史



说明

发布历史支持版本对比、还原版本和删除操作：

- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前单智能体应用至该历史版本的状态。当前最新配置将被覆盖，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前单智能体应用发布历史。

常见问题

发布后API调用返回错误码AgentArts.02001017如何处理?

原因：表示您的智能体API调用次数已达到当前套餐的额度限制。系统检测到您的API调用量已超过套餐包含的调用次数上限。

解决办法：

1. 登录AgentArts控制台，检查当前套餐的API调用额度及使用情况。
2. 如需继续使用API服务，请升级至更高套餐以获得更大的API调用额度。
3. 升级套餐后，API调用功能将恢复正常。

版权所有 © 华为云计算技术有限公司

3.11 通过 API 调用单智能体应用

AgentArts提供OpenAPI请求方式，可通过调用路径发送请求，程序将调用应用并返回预期结果。

AgentArts的API调用是应用开发中的强大工具，可以帮助用户快速集成功能和服务，同时支持与其他系统或服务进行交互，提升应用性能和用户体验。合理设计和管理API是确保应用安全和稳定的关键。通过API，用户可以构建功能丰富、高效的应用，满足多样化的用户需求。

前提条件

在调用应用前，须确保应用已发布，具体请参考**发布应用为API服务**。

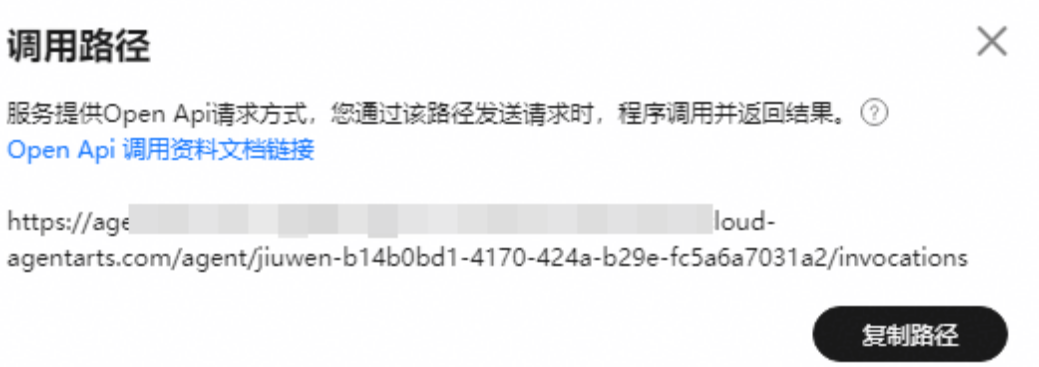
获取调用路径

步骤1 登录**AgentArts智能体平台**。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击“调用路径”，在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如**图3-66**所示。

图 3-66 获取应用调用路径



----结束

使用 API 调用单智能体应用

使用API调用单智能体的操作，请参考**使用API调用单/多智能体**。

3.12 管理应用

在AgentArts中创建应用之后，可以管理单智能体应用界面中的应用，可执行删除、复制创建的应用、复制应用ID、导入、导出及查看调用路径等。

版权所有 © 华为云计算技术有限公司

复制应用

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 鼠标移动到目标应用，单击应用下方的“复制”，即可复制该应用。

图 3-67 复制 Agent 应用



----结束

删除应用

警告

- 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，但不可恢复，请谨慎操作。

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 鼠标移动到目标应用，单击应用下方的“更多 > 删除”。

图 3-68 删除 Agent 应用



版权所有 © 华为云计算技术有限公司

步骤4 弹出的对话框中单击“确定”。

图 3-69 删除确认



----结束

复制 ID

对于Agent应用除了具有页面操作的能力之外，还具有Chat API调用能力，因此获取AppID十分必要。该ID为调用Agent应用接口的agent_id字段。

```
POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}
```

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 鼠标移动到目标应用，单击应用下方的“更多 > 复制ID”。

图 3-70 复制 ID



步骤4 弹出复制成功对话框，表示已成功复制该应用的agent_id，可将其粘贴至调用Agent应用接口的对应字段中。

图 3-71 复制 ID



----结束

版权所有 © 华为云计算技术有限公司

调用路径

调用路径可为Agent应用的API接口。详细API调用过程请参见**使用API调用单智能体应用**。

图 3-72 获取调用路径



导入应用

平台支持导入单智能体应用。导入单智能体应用时，将同步导入单智能体应用关联的插件等配置。

步骤1 登录**AgentArts智能体平台**。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 导入单智能体应用。

1. 单击页面右上角“导入”。
2. 在“导入”页面，单击“选择文件”选择需要导入的jsonl格式文件。

版权所有 © 华为云计算技术有限公司

图 3-73 导入



图 3-74 导入单智能体



版权所有 © 华为云计算技术有限公司

说明

- 如果应用关联的资源（如插件、MCP、知识库）导入失败，应用仍会正常导入，但这些资源将不会被导入。
- 如果导入失败，可在操作列查看原因。
- 如果导入的智能体中存在资源未配置鉴权，系统会标注“未配置鉴权”，单击操作列中的“配置鉴权”后可在配置鉴权窗口中进行操作。

3. 选择导入文件后，选择解析内容。

   平台将自动解析jsonl文件。如果解析的文件已存在，勾选该文件将自动覆盖平台现有文件。

4. 单击“导入”，导入成功的单智能体应用将在“开发中心 > 应用管理 > 单智能体应用”页面中展示。

说明

仅支持上传jsonl格式文件，文件的最大导入大小为128MB。

----结束

导出应用

平台支持导出单智能体应用。导出单智能体应用时，将同步导出单智能体应用关联的插件等配置。

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 导出单智能体应用。

1. 单击页面右上角“导出”，或鼠标移动到目标单智能体，单击“更多 > 导出”。
2. 在“导出”页面选择单智能体应用，单击“导出”。在导出界面单击“下载”，智能体将以一个jsonl格式的文件下载至本地。

图 3-75 导出单智能体



版权所有 © 华为云计算技术有限公司

说明

- 如果应用关联的资源（如插件、MCP、知识库）导出失败，则导出的文件中将不包含对应资源。
- 如果导出失败，可在操作列查看原因。

----结束

渠道管理

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 鼠标移动到目标应用，单击应用下方的“渠道管理”。

图 3-76 渠道管理



步骤4 跳转至“发布管理”页面。详细操作可参见发布应用。

----结束

查看智能体运行时数据

- 在“智能体运行时”页面，通过搜索智能体ID搜索查看智能体运行时数据。
  a. 登录AgentArts智能体平台。
  b. 在左侧导航选择“部署运行 > 智能体运行时”，在搜索框中输入已获取的ID进行搜索。
  c. 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。
- 在“运营运维 > 观测 > 智能体运行分析”页面，通过搜索智能体ID查看智能体运行时数据。
  a. 登录AgentArts智能体平台。
  b. 在左侧导航选择“运营运维 > 观测 > 智能体运行分析”，在搜索框中输入已获取的ID进行搜索。
  c. 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。

版权所有 © 华为云计算技术有限公司

4 开发工作流应用

4.1 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。

工作流的核心是一个由人工设计和编排的智能系统，由多个AI Agent通过利用自然语言处理（NLP）和大型语言模型（LLM）协作完成任务。这些智能体在预设的逻辑框架下工作，能够根据既定规则自主感知、推理和行动，以追求特定目标，形成强大的集体智慧，可以打破信息孤岛，集成不同的数据源，并提供无缝的端到端自动化。

AgentArts平台提供一个可视化的画布，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。

费用说明

运行工作流将调用工作流内各节点对应的服务（如大模型调用），这可能产生相应费用。具体计费规则请参考**计费项**。

工作流节点介绍

AgentArts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，根据功能分为基础节点、通用节点、逻辑节点、工具节点、消息管理节点和变量&知识节点，具体节点功能详见**表4-1**。

表 4-1 配置节点

分类	节点名称	节点说明
基础节点	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，详细配置参见**开始节点**。
	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，详细配置参见**结束节点**。

版权所有 © 华为云计算技术有限公司

分类	节点名称	节点说明
通用节点	大模型节点	用于在工作流中引入大模型能力，详细配置参见**大模型**。
	工作流节点	实现工作流嵌套工作流的效果，详细配置参见**工作流**。
逻辑节点	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，详细配置参见**判断**。
	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，详细配置参见**意图识别**。
	高级意图识别节点	用于根据用户大量可归类的输入进行意图分类并导向后续不同的处理流程。适用于编排20个以上意图的分支逻辑。详细配置参见**高级意图识别**。
	代码节点	用于引入代码执行器，根据节点的输入，执行Python代码，节点的输出是代码执行的结果信息，详细配置参见**代码**。
	循环节点	循环节点用于重复执行一系列任务，详细配置参见**循环**。
工具节点	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，详细配置参见**插件**。
	MCP服务节点	支持从MCP服务中选择您已配置好的或预置MCP服务，并选择所需的工具完成调用，详细配置参见**MCP服务**。
	HTTP请求节点	允许用户通过HTTP协议发送请求到外部服务，实现数据的获取、提交和交互。支持多种http请求方法，并允许用户配置请求参数、请求头、鉴权信息、请求体等，以满足不同的数据交互需求，详细配置请参见**HTTP请求**。
消息管理节点	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，详细配置参见**消息**。
	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，详细配置参见**提问器**。
	输入节点	输入节点用于在工作流运行时收集用户输入，详细配置参见**输入**。
	问答节点	问答节点可提供中间过程的向用户提问的能力，详细配置参见**问答**。
	异常节点	异常节点允许用户根据业务需求灵活设置和抛出详细的异常信息，详细配置参见**异常**。
变量&知识节点	变量赋值节点	变量赋值节点用于在循环执行过程中动态设置中间变量，详细配置参见**变量赋值**。

版权所有 © 华为云计算技术有限公司

分类	节点名称	节点说明
	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，详细配置参见**变量聚合**。
	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，详细配置参见**知识检索**。

配置方式

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，根据功能需要使用节点进行设计。

4.2 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：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。

费用说明

任务型工作流常用于批量处理，运行过程中会产生大模型调用费用。进行批量数据处理前，建议先用小样本进行试运行以估算成本。

应用限制

任务型工作流旨在实现端到端的自动化处理，任何需要等待用户中间输入或交互的节点（如提问器、消息节点）都会中断自动化流程，因此任务型工作流不支持配置**输入节点、消息节点、提问器节点、问答节点**。

对话型工作流和任务型工作流差异

表 4-2 对话型工作流和任务型工作流区别说明

差异项	对话型工作流	任务型工作流
适用场景	AI客服助手、虚拟助手、工单助手、娱乐互动等多轮交互的场景。	数据处理、批量生成、自动化报告、批量翻译、数据分析等场景。

版权所有 © 华为云计算技术有限公司

差异项	对话型工作流	任务型工作流
节点	支持输入节点、消息节点、提问器节点、问答节点	不支持输入节点、消息节点、提问器节点、问答节点。
试运行方式	试运行界面与任务型工作流不同。 如果“开始”节点有多个参数，query参数为默认参数，可先对除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。	如果“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。

相关文档

对话型工作流、任务型工作流是否可以相互转换？请参考**对话型工作流、任务型工作流是否可以相互转换？**。

4.3 工作流使用限制

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。在使用过程中注意以下限制。

工作流使用限制

表 4-3 工作流使用限制

限制	说明
超时时间	工作流超时时间15分钟，插件超时时间50秒， 模型超时时间15分钟，其他单节点无限制。
运行次数	循环节点自身最大循环次数为1000次。
节点总数	工作流中非游离节点个数最多为150个。
请求大小	请求查询大小上限为100000字符。

4.4 工作流搭建范式

4.4.1 工作流搭建范式简介

工作流的搭建范式，是指在AgentArts工作流编排中，节点之间的连接与协作方式所形成的固定模式。不同的范式决定了数据如何在节点间流转、任务如何被拆解执行、以及最终结果如何汇聚输出。

理解搭建范式，有助于在面对复杂业务场景时，能够快速选择合适的编排结构，而不是凭直觉随意连线。AgentArts工作流编辑器提供了开始节点、结束节点、大模型节

版权所有 © 华为云计算技术有限公司

点、判断节点、循环节点、意图识别节点、插件节点、工作流节点、变量聚合节点等丰富的节点类型。这些节点经过组合可以形成多种搭建范式：

表 4-4 工作流搭建范式

范式	核心节点	说明
顺序链式执行	大模型节点、插件节点、代码节点	节点首尾相连，形成一条有向执行链。前一节点的输出是后一节点的输入，任务被拆解为若干顺序步骤，逐步加工完成。
条件路由执行	判断节点、意图识别节点	工作流在运行过程中遇到分支，根据结构化条件或用户自然语言意图，将执行流程导向不同的处理分支，实现一个入口、多条路径。
并行聚合执行	判断节点 + 变量聚合节点	多个分支各自独立处理，最终由变量聚合节点收口，将不同分支的输出合并为统一变量，传递给下游节点。
循环迭代执行	循环节点、变量赋值节点	一组节点被重复执行多次，每次处理数组中的一个元素或完成一轮任务，适合批量处理与多轮迭代场景。
嵌套编排执行	工作流节点（子工作流）	将已发布的工作流作为子程序嵌入主流程，主流程负责总体调度，子流程专注单一功能，实现复杂任务的模块化拆分与复用。
反思精炼循环	循环节点 + 大模型节点 + 判断节点	生成结果后引入评估环节，若结果未达标则触发修订，形成“生成→评估→修订”的闭环，持续提升输出质量。

以上范式并非互斥，现实中的工作流往往是多种范式的组合。本章将逐一介绍每种范式的定义、工作原理，并提供可直接上手操作的示例。

面对一个工作流搭建需求，可以参考以下思路快速定位所需的范式：

```
任务是否需要重复处理多个数据项?
    └─ 是 → 是否需要在多轮处理中评估并改进输出质量?
            ├─ 是 → 【反思精炼循环】
            └─ 否 → 【循环迭代执行】

任务是否需要根据不同输入走不同路径?
    └─ 是 → 多个分支的结果是否需要汇聚给同一个下游节点?
            ├─ 是 → 【并行聚合执行】（条件路由 + 变量聚合）
            └─ 否 → 【条件路由执行】

流程是否包含可独立封装、未来可复用的功能模块?
    └─ 是 → 【嵌套编排执行】

以上都不是，任务可以被拆解为若干固定顺序步骤?
    └─ 是 → 【顺序链式执行】
```


版权所有 © 华为云计算技术有限公司

表 4-5 常见范式组合

组合方式	典型场景
顺序链 + 条件路由	先完成数据预处理，再根据处理结果路由到不同输出逻辑。
条件路由 + 并行聚合	多分支处理后用变量聚合统一结果，再交给下游节点处理。
嵌套编排 + 顺序链	主流程调用多个子流程，每个子流程内部是一条顺序链。
循环迭代 + 反思精炼	在循环体内嵌入"生成→评估→修订"的精炼逻辑。

4.4.2 顺序链式执行

简介

顺序链式执行是最基础的工作流结构。节点之间形成一条有向的执行链，前一节点的输出作为后一节点的输入，逐步完成一个完整任务。

这种范式适用于**任务可以被自然地拆解为若干顺序步骤的场景**：每一步的结果是下一步的前提，整体流程方向固定，不存在分叉或循环。

相比将所有事情塞进一个大模型节点的单步提示，顺序链式执行的优势在于：

- 每步职责单一，中间结果可审计。
- 不同步骤可选择最合适的节点类型（不一定都是大模型）。
- 中间结果可以被后续多个节点复用。

原理

```
[开始节点] → [节点A] → [节点B] → [节点C] → [结束节点]
```

数据流转规则

- 每个节点通过引用上游节点的输出参数获取输入。
- 节点按照连线顺序同步串行执行，前一节点完成才触发下一节点。
- 任意中间节点的输出，都可以被链路后方的任意节点引用。

常用的链式组合模式

表 4-6 工作流链路结构示例

链路结构	典型用途
开始 → 大模型 → 结束	简单问答、内容生成
开始 → 知识检索 → 大模型 → 结束	RAG知识库问答
开始 → 插件/MCP → 大模型 → 结束	实时数据获取后智能分析

版权所有 © 华为云计算技术有限公司

链路结构	典型用途
开始 → 大模型A → 大模型B → 结束	多步推理（提取 → 分析 → 输出）
开始 → 插件/MCP → 代码 → 大模型 → 结束	数据处理后智能生成

操作示例：热点话题生成工作流

场景描述：用户输入一个感兴趣的话题关键词，工作流通过智谱搜索插件实时搜索该话题的最新资讯，再由大模型将搜索结果提炼为一份结构清晰、适合快速阅读的简报。

这个示例用户只需输入一个关键词即可运行，插件节点解决了大模型不知道最新信息的核心痛点，效果直观。

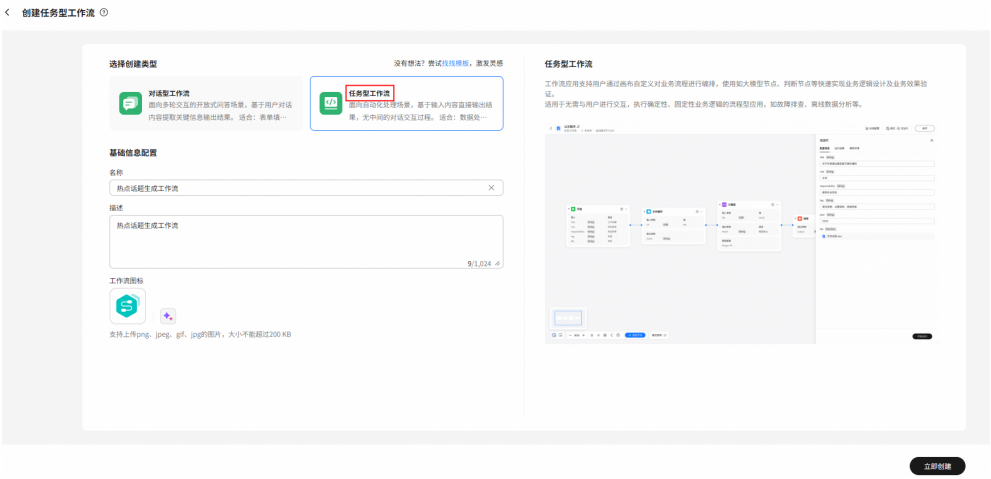
工作流结构：

```
[开始节点]
    ↓ keyword（话题关键词）
[插件节点：智谱搜索]
    ↓ search_results（搜索结果原文）
[大模型节点：提炼生成简报]
    ↓ brief（简报正文）
[结束节点]
```

创建工作流：

在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-1 任务型工作流



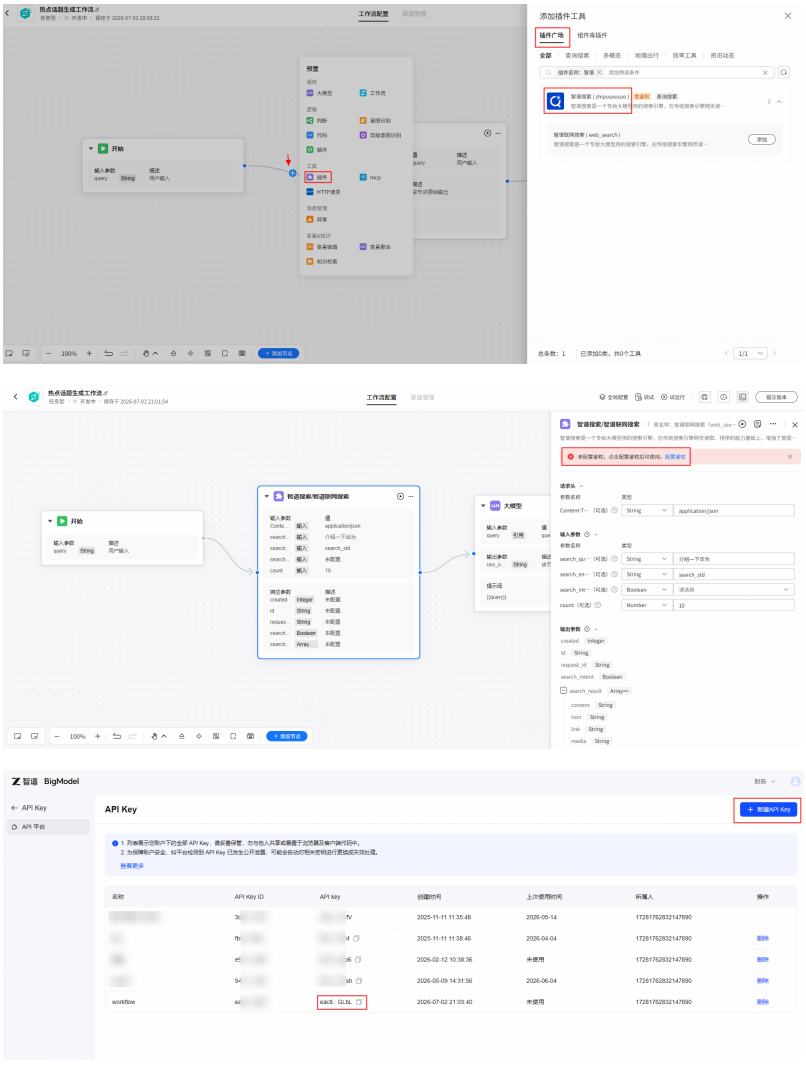
节点配置说明：

版权所有 © 华为云计算技术有限公司

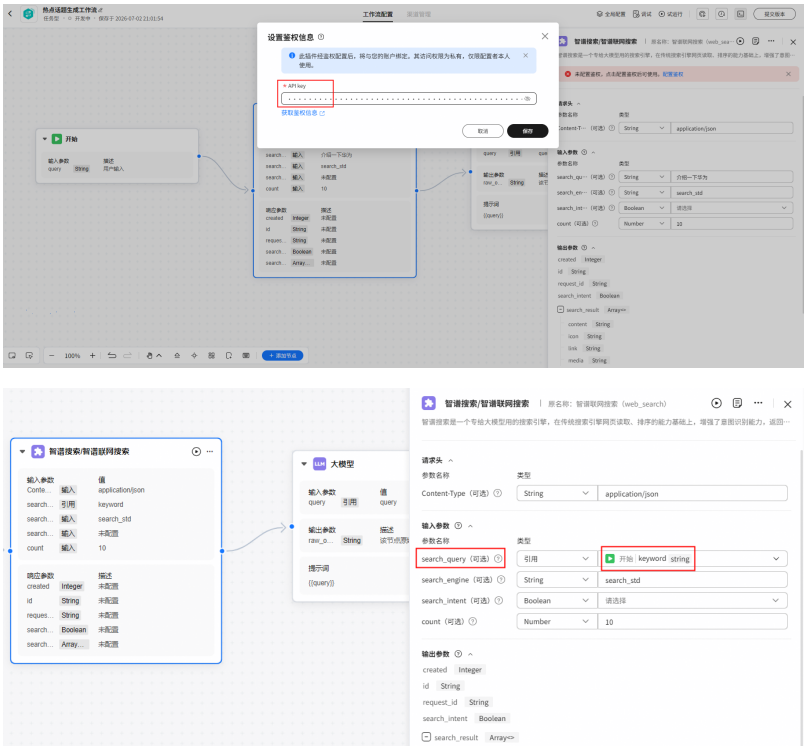
表 4-7 工作流各节点配置说明

节点	配置说明
开始节点	原有基础上新增一个参数。 ● 参数名称：keyword ● 参数类型：String ● 描述：想了解的话题关键词，例如“AI大模型”、“新能源汽车”、“世界杯” ● 必填：勾选 

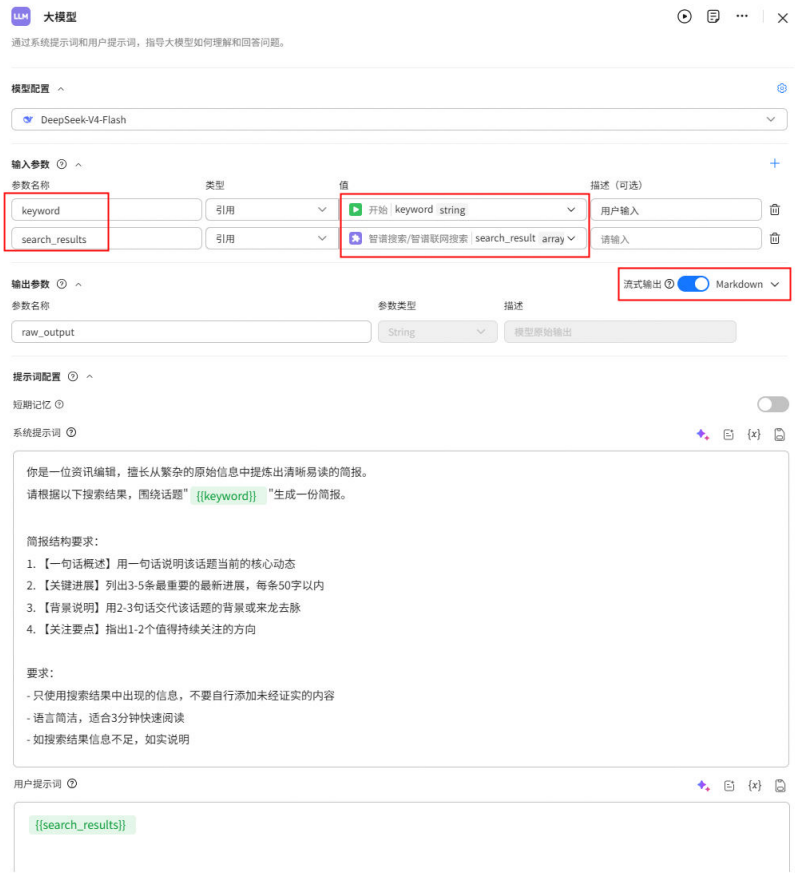
版权所有 © 华为云计算技术有限公司

节点	配置说明
插件节点	使用智谱搜索插件。 如果此前未安装智谱搜索插件，单击“添加节点 > 插件 > 插件广场”，按页面提示创建智谱搜索插件。智谱搜索插件需要进行鉴权，添加完成后，单击插件按照页面提示进行鉴权。 安装完成后配置： ● 工具：选择智谱搜索插件中的搜索工具。 ● **输入参数search_query：引用开始节点的keyword。** 输出参数由插件元信息自动导入，包含搜索返回的结果文本。 

版权所有 © 华为云计算技术有限公司

节点	配置说明
	

版权所有 © 华为云计算技术有限公司

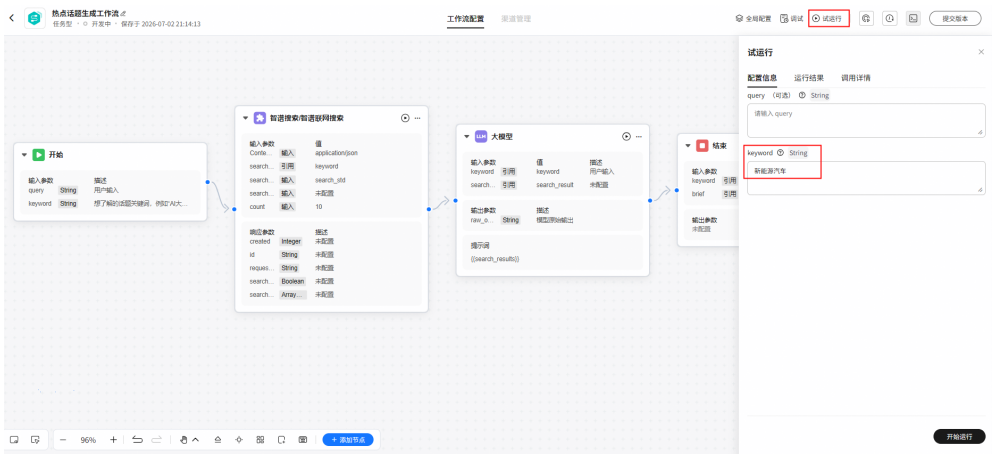
节点	配置说明
大模型节点	用于提炼内容生成简报。 • 模型：DeepSeek-V4-Flash • 输出格式：Markdown，开启流式输出 • 输入参数： – keyword：引用开始节点的keyword – search_results：引用智谱搜索节点的search_results • 系统提示词： 你是一位资讯编辑，擅长从繁杂的原始信息中提炼出清晰易读的简报。 请根据以下搜索结果，围绕话题"{{keyword}}"生成一份简报。 简报结构要求： 1.【一句话概述】用一句话说明该话题当前的核心动态 2.【关键进展】列出3-5条最重要的最新进展，每条50字以内 3.【背景说明】用2-3句话交代该话题的背景或来龙去脉 4.【关注要点】指出1-2个值得持续关注的方向 要求： - 只使用搜索结果中出现的信息，不要自行添加未经证实的内容 - 语言简洁，适合3分钟快速阅读 - 如搜索结果信息不足，如实说明 • 用户提示词：搜索结果原文：{{search_results}} 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： – keyword：引用开始节点的 keyword – brief：引用大模型节点的 raw_output • 指定回复： {{keyword}}·快速简报 {{brief}} 

试运行：

在开始节点的keyword输入“新能源汽车”，工作流将实时搜索该话题的最新资讯，大模型提炼后输出结构化简报。换一个关键词再跑一次，可以对比不同话题的简报效果。



4.4.3 条件路由执行

介绍

条件路由执行是指工作流在运行过程中，根据某个条件的判断结果，**将执行流程导向不同的处理分支**。

版权所有 © 华为云计算技术有限公司

这种范式解决了“不同输入需要不同处理逻辑”的问题：同一个工作流入口，根据运行时的实际情况，流向完全不同的执行路径，每条路径针对其场景做最合适的处理。

AgentArts提供了两种路由节点，分别适用于不同的判断场景：

表 4-8 路由节点

节点	判断依据	适用场景
判断节点	结构化变量（数字大小、字符串是否相等、布尔值等）	评分高低分档、状态码路由、数值范围判断
意图识别节点	用户自然语言的语义意图	多功能助手的能力分发、用户诉求分类

原理

```
                ┌──→ [分支A节点链]
[上游节点] → [路由节点] ──→ [分支B节点链]
                └──→ [兜底分支]
```

路由节点的核心逻辑：

- **判断节点：**依次检查 IF → ELSE IF → ELSE，命中第一个满足条件的分支即执行，后续分支不再检查；条件越具体越应排在越前面。
- **意图识别节点：**大模型对用户输入进行语义分类，将执行流导向匹配的意图分支，未命中任何意图时进入“其他意图”兜底分支。

每个分支是互斥的：同一次执行只有一条分支会运行。务必设置兜底分支，避免未覆盖的输入导致流程无法继续。

操作示例：作文智能点评工作流

场景描述： 用户输入一篇作文，工作流按照作文评分标准进行评分。大模型节点分别评定基础等级分（满分40分）和发展等级分（满分20分），并计算错别字、字数不足等扣分项，代码节点汇总计算最终得分（满分60分），判断节点按得分档次路由到对应的点评策略，给出最有针对性的反馈意见。

用户只需输入一篇作文即可运行，内容长短不限，几段话也能评分。

工作流结构：

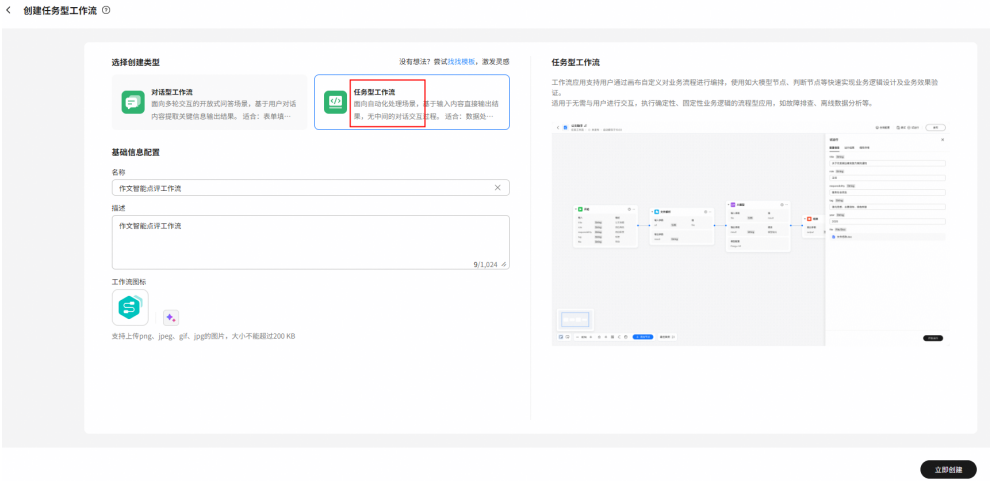
```
[开始节点]
   ↓ essay（作文内容）
[大模型节点：按高考标准评分]
   ↓ basic_score（基础等级分0-40）、dev_score（发展等级分0-20）
     deduction（扣分项合计）、score_detail（各维度评分说明）
[代码节点：汇总计算最终得分]
   ↓ total_score（最终得分）、score_breakdown（得分构成说明）
[判断节点：按档次路由]
   ├──→ IF: total_score ≥ 50  → [大模型节点A：一类卷点评]
   ├──→ ELSE IF: total_score ≥ 35 → [大模型节点B：二类卷点评]
   ├──→ ELSE IF: total_score ≥ 20 → [大模型节点C：三类卷点评]
   └──→ ELSE: total_score < 20 → [大模型节点D：四类卷点评]
   ↓
[变量聚合节点：汇聚四条分支输出]
   ↓ Group1（统一点评内容）
[结束节点]
```

创建工作流：

 版权所有 © 华为云计算技术有限公司

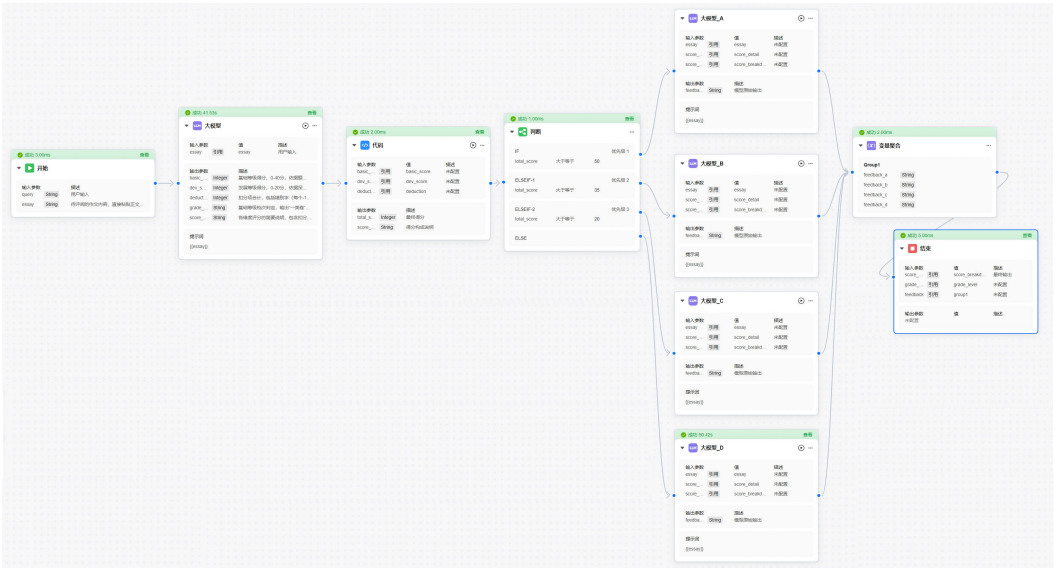
在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-2 任务型工作流



节点配置说明：

图 4-3 作文智能点评工作流



版权所有 © 华为云计算技术有限公司

表 4-9 工作流各节点配置说明

节点	配置说明
开始节点	原有基础上新增一个参数。 ● 参数名称：essay ● 参数类型：String ● 描述：待评阅的作文内容，直接粘贴正文即可，建议包含标题 ● 必填：勾选 

版权所有 © 华为云计算技术有限公司

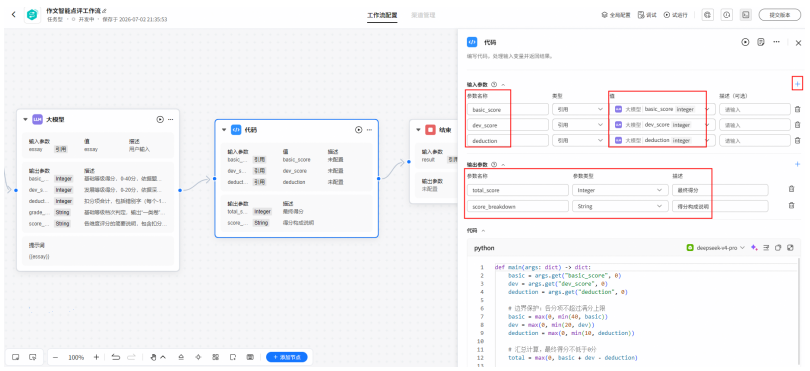
节点	配置说明
大模型节点	用于作文评分。 ● 模型：DeepSeek-V4-Flash ● 输入参数：essay，引用开始节点的 essay ● 输出格式：JSON，关闭流式输出 ● 输出参数定义： – basic_score，类型Integer，基础等级得分，0-40分，依据题意、内容、语言、文体四项综合评定 – dev_score，类型Integer，发展等级得分，0-20分，依据深刻、丰富、有文采、有创意四项累计评定 – deduction，类型Integer，扣分项合计，包括错别字（每个-1分最多-5分）、字数不足、无标题等 – grade_level，类型String，基础等级档次判定，输出“一类卷”、“二类卷”、“三类卷”或“四类卷”之一 – score_detail，类型String，各维度评分的简要说明，包含扣分原因，200字以内 ● 系统提示词： 你是一位语文阅卷专家，请严格按照以下作文评分标准对作文进行评分。 【基础等级 满分40分】 一类卷（32-40分）：符合题意，中心突出，内容充实，思想健康，感情真挚； 文体规范，结构严谨，语言流畅，字迹工整。 二类卷（26-31分）：符合题意，中心明确，内容较充实，思想健康，感情真实； 文体规范，结构完整，语言通顺，字迹清楚。 三类卷（20-25分）：基本符合题意，中心基本明确，内容单薄，思想基本健康； 文体基本规范，结构基本完整，语言基本通顺，字迹较清楚。 四类卷（0-19分）：偏离题意，中心不明或立意不当，思想不健康； 文体不规范，结构混乱，语言不通顺，字迹难辨。 【发展等级 满分20分】 深刻（0-5分）：透过现象看本质，揭示事物内在联系，观点具有启发性。 丰富（0-5分）：材料丰富，论据充实，形象丰满，意境深远。 有文采（0-5分）：词语生动，句式灵活，善于运用修辞手法，文句有意蕴。 有创意（0-5分）：见解新颖，材料新鲜，构思精巧，推理想象有独到之处。 【扣分项】 错别字：每个错别字扣1分，重复不计，最多扣5分。 字数不足：少于800字每少50字扣1分最多扣3分；不足400字按低分处理（basic_score不超过19分）。 无标题：扣2分（计入deduction）。 违规：抄袭、严重偏题等，basic_score不超过20分，dev_score不超过5分。 【评分说明】 - basic_score 和 dev_score 分开评定，deduction 单独统计 - deduction 为正整数，代表需要扣除的总分数 - score_detail 中需说明基础等级判定依据、发展等级各项得分、扣分原因 仅输出原始JSON，不要包含任何Markdown代码块。 ● 用户提示词：{{essay}}

版权所有 © 华为云计算技术有限公司

节点	配置说明
	

 版权所有 © 华为云计算技术有限公司

节点	配置说明

代码节点	用于汇总最终的计算得分。
	• 输入参数：
	– basic_score：引用大模型节点的 basic_score
	– dev_score：引用大模型节点的 dev_score
	– deduction：引用大模型节点的 deduction
	• 输出参数：
	– total_score，类型Integer，最终得分
	– score_breakdown，类型String，得分构成说明
	• 代码内容：
	`def main(args: dict) -> dict:`
	`  basic = args.get("basic_score", 0)`
	`  dev = args.get("dev_score", 0)`
	`  deduction = args.get("deduction", 0)`
	`  # 边界保护：各分项不超过满分上限`
	`  basic = max(0, min(40, basic))`
	`  dev = max(0, min(20, dev))`
	`  deduction = max(0, min(10, deduction))`
	`  # 汇总计算，最终得分不低于0分`
	`  total = max(0, basic + dev - deduction)`
	`  # 生成得分构成说明`
	`  breakdown = (`
	`  f"基础等级：{basic}/40分"`
	`  f"发展等级：{dev}/20分"`
	`  f"扣分项：-{deduction}分"`
	`  f"最终得分：{total}/60分"`
	`  )`
	`  return {`
	`  "total_score": total,`
	`  "score_breakdown": breakdown`
	`  }`
	 

版权所有 © 华为云计算技术有限公司

节点	配置说明
判断节点	引用代码节点的 total_score，按以下顺序配置条件（注意从高到低排列，避免宽泛条件提前拦截）： • IF 条件：代码节点的total_score大于等于 50。用于统计一类卷水平（基础32+以上，发展分较高） • ELSE IF 1 条件：total_score大于等于 35。用于统计二三类卷水平（基础分达标，发展分一般） • ELSE IF 2 条件：total_score大于等于 20。用于统计三四类卷交界（基础分偏低，问题较多） • ELSE，表示其余（total_score < 20）。用于统计四类卷水平（需要重点指导） 
大模型节点A	放在IF分支后面，用于一类卷点评，≥50分 • 模型：DeepSeek-V4-Flash，输出格式：文本，**流式输出** • 输入参数： – essay：引用开始节点的 essay – score_detail：引用大模型评分节点的 score_detail – score_breakdown：引用代码节点的 score_breakdown • 输出参数：feedback_a • 系统提示词： 你是一位高考语文阅卷组长，正在为一篇高水平作文撰写精细点评。 本文评分情况：{{score_breakdown}} 阅卷评语：{{score_detail}} 该文已达到一类卷水平，请从以下角度给出进阶点评： 1.【亮点赏析】指出2-3处最出彩的表达或立意，说明好在哪里 2.【精益求精】找出1-2处仍可进一步打磨的细节，给出具体修改示范 3.【发展等级提升】结合"深刻、丰富、有文采、有创意"四个维度， 指出哪个维度最有潜力继续提升，并给出具体方向 语气：以赏析为主，建议具体可操作，激励考生冲击更高分。 • 用户提示词：{{essay}}

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点B	放在ELSE IF 1分支后面，用于二三类卷点评，35-49分 ● 模型：DeepSeek-V4-Flash，输出格式：文本，**流式输出** ● 输入参数： – essay：引用开始节点的 essay – score_detail：引用大模型评分节点的 score_detail – score_breakdown：引用代码节点的 score_breakdown ● 输出参数：feedback_b ● 系统提示词： 你是一位高考语文阅卷教师，正在为一篇中等水平作文撰写点评。 本文评分情况：{{score_breakdown}} 阅卷评语：{{score_detail}} 该文处于二三类卷水平，请从以下角度给出有针对性的改进意见： 1.【基础等级短板】结合"题意、内容、语言、文体"四项， 指出得分最低的1-2项，说明具体问题在哪里 2.【内容提升建议】指出论据或素材的不足之处， 推荐1-2个可以替换或补充的典型论据方向 3.【语言改进示范】摘出原文中1-2处表达平淡的句子， 给出更生动有力的改写示例 4.【发展等级入门】说明该文在"深刻、丰富、有文采、有创意" 中最容易突破的一项，给出具体建议 语气：客观指出问题，同时给出明确可操作的改进路径。 ● 用户提示词：{{essay}}

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	

 版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点C	放在ELSE IF 2后面，用于三四类卷点评，20-34分 ● 模型：DeepSeek-V4-Flash，输出格式：文本，**流式输出** ● 输入参数： – essay：引用开始节点的 essay – score_detail：引用大模型评分节点的 score_detail – score_breakdown：引用代码节点的 score_breakdown ● 输出参数：feedback_c ● 系统提示词： 你是一位高考语文辅导老师，正在帮助一位作文基础薄弱的学生进行针对性辅导。 本文评分情况：{{score_breakdown}} 阅卷评语：{{score_detail}} 该文处于三类卷水平，存在较明显的基础问题，请从以下角度给出辅导意见： 1.【最紧迫的问题】指出当前得分最低的核心问题（如跑题、结构混乱、语言不通顺）， 用一句话概括，让学生清楚知道首先要解决什么 2.【结构修复建议】分析文章开头、主体、结尾的问题， 给出一个适合本文题目的简单写作框架（三段式或五段式） 3.【语言通顺改写】摘出1处最不通顺的句子，展示如何修改使其通顺表达 4.【扣分项提醒】明确指出哪些扣分项可以通过习惯改正（如检查错别字、字数、标题） 语气：耐心温和，聚焦最核心的问题，避免一次提出太多要求让学生不知所措。 ● 用户提示词：{{essay}}

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	

 版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点D	放在ELSE后面，用于四类卷点评，<20分 ● 模型：DeepSeek-V4-Flash，输出格式：文本，**流式输出** ● 输入参数： – essay：引用开始节点的 essay – score_detail：引用大模型评分节点的 score_detail – score_breakdown：引用代码节点的 score_breakdown ● 输出参数：feedback_d ● 系统提示词： 你是一位有丰富经验的高考语文补习老师，正在帮助一位作文存在严重问题的学生重新起步。 本文评分情况：{{score_breakdown}} 阅卷评语：{{score_detail}} 该文处于四类卷水平，请从最基础的角度给出重建性指导： 1.【核心问题诊断】用1-2句话指出最根本的问题（如：严重偏题/无法成篇/立意有误） 2.【基础框架重建】不针对原文修改，而是给出一个全新的写作框架： - 建议的立意方向（确保审题正确） - 推荐的文章结构（开头应写什么、主体如何展开、结尾如何收束） 3.【可立即执行的改进】列出2-3项最基础、最容易做到的改进动作 （如：先拟好标题、先写好首段点题、每段首句先点中心） 4.【鼓励性总结】用1-2句话给予正向激励，帮助学生建立继续写作的信心 语气：温和坚定，从零开始引导，不批评，只建设。 ● 用户提示词：{{essay}}

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	

 版权所有 © 华为云计算技术有限公司

节点	配置说明
变量聚合节点	设置分组Group1，按以下顺序添加四个变量（均为 String 类型）： • 优先级1：变量来源大模型节点A（一类卷），变量名feedback_a • 优先级2：变量来源大模型节点B（二三类卷），变量名feedback_b • 优先级3：变量来源大模型节点C（三四类卷交界），变量名feedback_c • 优先级4：变量来源大模型节点D（四类卷），变量名feedback_d 

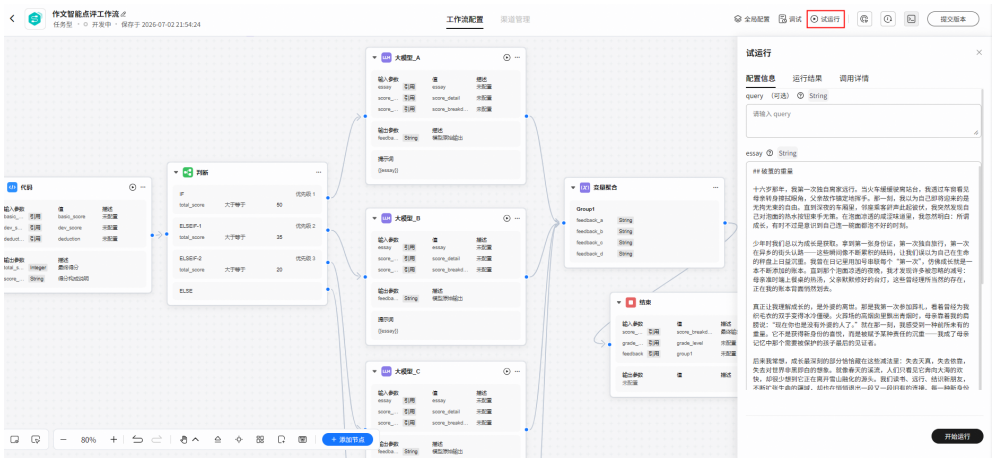
 版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： – score_breakdown：引用代码节点的 score_breakdown – grade_level：引用大模型评分节点的 grade_level – feedback：引用变量聚合节点的 Group1 • 指定回复： 【评分结果】{{score_breakdown}} 【档次判定】{{grade_level}} 【点评意见】 {{feedback}} 

试运行：

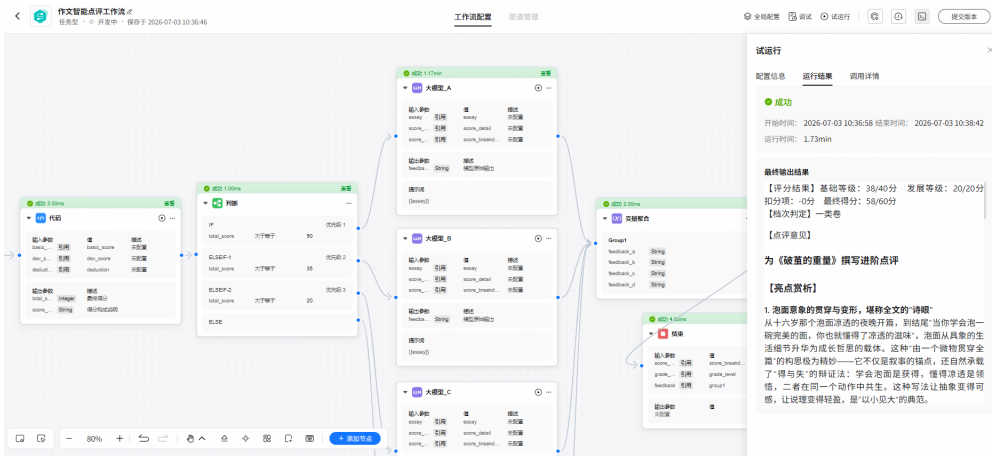
在开始节点的essay参数中输入一篇作文，工作流会先打分，再自动进入对应分段的点评逻辑，输出针对性的反馈。

图 4-4 试运行



版权所有 © 华为云计算技术有限公司

图 4-5 查看运行结果



4.4.4 并行聚合执行

介绍

并行聚合执行是指工作流通过判断节点拆分为多个并行分支，每个分支独立处理，最终由**变量聚合**节点将所有分支的输出合并为统一变量，供下游节点统一引用。

这种范式的核心价值在于：解决多分支执行后下游需要统一接口的问题。在工作流中，判断节点将请求导入不同分支后，同一时刻只有一条分支会执行并产生有效输出，其他分支的输出为空。变量聚合节点按照预设的优先级顺序，自动取出第一个非空值，让下游节点无需关心实际走了哪条分支，始终从同一个变量名获取结果。

原理

```
                ┌──→ [分支A] → output_A (有值) ──┐
[路由节点] ──┼──→ [分支B] → output_B (为空) ──┼──→ [变量聚合节点] → Group1 → [下游节点]
                └──→ [分支C] → output_C (为空) ──┘
                （实际执行了分支A，Group1 = output_A）
```

变量聚合节点的工作机制：

1. 将多个分支的输出变量放入同一个分组，按顺序排列（排在越前面优先级越高）。
2. 运行时依次检查：第一个变量非空则取之，否则检查第二个，以此类推。
3. 将第一个非空值赋给聚合输出变量（如Group1），供下游统一引用。

关键规则：

- 同一分组内的变量类型必须一致（如都是 String，或都是 Integer）。
- 如需聚合多个不同用途的结果，可创建多个分组（Group1、Group2……）。
- 若所有变量均为空，聚合结果也为空，建议确保至少有一条分支能产生输出。

操作示例：多平台文案生成工作流

场景描述： 用户输入一个产品名称和目标发布平台（小红书/微博/微信公众号），工作流根据平台将请求路由到对应风格的文案生成分支，三条分支的输出通过变量聚合统一后，再交给大模型添加 emoji 和话题标签，最终输出完整的发布内容。

这个示例用户只需输入产品名和平台名即可运行，无需任何准备。

版权所有 © 华为云计算技术有限公司

工作流结构：

```
[开始节点]
   ↓ product（产品名）、platform（目标平台）
[判断节点：平台路由]
   ├──→ IF: platform == "小红书" → [大模型节点A：小红书风格] → content_xhs
   ├──→ ELSE IF: platform == "微博" → [大模型节点B：微博风格] → content_wb
   └──→ ELSE（微信公众号）──────→ [大模型节点C：公众号风格] → content_wx
   ↓
[变量聚合节点：统一内容变量]
   ↓ Group1（统一文案）
[大模型节点：添加话题标签]
   ↓
[结束节点]
```

创建工作流：

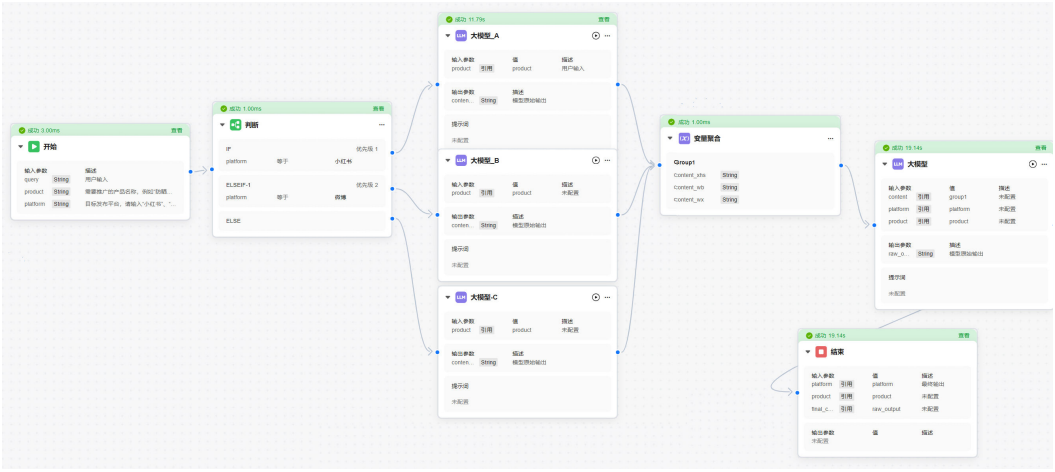
在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-6 任务型工作流



节点配置说明：

图 4-7 多平台文案生成工作流



版权所有 © 华为云计算技术有限公司

表 4-10 工作流各节点配置说明

节点	配置说明
开始节点	新增2个参数。 • 参数一： – 参数名称：product – 参数类型：String – 描述：需要推广的产品名称，例如“防晒霜”、“蓝牙耳机”、“手账本” – 必填：勾选 • 参数二： – 参数名称：platform – 参数类型：String – 描述：目标发布平台，请输入“小红书”、“微博”或“微信公众号”之一 – 必填：勾选 
判断节点	• IF 分支：platform（引用开始节点）等于“小红书” • ELSE IF 1 分支：platform等于“微博” • ELSE 分支：其余情况（视为微信公众号） 

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点A	放在IF分支后面，生成**小红书风格**内容 • 模型：DeepSeek-V4-Flash，输出格式：文本，开启流式输出 • 输入参数：product，引用开始节点的 product • 输出参数：命名为content_xhs • 系统提示词： 你是一位小红书爆款内容创作者。 请为产品"{{product}}"写一篇小红书种草笔记。 风格要求： - 开头用一句吸引眼球的问句或感叹句 - 正文分3-4个小点，每点用简短标题 - 语气亲切、真实，像闺蜜推荐一样 - 字数200字左右 直接输出笔记正文内容。 • 用户提示词：不填，删除原有内容 

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点B	放在ELSE IF 1分支后面，生成**微博风格**内容 • 模型：DeepSeek-V4-Flash，输出格式：文本，开启流式输出 • 输入参数：product，引用开始节点的 product • 输出参数：命名为content_wb • 系统提示词： 你是一位擅长写微博的内容运营。 请为产品"{{product}}"写一条微博推广文案。 风格要求： - 语言简洁有力，适合碎片化阅读 - 突出1-2个核心卖点 - 结尾留有互动引导（如：你们觉得呢？） - 字数控制在140字以内 直接输出微博正文内容。 • 用户提示词：不填，删除原有内容 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点C	放在ELSE分支后面，生成**微信公众号风格**内容 • 模型：DeepSeek-V4-Flash，输出格式：文本，开启流式输出 • 输入参数：product，引用开始节点的 product • 输出参数：命名为content_wx • 系统提示词： 你是一位微信公众号编辑。 请为产品"{{product}}"写一段公众号推文的核心段落。 风格要求： - 语言专业、有信任感 - 结构清晰，包含场景引入和产品优势说明 - 字数300字左右 直接输出段落正文内容。 • 用户提示词：不填，删除原有内容 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
变量聚合节点	设置分组 Group1，按以下顺序添加三个变量（均为 String 类型）： • 优先级1：变量来源大模型节点A（小红书），变量名content_xhs • 优先级2：变量来源大模型节点B（微博），变量名content_wb • 优先级3：变量来源大模型节点C（公众号），变量名content_wx 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点	放在变量聚合节点后面。 • 模型：DeepSeek-V4-Flash，输出格式：文本，开启流式输出 • 输入参数： – content：引用变量聚合节点的 Group1 – platform：引用开始节点的 platform – product：引用开始节点的 product • 系统提示词： 你是一位社媒内容优化专家。 请在以下{{platform}}文案的末尾，添加3-5个适合该平台的热门话题标签，以及2-3个相关emoji装饰。 要求：标签自然融入，emoji点缀在文案中适当位置（不要全堆在结尾）。 文案内容：{{content}} 直接输出优化后的完整内容，不要添加任何说明前缀。 • 用户提示词：不填，删除原有内容 

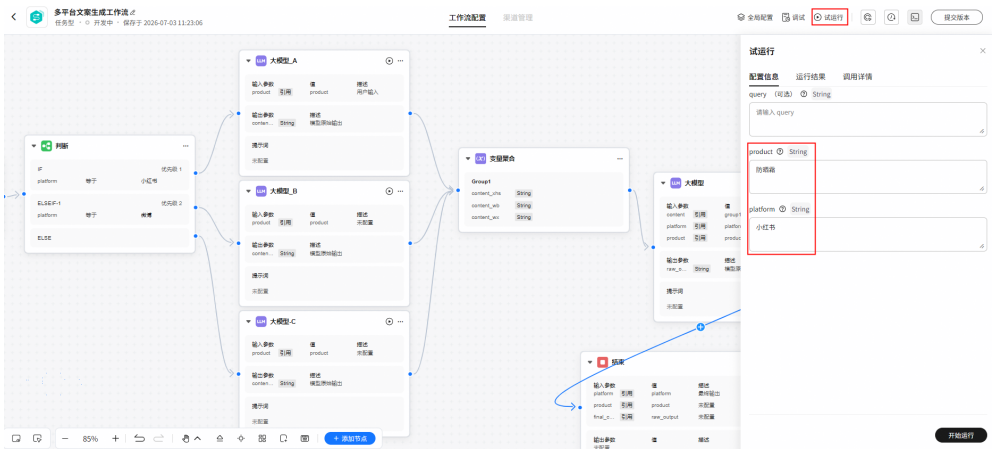
版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： – platform：引用开始节点的 platform – product：引用开始节点的 product – final_content：引用大模型节点的 raw_output • 指定回复： 【{{platform}} · {{product}} 推广文案】 {{final_content}} 

试运行：

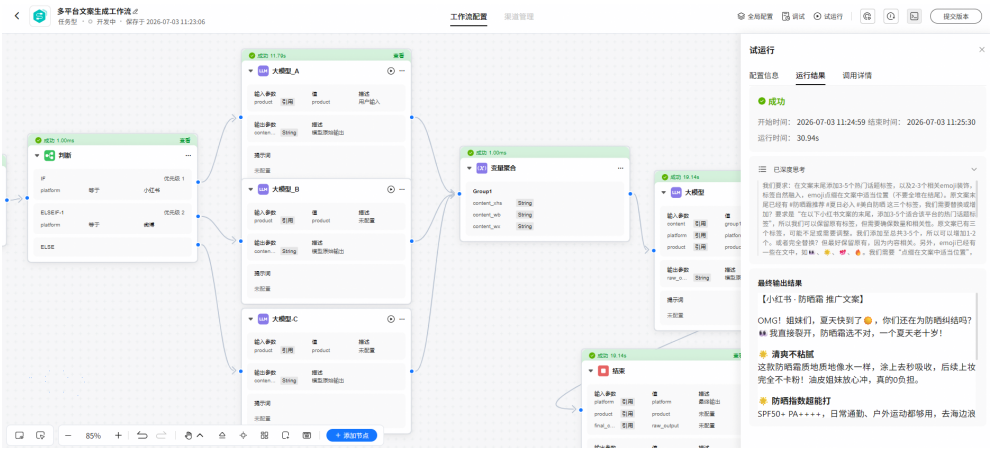
开始节点的product输入“防晒霜”，platform输入“小红书”，工作流会路由到小红书分支生成种草风格文案，经变量聚合后再由大模型节点添加标签和emoji，输出完整的可发布内容。将 platform 改为“微博”后重新运行，可以对比不同平台的风格差异。

图 4-8 试运行



版权所有 © 华为云计算技术有限公司

图 4-9 查看运行结果



4.4.5 循环迭代执行

介绍

循环迭代执行是指工作流中的某一组节点被**重复执行多次**，每次迭代处理数组中的一个元素或完成一轮任务，直到遍历完所有数据或达到指定次数为止。

这种范式解决了**批量处理问题**：当需要对一组数据逐条执行相同操作，或需要某个任务重复运行 N 轮时，循环迭代能够替代人工逐一触发，实现自动化批量处理。

AgentArts 的循环节点支持两种模式：

- **数组循环**：遍历一个已知数组，数组有多少元素就循环多少次，常用于批量处理场景。
- **指定次数循环**：固定执行 N 次，常用于多轮迭代优化场景。

原理

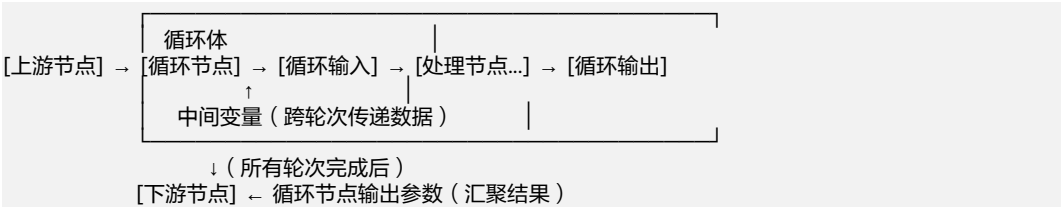


表 4-11 关键概念

概念	说明
item	数组循环中当前轮次处理的数组元素。
index	当前轮次的序号（从0开始）。
中间变量	跨轮次持久存在的变量，每轮循环后可通过变量赋值节点更新其值。
输出参数	循环结束后，将指定的最终结果传递给下游节点。

版权所有 © 华为云计算技术有限公司

概念	说明
终止条件	每轮循环体执行完后检查，满足则提前退出循环。

操作示例：多平台文案批量生成工作流

场景描述： 用户输入一款产品的名称，工作流自动为小红书、微信朋友圈、微博三个平台分别生成针对性的营销文案，每轮循环处理一个平台，将所有文案逐步拼接，最终输出一份完整的多平台文案合集。

这个示例用户只需输入一个产品名称即可运行，效果直观，便于验证循环执行的过程。

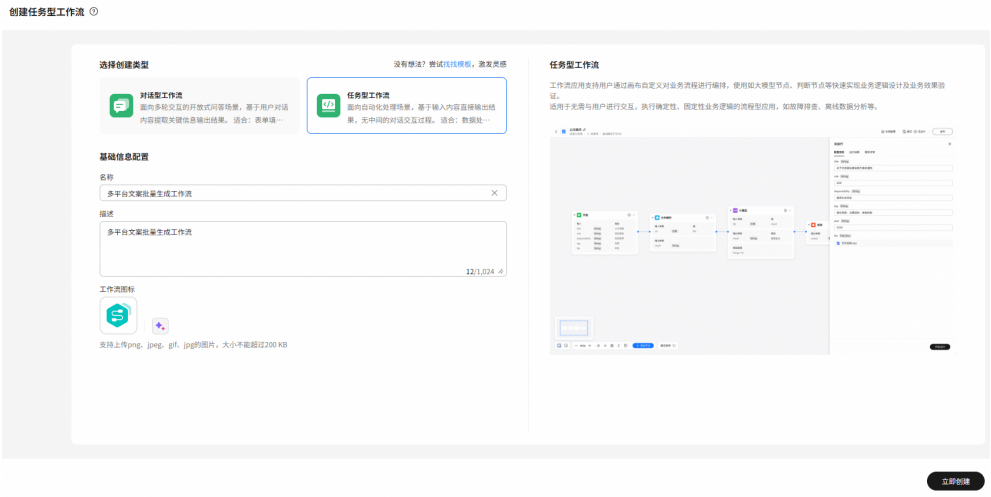
工作流结构：

```
[开始节点]
   ↓ product（产品名）、platforms（平台数组）
[循环节点（数组循环，遍历 platforms）]
   ├── 中间变量：all_copies（汇总文案，初始为空字符串）
   ├── 循环体：
   │    [大模型节点：生成当前平台文案] → copy（String）
   │    [代码节点：拼接到汇总] → merged（String）
   │    [变量赋值节点：更新 all_copies]
   │    ↓（连接循环输出）
   └── 输出参数：output（引用中间变量 all_copies）
   ↓
[结束节点]
```

创建工作流：

在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-10 任务型工作流



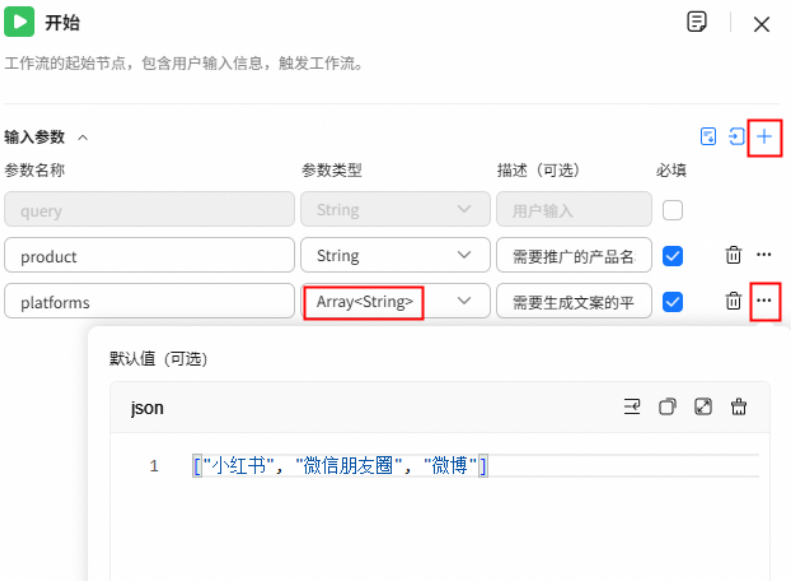
节点配置说明：

版权所有 © 华为云计算技术有限公司

图 4-11 多平台文案批量生成工作流



表 4-12 工作流各节点配置说明

节点	配置说明
开始节点	新增2个参数。 • 参数一： – 参数名称：product – 参数类型：String – 描述：需要推广的产品名称，例如降噪耳机、保温杯 – 必填：勾选 • 参数二： – 参数名称：platforms – 参数类型：Array<String> – 描述：需要生成文案的平台列表 – 默认值：["小红书", "微信朋友圈", "微博"] – 必填：勾选 

版权所有 © 华为云计算技术有限公司

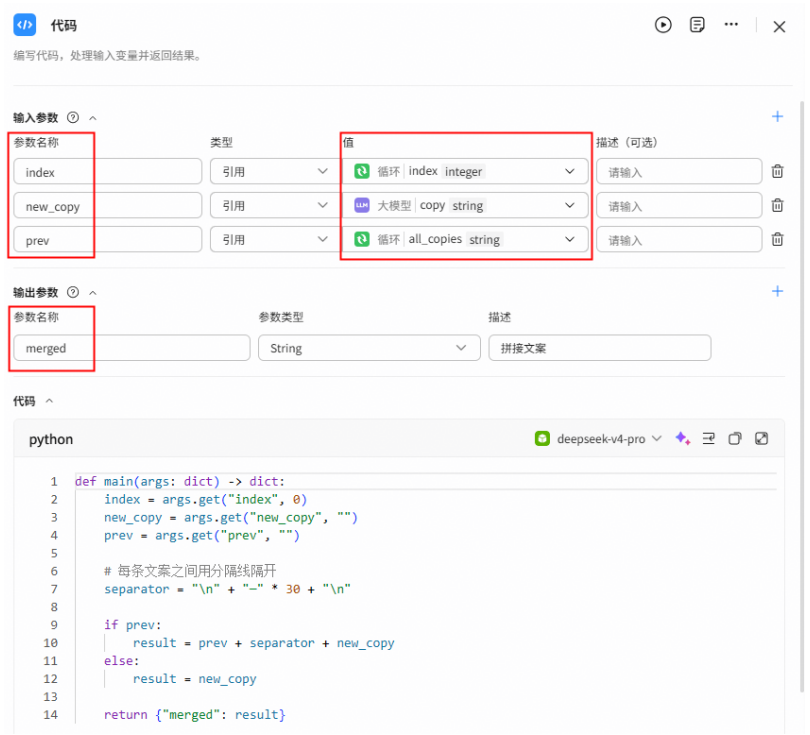
节点	配置说明
循环节点	• 循环类型：**使用数组循环** • 循环数组 arr_loop_var：引用开始节点的 platforms • 中间变量： – 名称：all_copies，初始值：输入固定值""（空字符串） • 输出参数： – 名称：output，引用中间变量 all_copies 

 版权所有 © 华为云计算技术有限公司

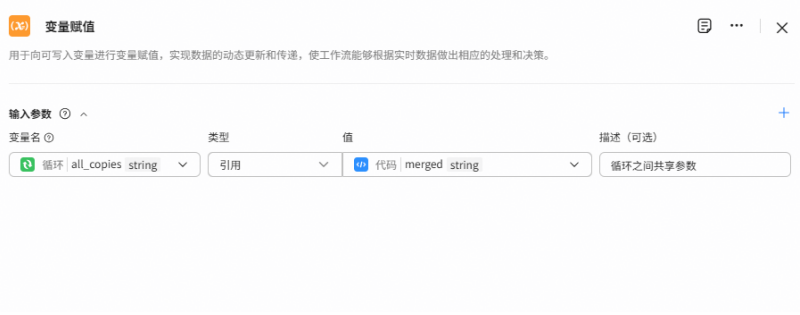
节点	配置说明
	循环体-大模型节点 此节点放置于循环节点内部，连接循环输入 → 此节点。 • 模型：DeepSeek-V4-Flash，将生成随机性调高至 0.8（确保每轮生成不同风格） • 输入参数： – product：引用开始节点的 product – platform：引用循环节点的 item（当前轮次处理的平台名称） • 输出格式：JSON，关闭流式输出 • 输出参数：copy（String，描述：针对该平台风格生成的营销文案正文） • 系统提示词： 你是一位资深百万粉自媒体爆款操盘手。 我的新产品是一款{{product}}。 请针对{{platform}}平台的受众偏好和语言风格，写一篇吸引人的种草文案。 需要按照以下格式输出： {{platform}}风格文案：[文案内容] 仅输出原始JSON，严格按照示例：{"copy": "xxx风格文案：[内容]"} • 用户提示词：不填，删除原有内容 

版权所有 © 华为云计算技术有限公司

节点	配置说明

	循环体 - 代码节点（拼接文案）
	此节点连接大模型节点 → 此节点。
	• 输入参数：
	– index：引用循环节点的 index（当前轮次序号）
	– new_copy：引用大模型节点的 copy
	– prev：引用循环节点的中间变量 all_copies
	• 输出参数：merged（String，表述为拼接文案）
	代码内容：
	`def main(args: dict) -> dict`
	`  index = args.get("index", 0)`
	`  new_copy = args.get("new_copy", "")`
	`  prev = args.get("prev", "")`
	`  # 每条文案之间用分隔线隔开`
	`  separator = "\n" + "—" * 30 + "\n"`
	`  if prev:`
	`  result = prev + separator + new_copy`
	`  else:`
	`  result = new_copy`
	`  return {"merged": result}`
	 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	循环体 - 变量赋值节点 此节点连接代码节点 → 此节点 → 循环输出。 • 赋值目标：循环节点中间变量 all_copies • 赋值方式：引用 • 值：引用代码节点的 merged 
结束节点	• 输入参数： – product：引用开始节点的 product – copies：引用循环节点的 output • 指定回复： {{product}} 多平台营销文案合集 {{copies}} 

试运行：

在开始节点的product中输入“降噪耳机”，platforms使用默认值["小红书", "微信朋友圈", "微博"]，工作流将循环执行3次，每次生成一个平台的文案，通过代码节点拼接后，输出三平台文案合集。在调试视图中可以清晰地看到3次循环的执行过程。

版权所有 © 华为云计算技术有限公司

图 4-12 试运行

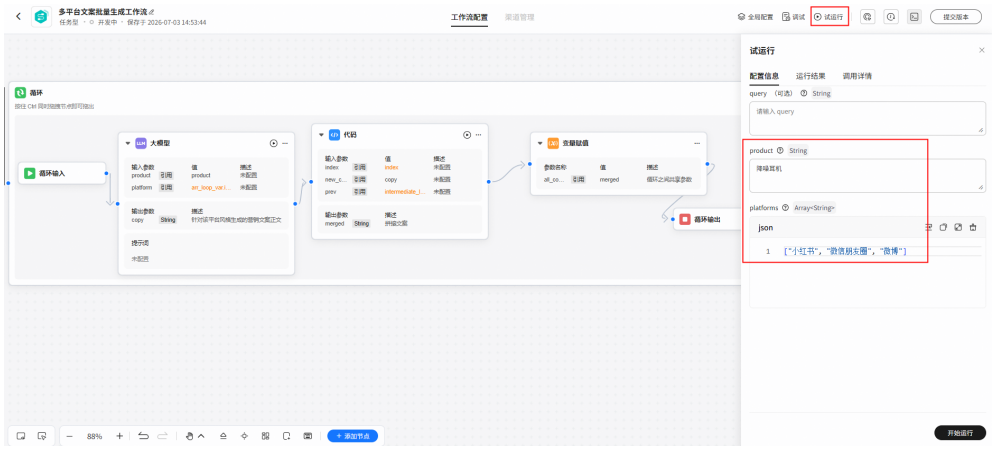
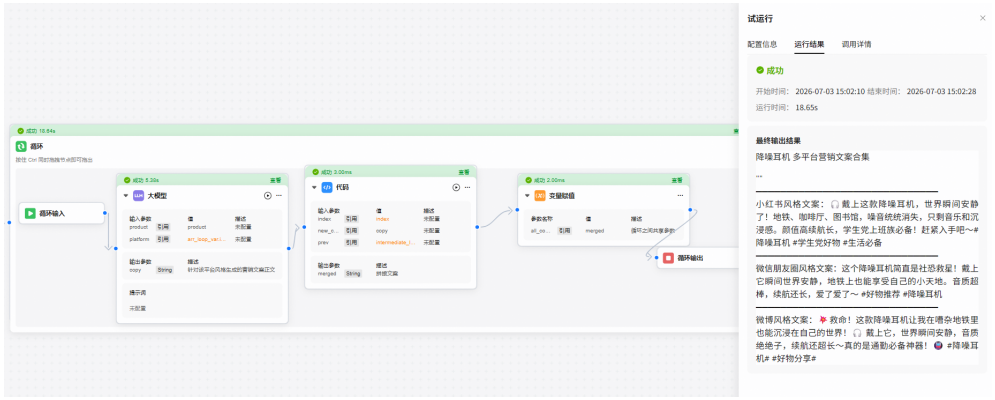


图 4-13 查看运行结果



4.4.6 嵌套编排执行

介绍

嵌套编排执行是指将一个**已发布的工作流作为子程序**，嵌入到另一个主工作流中调用。主工作流充当总调度角色，将特定任务委派给子工作流处理，子工作流完成后将结果返回主流程继续执行。

这种范式解决了大型工作流的模块化管理问题。当业务流程涉及多个相对独立的功能模块，且这些模块未来可能被多个流程复用时，嵌套编排能够：

- 让主工作流保持简洁，只关注总体流程调度。
- 将可复用的逻辑封装为独立子工作流，一处修改处处生效。
- 不同的子工作流可由团队成员分工开发，互不干扰。

原理

```
主工作流：
[开始节点] → [工作流节点A] → [工作流节点B] → [结束节点]
                ↓                ↓
          （调用子流程1）    （调用子流程2）
          [开始]→[...]→[结束] [开始]→[...]→[结束]
              ↓                  ↓
        response_content   response_content
```


版权所有 © 华为云计算技术有限公司

调用机制：

- 工作流节点的输入参数由子工作流的开始节点决定，不可在主流程中修改。
- 工作流节点默认包含 response_content 输出参数，对应子工作流结束节点“指定回复”的内容。
- 子工作流结束节点“输出参数”区域定义的结构化参数，也可在主流程中通过工作流节点引用。
- 主工作流锁定子工作流的特定发布版本，子流程更新后需在主流程中手动刷新升级。

操作示例：文章多维度解读工作流

场景描述：用户粘贴任意一篇文章或新闻，两个子工作流分别独立完成"核心观点提炼"和"金句摘录与点评"，主工作流调用这两个子流程后，将两份结果整合为一份完整的阅读笔记输出。

用户只需复制任意一段文字即可运行——公众号文章、新闻报道、博客长文均可，无需任何前期准备。

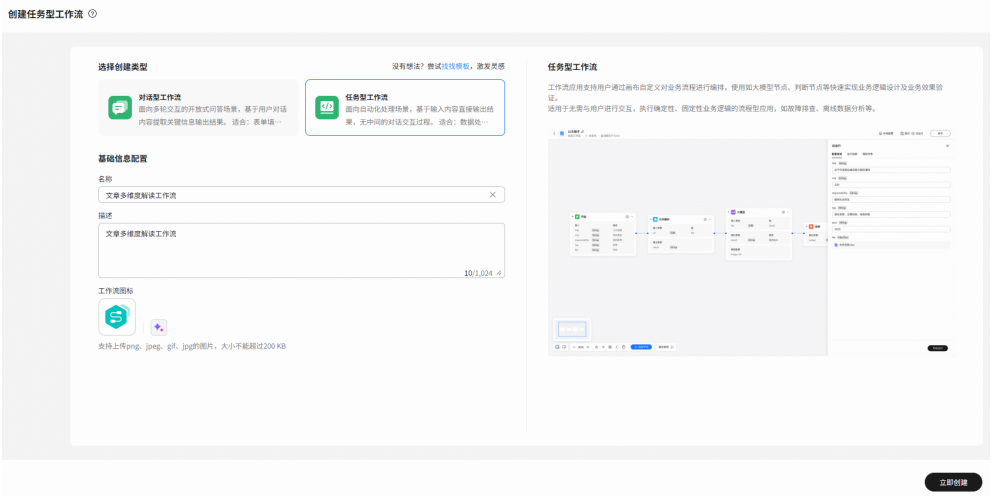
整体架构：

```
主工作流：
[开始节点] → [工作流节点A：核心观点提炼] → [工作流节点B：金句摘录与点评] → [大模型节点：整合阅读笔记] → [结束节点]
                    ↓                                  ↓
          子工作流1内部执行                    子工作流2内部执行
        [开始]→[大模型]→[结束]                [开始]→[大模型]→[结束]
                    ↓                                  ↓
                key_points                          quotes
```

创建工作流：

在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-14 任务型工作流



子工作流1：用于提炼文章的核心观点

创建任务型工作流。

版权所有 © 华为云计算技术有限公司

图 4-15 子工作流 1 提炼文章的核心观点



表 4-13 子工作流 1 配置说明

节点	配置说明
开始节点	新增一个参数。 ● 参数名称：article ● 参数类型：String ● 描述：需要提炼观点的文章原文内容 ● 必填：勾选 

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点	• 模型：DeepSeek-V4-Flash • 输出格式：JSON，关闭流式输出 • 输入参数：article，引用开始节点的 article • 输出参数： – key_points，类型String，描述为“文章的3-5个核心观点，每条观点单独一行，以序号开头”。 – main_theme，类型String，描述为“文章的整体主旨，用一句话概括，30字以内”。 • 系统提示词： 你是一位专业的文章分析师，擅长从文章中提炼核心思想。 请对用户提供的文章进行深度阅读，完成以下两项任务： 任务一：提炼核心观点（key_points） - 从文章中提炼3-5个最重要的核心论点或观点 - 每条观点用一句话表达，准确反映原文意思，不要过度演绎 - 每条观点单独一行，以"1. 2. 3."序号开头 - 每条观点控制在40字以内 任务二：概括整体主旨（main_theme） - 用一句话概括整篇文章的核心主旨 - 控制在30字以内 仅输出原始JSON，不要包含任何Markdown代码块。 • 用户提示词：{{article}}

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	

 版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： – key_points：引用大模型节点的 key_points，String – main_theme：引用大模型节点的 main_theme，String • 指定回复： 文章主旨：{{main_theme}} 核心观点： {{key_points}} 

完成配置后，发布并部署此子工作流，部署时选择API Key认证。

图 4-16 部署子工作流



 版权所有 © 华为云计算技术有限公司

图 4-17 选择 API Key 认证



子工作流2：用于金句摘录与点评

创建任务型工作流。

图 4-18 子工作流 2 金句摘录点评



版权所有 © 华为云计算技术有限公司

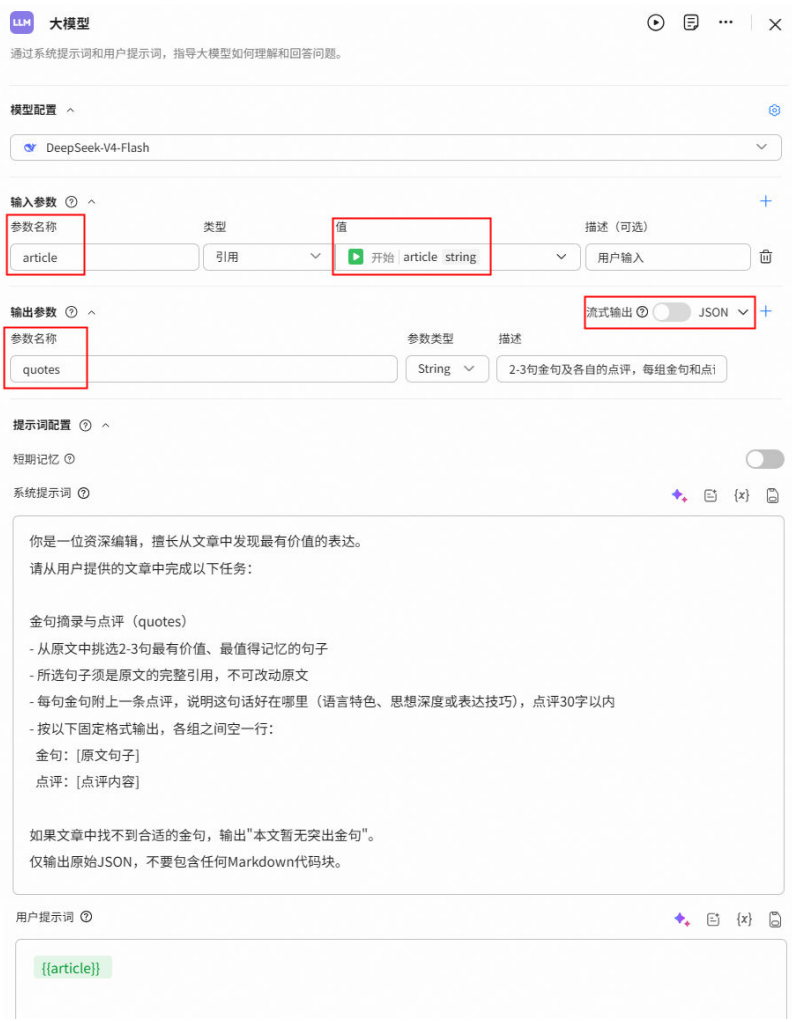
表 4-14 子工作流 2 配置说明

节点	配置说明
开始节点	新增一个参数。 ● 参数名称：article ● 参数类型：String ● 描述：需要摘录金句的文章原文内容 ● 必填：勾选 

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点	• 模型：DeepSeek-V4-Flash • 输出格式：JSON，关闭流式输出 • 输入参数：article，引用开始节点的 article • 输出参数：quotes，类型String，描述为“2-3句金句及各自的点评，每组金句和点评成对出现，格式为"金句：[原文] \n点评：[点评内容]"，各组之间空一行”。 • 系统提示词： 你是一位资深编辑，擅长从文章中发现最有价值的表达。 请从用户提供的文章中完成以下任务： 金句摘录与点评（quotes） - 从原文中挑选2-3句最有价值、最值得记忆的句子 - 所选句子须是原文的完整引用，不可改动原文 - 每句金句附上一条点评，说明这句话好在哪里（语言特色、思想深度或表达技巧），点评30字以内 - 按以下固定格式输出，各组之间空一行： 金句：[原文句子] 点评：[点评内容] 如果文章中找不到合适的金句，输出"本文暂无突出金句"。 仅输出原始JSON，不要包含任何Markdown代码块。 • 用户提示词：{{article}}

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	
结束节点	• 输入参数：quotes：引用大模型节点的 quotes • 指定回复： 金句摘录： {{quotes}} 

 版权所有 © 华为云计算技术有限公司

完成配置后，发布并部署此子工作流，部署时选择API Key认证。

图 4-19 发布子工作流



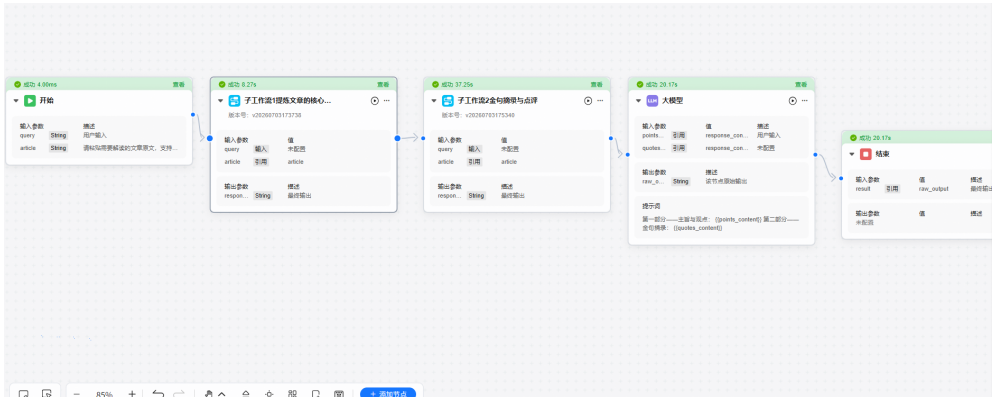
图 4-20 选择 API Key 认证



搭建主工作流

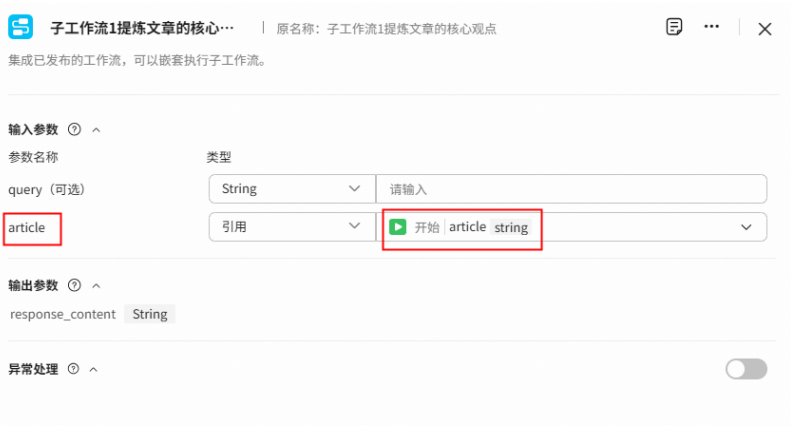
创建任务型工作流。

图 4-21 主工作流



 版权所有 © 华为云计算技术有限公司

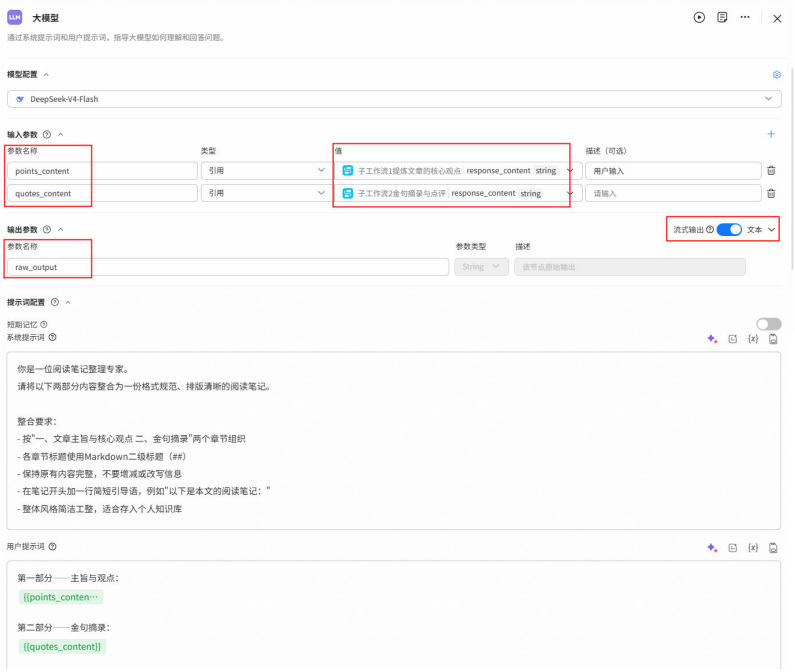
表 4-15 主工作流配置说明

节点	配置说明
开始节点	新增一个参数。 • 参数名称：article • 参数类型：String • 描述：请粘贴需要解读的文章原文，支持新闻、公众号文章、博客长文等任意文本内容 • 必填：勾选 
工作流节点1	选择子工作流1，用于提炼核心观点。 • 选择子工作流1 • 输入参数配置：article：引用主流程开始节点的 article 

版权所有 © 华为云计算技术有限公司

节点	配置说明
工作流节点2	选择子工作流2，用于金句摘录和点评。 ● 选择子工作流2 ● 输入参数配置：article：引用主流程开始节点的 article 工作流节点1和工作流节点2都只依赖主流程开始节点的article参数，两个子流程之间没有数据依赖，各自独立执行。 

版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点	用于整合信息。 • 模型：DeepSeek-V4-Flash • 输出格式：文本，开启流式输出 • 输入参数： – points_content：引用工作流节点A的 response_content – quotes_content：引用工作流节点B的 response_content • 系统提示词： 你是一位阅读笔记整理专家。 请将以下两部分内容整合为一份格式规范、排版清晰的阅读笔记。 整合要求： - 按"一、文章主旨与核心观点 二、金句摘录"两个章节组织 - 各章节标题使用Markdown二级标题（##） - 保持原有内容完整，不要增减或改写信息 - 在笔记开头加一行简短引导语，例如"以下是本文的阅读笔记：" - 整体风格简洁工整，适合存入个人知识库 • 用户提示词： 第一部分——主旨与观点： {{points_content}} 第二部分——金句摘录： {{quotes_content}} 

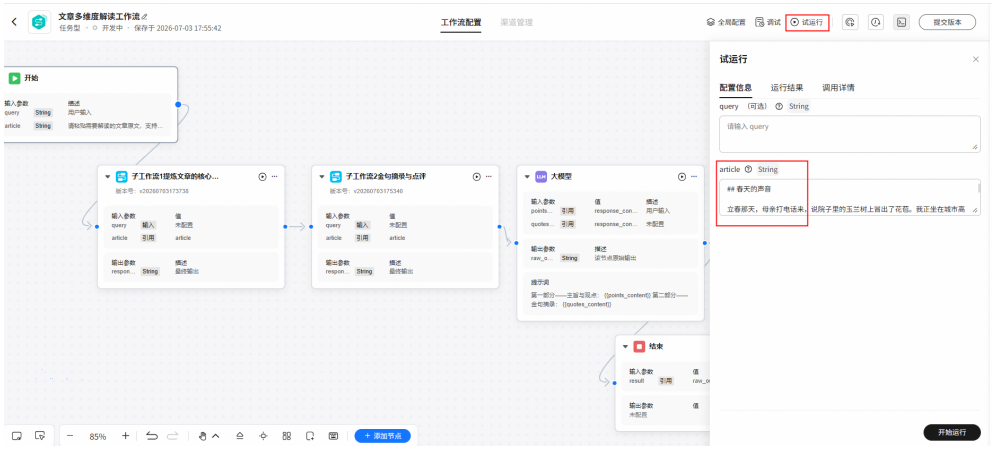
版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： result：引用大模型节点的 raw_output • 指定回复：{{result}} 

试运行：

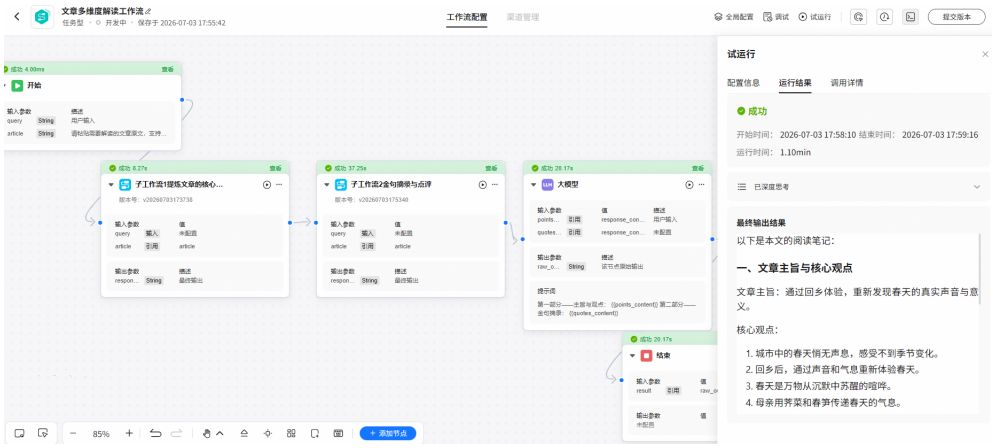
在开始节点的article参数中粘贴任意一篇文章，工作流将依次调用两个子流程分别完成观点提炼和金句摘录，再由大模型整合为一份完整的阅读笔记输出。两个子流程的执行过程在调试视图中清晰可见。

图 4-22 试运行



 版权所有 © 华为云计算技术有限公司

图 4-23 查看运行结果



4.4.7 反思精炼循环

介绍

反思精炼循环是一种闭环**自我优化**的搭建范式：工作流不满足于大模型的第一次输出，而是引入一个独立的评估环节对结果进行审查，若结果未达标则触发修订，如此循环，直到输出质量满足要求或达到最大迭代次数为止。

这种范式**解决了单次大模型调用不够可靠的问题**，特别适用于：

- 输出需要满足明确可检验标准的场景（格式、字数、关键信息完整性）。
- 对输出质量有较高要求、需要反复打磨的场景（正式文书、营销标语）。
- 希望系统自主完成质量把关、减少人工审核的场景。

原理

```
[初始输入]
   ↓
[生成节点：初次生成内容] → draft（草稿）
   ↓
[循环节点（指定最大轮次）]
   ├── 中间变量：current_draft（当前草稿）、is_passed（布尔值，初始 false）
   ├── 终止条件：is_passed 为 true
   ├── 循环体：
   │     [大模型节点A：评估当前草稿]
   │           ↓ pass（Boolean）+ feedback（String）
   │     [判断节点：是否通过？]
   │     ├──→ 通过 → [变量赋值：is_passed = true] → 循环输出（提前终止）
   │     └──→ 不通过 → [大模型节点B：依据反馈修订] → [变量赋值：更新 current_draft] → 循环输出
   └── 输出参数：final_draft（引用中间变量 current_draft）
   ↓
[下游节点 / 结束节点]
```

表 4-16 关键机制

机制	说明
循环最大次数	防止无限迭代，保证工作流必然终止。
中间变量 current_draft	每轮修订都基于上一轮的草稿，而非重新生成。

版权所有 © 华为云计算技术有限公司

机制	说明
终止条件 is_passed == true	质量达标后立即退出，不必跑满所有轮次。

操作示例：产品宣传语打磨工作流

场景描述：用户输入产品名称和核心卖点，工作流先生成一条宣传标语草稿，再由评估节点检查是否满足三项标准（字数在20字以内/包含核心卖点/朗朗上口），不满足则自动根据评估意见修改，最多迭代3轮，输出最终版标语。

这个示例用户只需输入产品名称和一两个卖点即可运行，每轮迭代都能看到草稿如何在评估反馈下逐步改进，直观体现了反思精炼的过程。

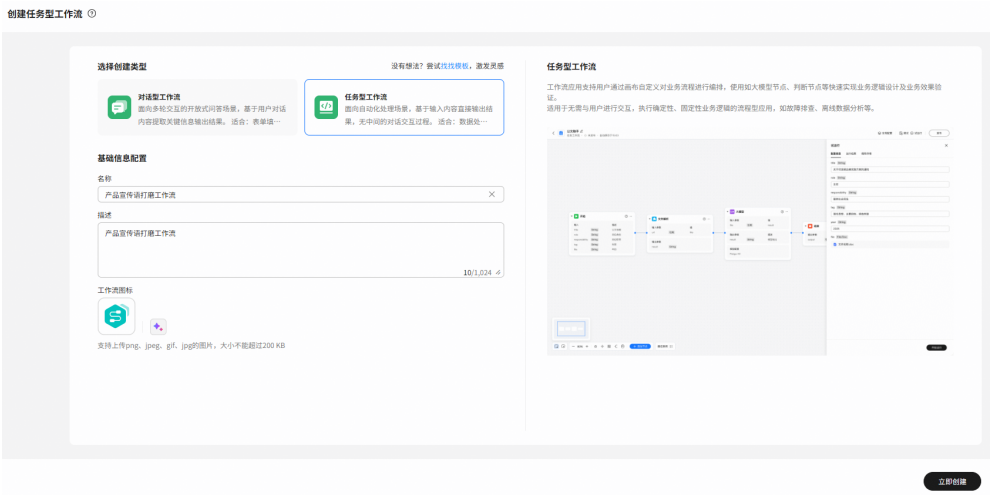
工作流结构：

```
[开始节点]
   ↓ product（产品名）、selling_points（核心卖点）
[大模型节点：初稿生成]
   ↓ draft（第一版标语）
[循环节点（指定3次，终止条件：is_passed 为 true）]
   ├── 中间变量：current_draft（当前标语），is_passed（是否通过，初始 false）
   ├── 循环体：
   │    [大模型节点A：评估标语质量] → pass + feedback
   │    [判断节点]
   │     ├──→ 通过 → [变量赋值：is_passed = true] → 循环输出
   │     └──→ 不通过 → [大模型节点B：依据反馈修订标语] → [变量赋值：更新 current_draft] → 循环输出
   └── 输出参数：final_slogan（引用 current_draft）
   ↓
[结束节点]
```

创建工作流：

在AgentArts左侧导航栏中选择“开发中心 > 智能体管理”，在“工作流”页签，单击“创建工作流”。

图 4-24 任务型工作流



节点配置说明：

版权所有 © 华为云计算技术有限公司

图 4-25 产品宣传语打磨工作流

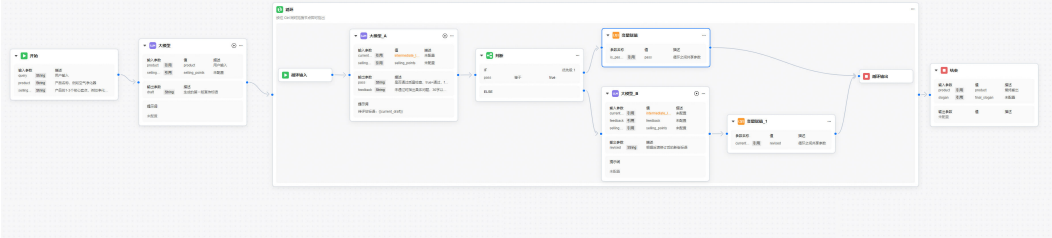


表 4-17 工作流各节点配置说明

节点	配置说明
开始节点	新增2个参数。 ● 参数一： – 参数名称：product – 参数类型：String – 描述：产品名称，例如空气净化器 – 必填：勾选 ● 参数二： – 参数名称：selling_points – 参数类型：String – 描述：产品的1-3个核心卖点，例如净化效率高、静音、一键操控 – 必填：勾选 

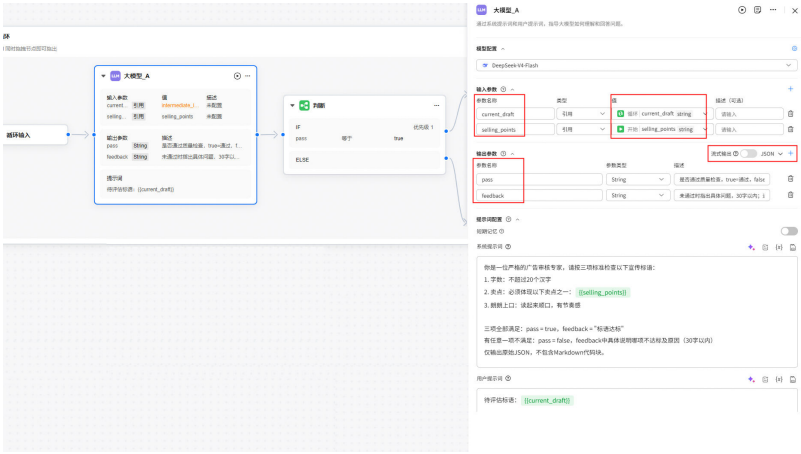
版权所有 © 华为云计算技术有限公司

节点	配置说明
大模型节点	用于生成初稿。 • 模型：DeepSeek-V4-Flash，输出格式：JSON，关闭流式输出 • 输入参数： – product：引用开始节点的 product – selling_points：引用开始节点的 selling_points • 输出参数：draft（String，描述：生成的第一版宣传标语） • 系统提示词： 你是一位品牌文案专家。 请为产品"{{product}}"生成一条宣传标语。 产品核心卖点：{{selling_points}} 要求：字数不超过20个汉字，突出核心卖点，简洁有力，易于记忆。 仅输出原始JSON，示例：{"draft": "标语内容"} • 用户提示词：不填，删除原有内容 

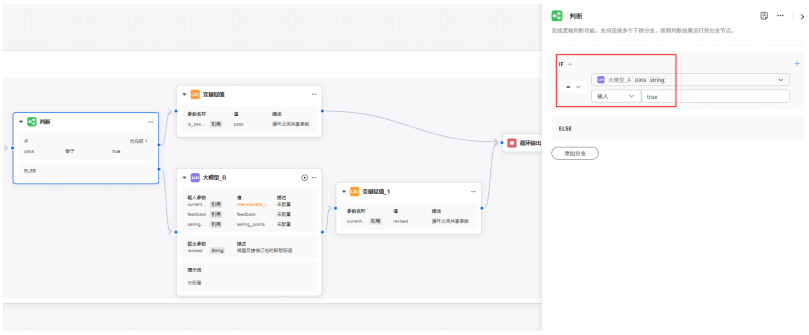
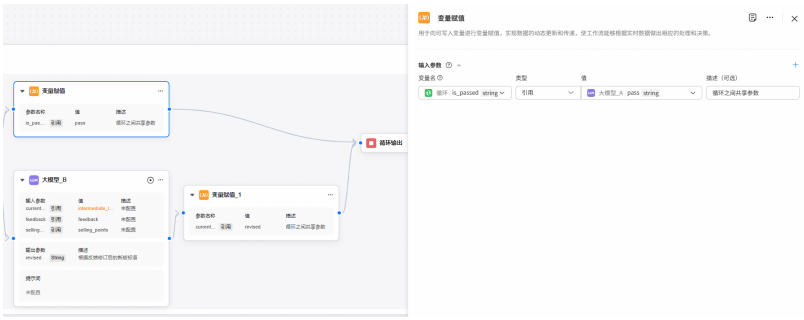
版权所有 © 华为云计算技术有限公司

节点	配置说明
循环节点	• 循环类型：指定循环次数，设置为 3次 • 中间变量： – current_draft，类型String，初始值引用大模型节点（初稿生成）的 draft。 – is_passed，类型Boolean，初始值输入固定值 false。 • 输出参数：final_slogan，引用中间变量 current_draft • 终止条件：is_passed 等于 true 

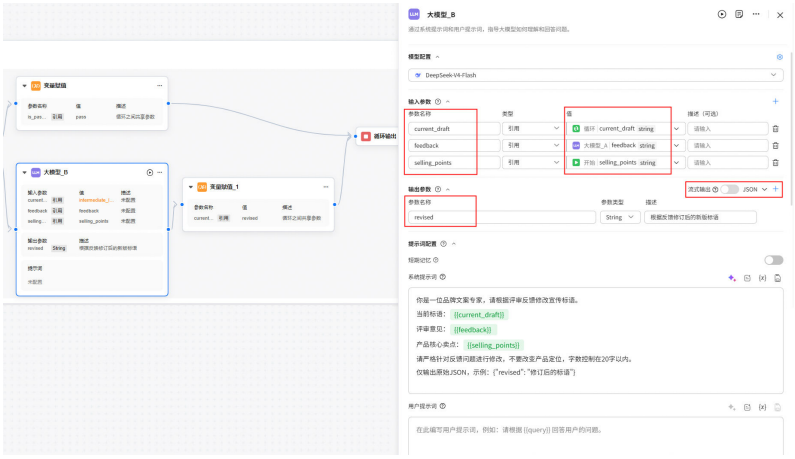
版权所有 © 华为云计算技术有限公司

节点	配置说明
	循环体 - 大模型节点A（评估标语质量） 连接循环输入 → 此节点。 ● 模型：DeepSeek-V4-Flash，生成随机性设为 0 ● 输入参数： – current_draft：引用循环节点中间变量 current_draft – selling_points：引用开始节点的 selling_points ● 输出格式：JSON，关闭流式输出 ● 输出参数： – pass，类型Boolean，描述为“是否通过质量检查，true=通过，false=不通过” – feedback，类型String，描述为“未通过时指出具体问题，30字以内；通过时输出标语达标” ● 系统提示词： 你是一位严格的广告审核专家，请按三项标准检查以下宣传标语： 1. 字数：不超过20个汉字 2. 卖点：必须体现以下卖点之一：{{selling_points}} 3. 朗朗上口：读起来顺口，有节奏感 三项全部满足：pass = true，feedback = "标语达标" 有任意一项不满足：pass = false，feedback中具体说明哪项不达标及原因（30字以内） 仅输出原始JSON，不包含Markdown代码块。 ● 用户提示词：待评估标语：{{current_draft}} 

 版权所有 © 华为云计算技术有限公司

节点	配置说明
	循环体 - 判断节点 ● IF 分支：大模型节点A的 pass 等于 true ● ELSE 分支：pass 为 false 
	循环体 - IF 分支（通过）：变量赋值节点 ● 变量名：中间变量 is_passed ● 类型：引用 ● 值：引用大模型节点A的 pass（此时为 true） 此节点连接到**循环输出**。 

版权所有 © 华为云计算技术有限公司

节点	配置说明
	循环体 - ELSE 分支（不通过）：大模型节点B（依据反馈修订） ● 模型：DeepSeek-V4-Flash，输出格式：JSON，关闭流式输出 ● 输入参数： – current_draft：引用循环节点中间变量 current_draft – feedback：引用大模型节点A的 feedback – selling_points：引用开始节点的 selling_points ● 输出参数：revised（String，描述：根据反馈修订后的新版标语） ● 系统提示词： 你是一位品牌文案专家，请根据评审反馈修改宣传标语。 当前标语：{{current_draft}} 评审意见：{{feedback}} 产品核心卖点：{{selling_points}} 请严格针对反馈问题进行修改，不要改变产品定位，字数控制在20字以内。 仅输出原始JSON，示例：{"revised": "修订后的标语"} ● 用户提示词：不填，删除原有内容 
	循环体 - ELSE 分支（不通过）：变量赋值节点，放在大模型节点B后面 ● 变量名：中间变量 current_draft ● 类型：引用 ● 值：引用大模型节点B的 revised 此节点连接到**循环输出**。 

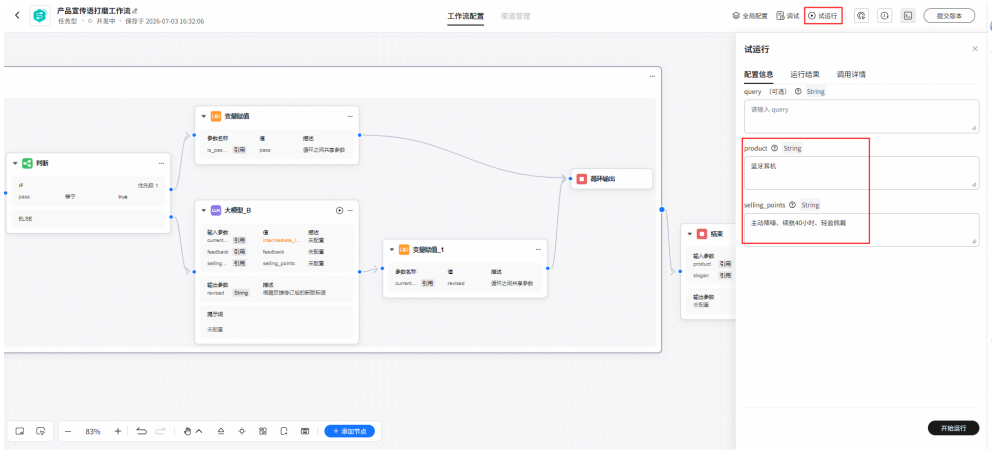
版权所有 © 华为云计算技术有限公司

节点	配置说明
结束节点	• 输入参数： – product：引用开始节点的 product – slogan：引用循环节点的 final_slogan • 指定回复： 【{{product}}】 最终宣传标语：{{slogan}} 

试运行：

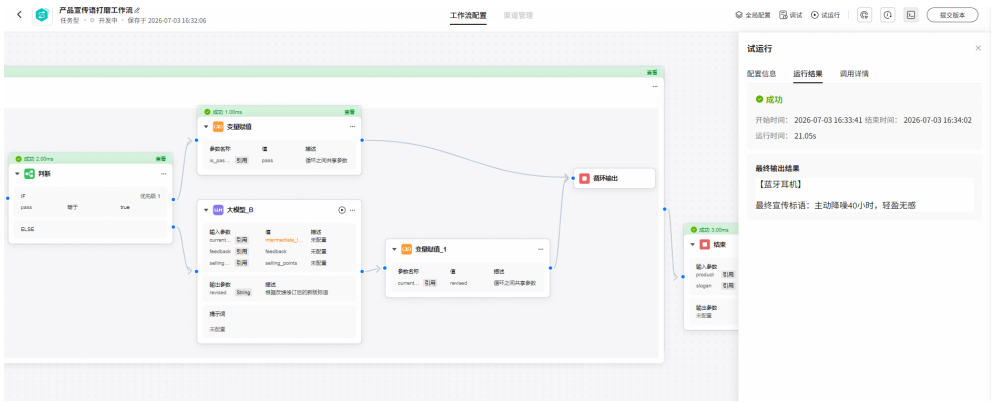
在开始节点的product中输入“蓝牙耳机”，selling_points 输入“主动降噪、续航40小时、轻盈佩戴”，观察工作流如何在最多3轮迭代中将初稿标语打磨至达标。在调试视图中可以看到每轮的草稿、评估结论和修订过程。

图 4-26 试运行



版权所有 © 华为云计算技术有限公司

图 4-27 查看运行结果



4.5 搭建工作流

4.5.1 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。

AgentArts支持创建工作流应用的方式如表4-18所示。

表 4-18 创建方式说明

创建方式	功能	优点	缺点	操作指导
从空白创建	基于平台可视化画布，可以通过拖拽节点和配置相关参数，迅速搭建工作流。	可控性强、透明度高。	开发成本高。	常规创建工作流
AI辅助创建	描述所需工作流具体应用场景和核心功能，平台会根据用户的需求，自动生成一个定制化工作流。	自适应能力强、用户体验好。	可控性差、数据依赖性强。	AI辅助创建工作流
使用预置应用创建	资产广场内置了工作流应用，用户可根据需要复制模板配置完全一样的工作流，并将其配置为符合自己需求的工作流应用。	高效的开发速度，低门槛。	定制化程度低，无法满足所有个性化需求。	-

选型建议：

- 如果您是初次使用或希望快速体验，推荐“使用预置应用创建”。
- 如果您有明确的复杂业务逻辑且需要精细控制，推荐“从空白创建”。

版权所有 © 华为云计算技术有限公司

- 如果您希望快速生成原型并通过自然语言描述需求，推荐“AI辅助创建”。

费用说明

不同套餐包允许创建的**应用数量限制**不同，具体可参考**计费说明**。

前提条件

- 已**开通AgentArts服务**。
- 模型已接入AgentArts平台，接入指导请参考**模型**。

常规创建工作流

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“工作流”页签，进入工作流应用管理界面。

步骤3 单击右上角“创建工作流”，进入“创建对话型工作流”或“创建任务型工作流”页面，相关区别请参见**表4-2**。

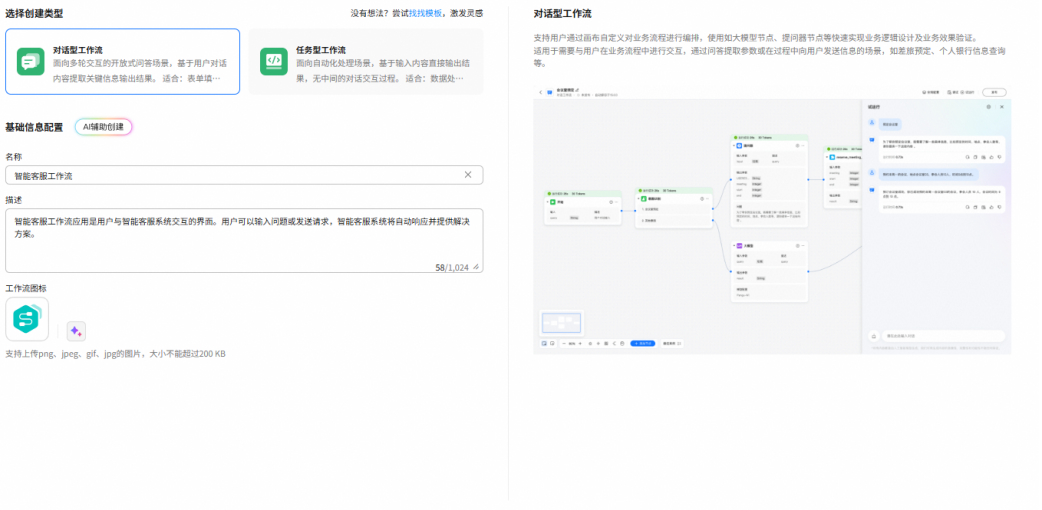
步骤4 选择完成后配置应用基础信息，参数说明如**表4-19**所示。

表 4-19 基础信息配置说明

参数	说明	示例
名称	输入内容支持中英文、数字、下划线和空格，长度2~64字符，但不允许以空格开头或结尾。	智能客服工作流
描述	描述工作流的功能，可直观呈现给用户，长度0~1024。	智能客服工作流应用是用户与智能客服系统交互的界面。用户可以输入问题或发送请求，智能客服系统将自动响应并提供解决方案。
工作流图标	系统默认工作流图标，用户也可以自定义图标或AI自动生成图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。支持jpg、jpeg、png、gif格式图片，且不大于200KB。	-

版权所有 © 华为云计算技术有限公司

图 4-28 创建工作流

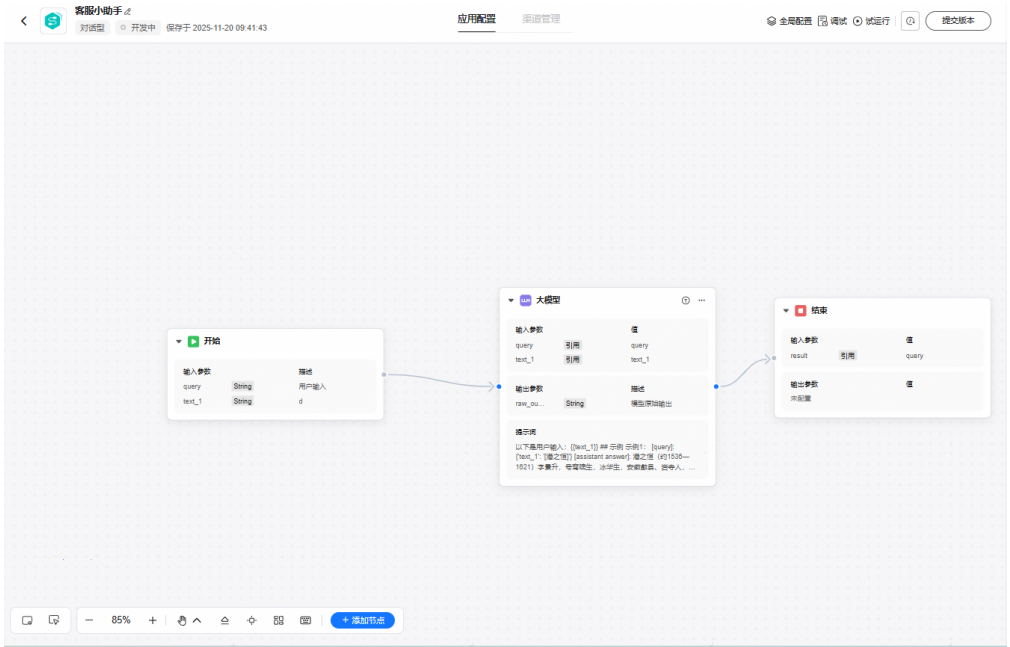


步骤5 配置完成后单击“立即创建”，进入工作流编排页面。

初始状态下工作流包含**开始**、**大模型**、**结束**节点。

- **开始**节点：用于启动工作流，详细配置请参考**开始节点**。
- **大模型**节点：（可选）提供了使用大模型的能力，可在节点中配置已部署的模型，用户可以通过编写提示词、设置参数让模型处理相应任务。如果无需配置，可单击右上角 “…” 删除节点，详细配置请参考**大模型**。
- **结束**节点：用于返回工作流的运行结果，详细配置请参考**结束节点**。

图 4-29 编排画布界面



----结束

版权所有 © 华为云计算技术有限公司

AI 辅助创建工作流

⚠ 注意

- AI辅助创建仅支持创建对话型工作流。
- AgentArts支持通过AI创建工作流，既可通过“开发中心 > 智能体管理 > 工作流”进行创建，也能在“总览 > 快速创建 > AI辅助创建”中进行创建。

在日常开发中，手动编写复杂流程图和代码逻辑既耗时又易出错。NL2Workflow通过自然语言描述业务需求，自动生成Mermaid格式流程图和DSL（JSON）代码，简化开发流程。系统支持多轮对话交互，确保流程准确无误，并允许用户调整和确认，确保最终生成的工作流既符合业务需求又具备可执行性。

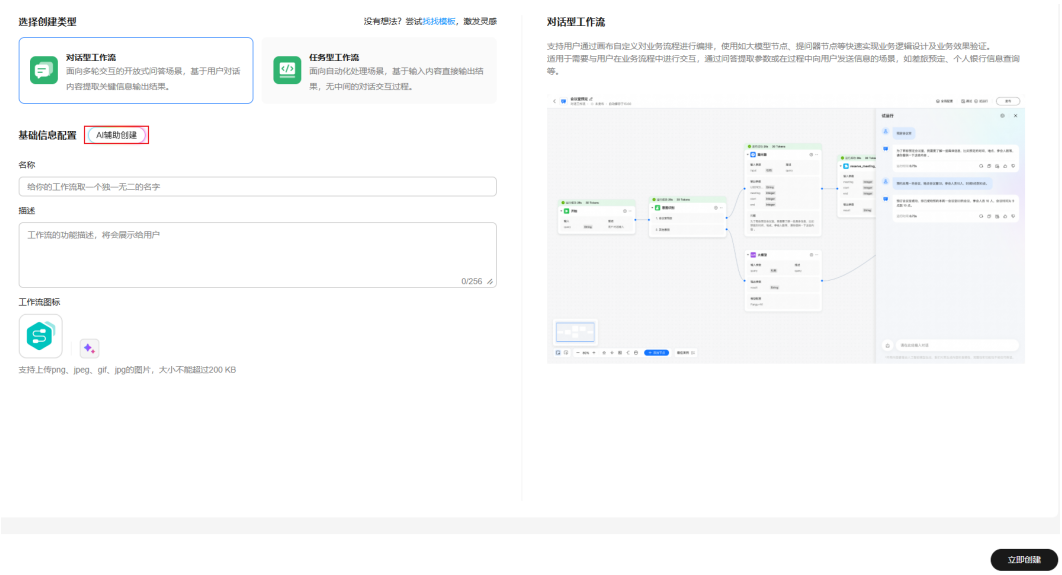
步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“工作流”页签，进入工作流应用管理界面。

步骤3 在工作流应用管理界面右上角单击“创建工作流”。

步骤4 在“创建对话型工作流”页面单击“AI辅助创建”。

图 4-30 AI 辅助创建



步骤5 选择输入框上方的任务，或在输入框中输入任务，并在输入框左下角选择“模型”和“工作流应用”，单击，以“创建假期审批工作流”为例。

版权所有 © 华为云计算技术有限公司

图 4-31 创建假期审批工作流

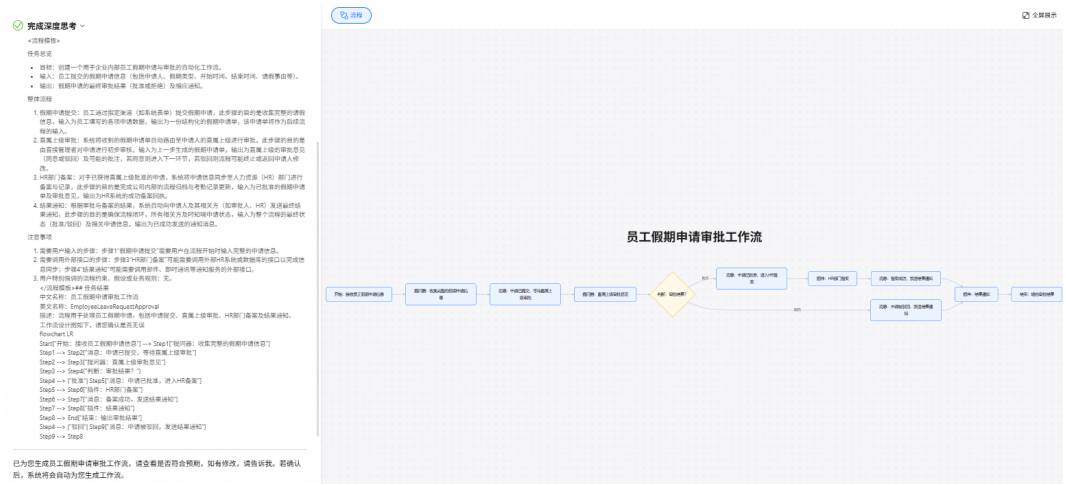


步骤6 在思考界面的输入框中提交需求，并在思考结果后单击“确定生成”。

说明

- 如果思考结果不满足要求，您可在输入框中重新输入指令。
- 在工作流编排页面您可以根据需要扩展工作流应用的能力。

图 4-32 思考结果



步骤7 AgentArts将根据工作流设计图生成工作流应用，可在预览界面查看，可单击界面右上角“前往编辑”跳转至工作流编排页面或单击“提交版本”。

说明

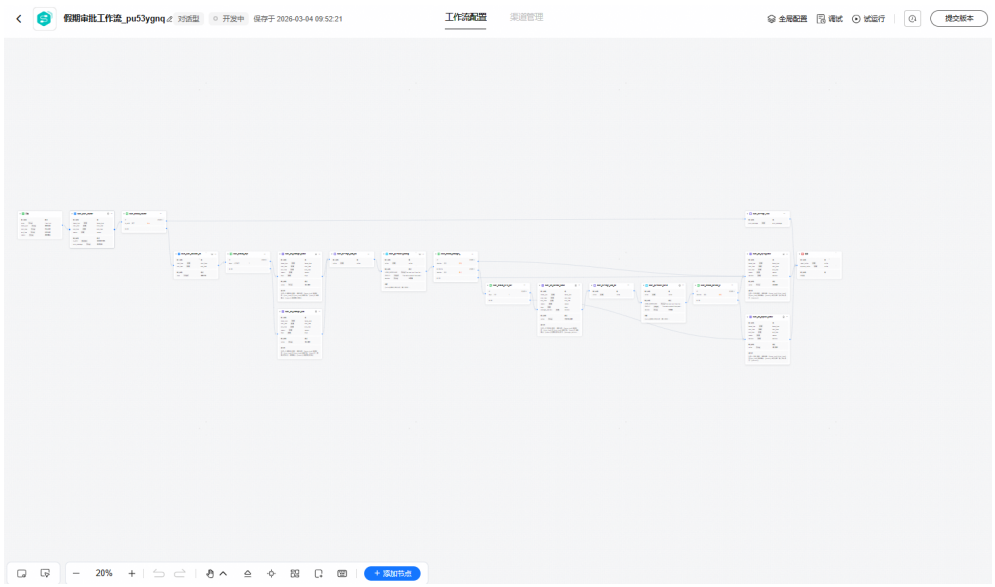
工作流名称在生成工作流应用时会自动生成，可单击 按钮修改。

版权所有 © 华为云计算技术有限公司

图 4-33 工作流预览



图 4-34 工作流应用编排页面



----结束

全局配置

在工作流编排界面，在画布右上角有全局配置入口，用于配置对话型工作流对话体验、默认模型、全局特性开关和定义的配置能力。

说明

工作流发布后的版本作为独立的资源，不支持修改全局配置。

版权所有 © 华为云计算技术有限公司

表 4-20 全局配置参数说明

参数	功能
默认模型	作为开场白、推荐问题的智能生成模型来源，新增节点默认使用该模型配置。 ● 单击模型配置下拉框，可配置默认模型，为新拖入节点提供默认的模型选项。 ● 勾选模型配置下的复选框 可将全局模型一键修改，提升模型配置效率。 **说明** 模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型服务**时的“选择标签”、“模型类型”。 ● 接入模型时的“选择标签”： ● 联网：表示该大模型具备联网搜索能力。 ● 思考：表示该大模型具备思维推理能力。 ● 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。 ● 限额体验：表示该平台预置大模型可以体验，会话轮数最大为20次。 ● “模型类型”包含： ● 文本：表示该大模型是文本对话类型。 ● 视觉：表示该大模型是图像理解类型。 ● 嵌入：表示该大模型是文本向量化类型。 ● 排序：表示该大模型是文本排序类型。 ● 模型状态： ● 未验证：表示该大模型未校验鉴权信息，不可使用。 ● 成功：表示该大模型鉴权信息校验成功，可以使用。 ● 失败：表示该大模型鉴权信息校验失败，不可使用。

版权所有 © 华为云计算技术有限公司

参数	功能
对话体验	该描述将在气泡内作为应用开场白展示给用户。最大支持输入226个字。 ● 支持在对话框中为对话型工作流中配置开场白、推荐问题。 ● 支持智能生成开场白和推荐问题。 – 智能生成开场白：开场白对话框中输入开场白概述，单击右上角 按钮，在“替换开场白”弹窗中单击“确定”，系统将自动生成开场白并替换当前开场白内容。 – 智能生成推荐问题：推荐问题对话框中输入问题概述，单击右上角 按钮，在“替换推荐问题”弹窗中单击“确定”，系统将自动生成推荐问题并替换当前对话框中的内容。 **说明** 仅支持添加3个推荐问题。 – 追问：当用户在全局配置中设置模型后，该功能将生效，以便在与工作流对话时主动提出跟进问题，以澄清需求或深化交互。 **用户自定义追问生成规则：**用户可以根据需求，自定义追问生成规则。通过这些规则，智能体在每轮回复后，能够根据对话内容智能地提供提问建议，从而挖掘用户的潜在需求。 **功能开启设置：**在“对话体验 > 追问”中，用户可以选择是否开启“追问”功能。如果开启，模型将在每轮回复后，默认根据对话内容提供提问建议，进一步优化对话体验。
记忆变量	记忆变量的节点赋值支持工作流节点的引用，同时支持引用对象模板和JSON导入。当全局变量配置了节点赋值，同时开始节点用户配置了其他输入参数时，可以在试运行界面中对这些参数进行调试。记忆变量支持会话级，当会话结束后记录的参数值将自动恢复为默认值。 记忆变量支持以下参数配置： ● 名称：支持字母、数字、下划线，且不能以数字开头。 ● 类型：支持配置string、number、boolean、object、integer、array多种类型的参数，其中object类型参数最多支持5层嵌套。 ● 描述：（可选）单击 *** 图标可配置描述参数信息，帮助理解传入参数的含义。 ● 默认值：（可选）单击 *** 图标您可以设置输入参数的默认值，其中Object类型参数的默认值需输入Json数据。 ● 单击 图标可删除记忆变量。

版权所有 © 华为云计算技术有限公司

参数	功能
内容审核配置	**注意** 审查用户输入内容和审查应用输出内容应至少开启一项。 内容审核配置功能旨在帮助用户对输入和输出内容进行严格审查与管理，确保内容符合预设的规则和标准。该功能支持自定义审核规则、关键词匹配、以及多种处理策略，适用于多种应用场景。 支持通过单击右侧的开关按钮“启动”或“关闭”内容审核配置功能。内容审核配置功能开启时，可通过单击 图标设置关键词匹配处理输入输出内容，支持添加多个审核规则。 ● **审核规则设置** 用户可创建多个审核规则，每个规则可独立配置关键词和处理策略。系统支持多规则并行审核，满足不同场景下的内容管理需求。 – 示例：系统中可设置“审核规则-1”、“审核规则-2”等，用户可根据实际需要添加更多规则。 – 每个规则可独立配置关键词和处理策略，互不干扰。 ● **关键词配置** 在每个审核规则中，用户可设置关键词用于内容匹配。关键词支持多行输入，每行一个关键词，最多支持**10000字符**。 – 系统将对用户输入或输出内容进行关键词匹配，如果匹配成功，将根据预设策略进行处理。 – 关键词支持中文、英文、数字及特殊字符，适用于多种内容类型。 ● **审查用户输入内容** 用户可单击右侧的开关按钮“启动”或“关闭”“审查用户输入内容”功能。当用户输入内容中包含关键词时，系统将自动根据预设策略进行处理。 – **预设回复内容**：用户可自定义回复文案，用于替代原始输入内容，输入审核关键词时，多个关键词之间需用逗号隔开。回复内容最多支持**1000字符**。 – 该功能可用于拦截不当输入、引导用户规范表达等场景。 ● **审查应用输出内容** 用户还可单击右侧的开关按钮“启动”或“关闭”“审查应用输出内容”功能。当系统输出内容中包含关键词时，系统将自动根据预设策略进行处理。 – **输出处理方式包括**： – **回复预设文案**：输出预设的固定内容，替代原始输出。 – **替换内容**：将匹配的关键词替换为指定内容。 – **屏蔽内容**：直接屏蔽匹配内容，不进行输出。 – 用户可根据实际需求选择合适的策略，灵活控制输出内容。 ● **添加规则** 用户可通过“添加规则”按钮，新增审核规则，以满足不同场景下的内容审核需求。 – 每个新增规则均可独立配置关键词和处理策略。 – 支持多规则并行审核，提升内容管理的灵活性和准确性。

版权所有 © 华为云计算技术有限公司

编排工作流

创建工作流后，初始状态下工作流包含**开始**、**大模型**、**结束**节点。在画布中添加节点，并按照任务执行顺序连接节点，同时按照工作流业务流向配置输入参数和输出参数。

工作流内置了多种基础节点，同时还可以添加“插件”节点来执行特定任务。插件节点使用方法详见**插件**。

画布界面操作详见**画布操作说明**。

步骤1 在工作流面板中单击“添加节点”，选择目标节点。

步骤2 将各个节点相连接，连接时需注意业务流向。

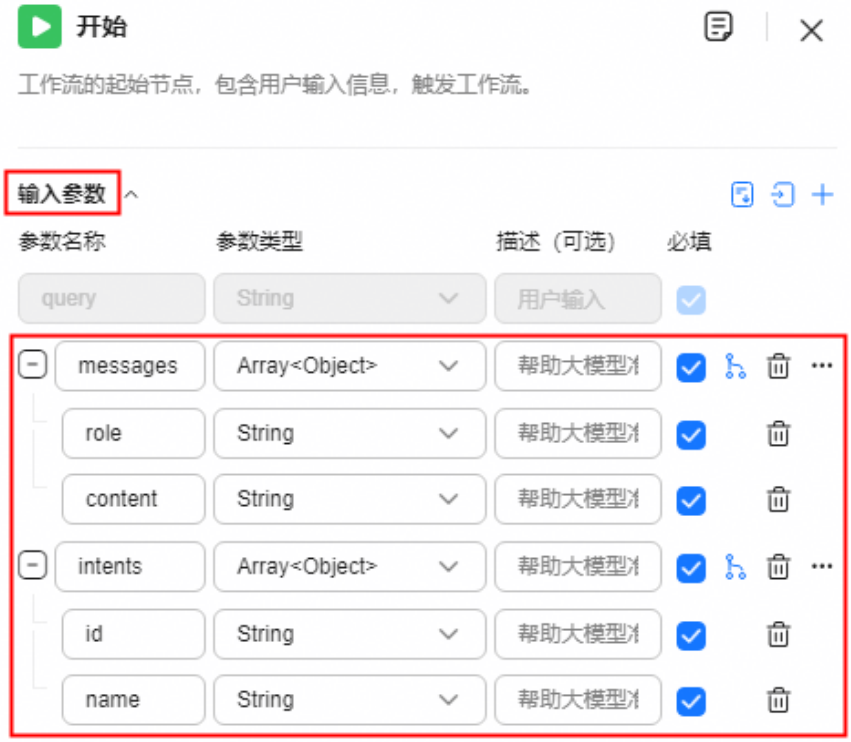
步骤3 配置节点的输入参数和输出参数。

各个节点的输入输出参数配置请参考**基础节点**、**通用节点**、**逻辑节点**、**工具节点**、**消息管理节点**、**变量&知识节点**。

如果此工作流用于多智能体应用的意图识别，则配置如下：

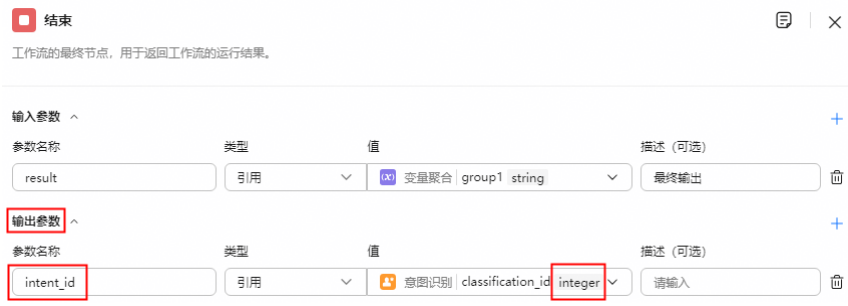
- 开始节点必须配置如**图4-35**所示参数，参数名称和参数类型必须保持一致。
- 结束节点必须配置如**图4-36**所示参数，输出参数名称以及引用的值必须为integer类型。

图 4-35 开始节点配置



版权所有 © 华为云计算技术有限公司

图 4-36 结束节点配置



----结束

画布操作说明

工作流构建过程中，画布中可以执行的操作如图4-37所示。

图 4-37 画布界面操作

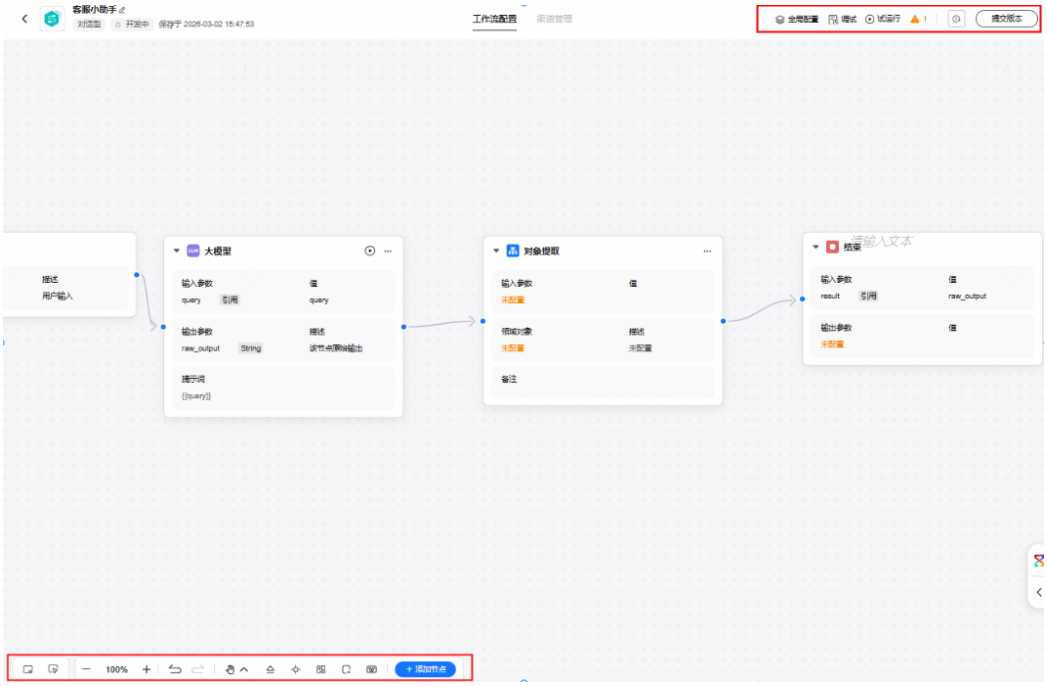


表 4-21 画布操作说明

操作	说明
删除节点	不支持删除起始节点和结束节点。 鼠标光标移至节点上，单击“ ··· > 删除”，即可删除节点。 **注意** 删除节点后将永久清除该节点的配置数据，请谨慎操作。

版权所有 © 华为云计算技术有限公司

操作	说明
复制节点	鼠标光标移至节点上，单击“ ··· > 复制”，画布中即可出现复制的节点。 **说明** 不支持跨画布复制节点。
重命名节点	鼠标光标移至节点上，单击“ ··· > 重命名”，在重命名窗口中输入节点名称。 **说明** 同一个工作流画布中节点名称不可重复。
显示缩略图	单击画布左下角 ，在画布中显示或取消工作流缩放图。
查看画布节点	单击画布左下角 ，查看画布节点，支持在查看画布节点界面输入节点名称搜索节点。
缩放工作流	单击画布左下角 − 100% + 两侧符号，调整工作流在画布中显示大小，步长10。
撤销	单击画布左下角 ，可撤销上一步对画布的操作。
恢复	单击画布左下角 ，可恢复上一步对画布的操作。
鼠标模式	单击画布左下角 ，可切换鼠标操作样式，支持切换“手模式”和“指针模式”。
全局节点折叠	单击画布左下角 ，折叠全局节点，折叠后画布内仅显示节点名称和业务流向。
全局节点打开	单击画布左下角 ，打开全局节点，打开后画布内显示节点配置，包括输入参数、输出参数、描述等。
居中	单击画布左下角 ，一键将工作流调整至画布中间位置。
布局优化	单击画布左下角 ，优化工作流编排样式。
添加注释	单击画布左下角 ，可为节点增加卡片注释，支持增加、删除、修改、查看、拖动、对齐和缩放等。

版权所有 © 华为云计算技术有限公司

操作	说明
快捷键	单击画布左下角，了解工作流画布快捷键操作。 • 手模式：英文输入法时，单击“H”或“h”键可切换鼠标模式为“手模式”，方便在画布上进行移动操作。 • 指针模式：英文输入法时，单击“V”或“v”键可切换鼠标模式为“指针模式”，用于选择和操作节点。 • 移动画布内容：“手模式”时可直接拖拽移动画布内容，以便查看不同区域的内容；“指针模式”时须按空格键后拖拽移动画布。 • 放大：通过滚动鼠标滚轮或按下“ ctrl”+ “+ ”键来放大画布，以便更详细地查看内容。 • 缩小：通过滚动鼠标滚轮或按下“ ctrl” +“ -”键来缩小画布，以便查看更大的范围。 • 撤销：按下 “ ctrl”+ “Z ”键可以撤销上一步操作，方便进行错误修正。 • 删除：单击“Delete”或“Backspace”键可以删除选中的节点或内容。 • 新增节点：单击“Tab”键或双击鼠标左键可以在当前位置新增一个节点。 • 多选：按住“Shift”键并单击节点，或者在“指针模式”下框选，可以进行多选操作。 • 全选：按下“ ctrl”+“A”键可以全选画布上的所有节点或内容。 • 连接到结束节点：英文输入法时，单击“Q”键可以将当前节点连接到结束节点，方便进行流程或路径的构建。
全局配置	单击画布右上角，可配置全局参数，包括模型配置、对话体验、记忆变量、节点赋值、内容审核配置、语音交互等。
调试	单击画布右上角，进入“调试”界面，支持查看“运行结果”和“调用详情”，详细操作可参考**试运行工作流**。
试运行	单击画布右上角，进入“试运行”界面，并在输入框中输入问题测试应用功能。
问题清单	单击画布右上角，可查看并快速定位工作流中的问题清单。
触发器配置	单击画布右上角，进入触发器配置页面，通过定时任务，调用应用按照指令要求执行，详细操作可参考**配置触发器**。 **说明** 仅任务型工作流支持配置触发器，对话型工作流不支持。

版权所有 © 华为云计算技术有限公司

操作	说明
发布历史	单击画布右上角  ，查看应用发布历史，支持还原版本和删除发布版本，详细操作可参考**查看发布历史**。 **注意** 删除发布版本后，该版本将不可恢复，且可能影响已发布渠道的调用，请谨慎操作。
渠道管理	单击画布上方渠道管理，支持查看已发布工作流的发布详情。
提交版本	单击“提交版本”，完善版本名称，可将工作流发布成不同发布渠道，详细操作可参考**发布并部署工作流**。

4.5.2 试运行工作流

开发者可以在工作流创建完成后，直接与工作流进行交互，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。平台提供的全链路调试功能，允许开发者查看每条用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而能够高效定位问题并快速调整配置。

AgentArts支持对整个工作流进行调试，也支持对工作流的单个节点进行调试。

费用说明

试运行工作流会实际调用大模型、插件或知识库等服务，将消耗相关资源额度或产生费用，具体计费规则请参考**计费项**。

前提条件

- 已**创建工作流**。

约束与限制

试运行工作流时，端到端运行时间最长可执行**10分钟**。

试运行工作流（必选）

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“工作流”页签，进入工作流应用管理界面。

步骤3 在工作流界面中找到目标工作流，并单击进入工作流编排页面。

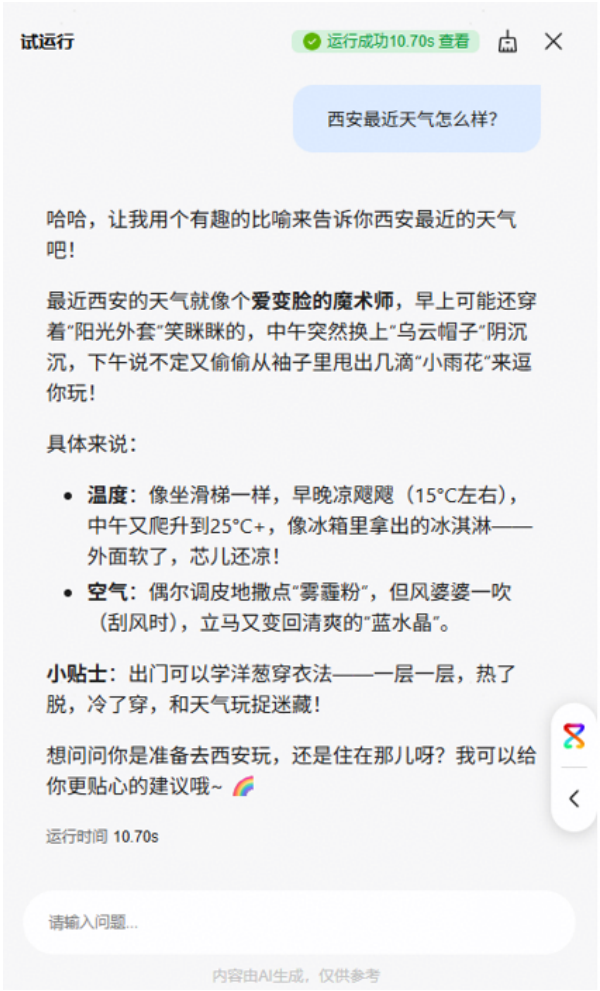
步骤4 试运行工作流。

在工作流的开始节点中，默认包含一个输入参数query，用于表示用户在本轮对话中输入的原始内容。您也可以根据需要添加其他参数，以供下游节点使用。因此，试运行工作流分为两种场景：一种是仅使用开始节点的默认输入参数query，另一种是开始节点用户添加了其他参数。

- **开始节点默认输入参数query时**，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。

版权所有 © 华为云计算技术有限公司

图 4-38 试运行工作流



说明

- 单击试运行页面右上角 ，一键清除预览调试界面内容。
- 文件输入：请参考**开始节点**配置参数，可增加“文件”或“文件数组”类型，并在试运行界面中上传文件。

- **开始节点用户添加了其他参数时**，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，单击“开始运行”，等待返回试运行结果。

 版权所有 © 华为云计算技术有限公司

图 4-39 试运行工作流



步骤5 在试运行结束后，可以单击右上角运行时间可查看调试结果，包括运行结果与调用详情，如图4-40所示。

如果试运行失败，常见报错与解决方案请详见**开发常见报错**。

版权所有 © 华为云计算技术有限公司

图 4-40 调试结果示例（列表）



图 4-41 调试结果示例（火焰图）



版权所有 © 华为云计算技术有限公司

说明

- 单击调用详情页面中的按钮，查看列表调用链，适用于查看具体的输入输出参数、排查逻辑错误。
- 单击调用详情页面中的按钮，查看火焰图调用链，适用于分析性能瓶颈、定位耗时较长的节点。

----结束

调试单节点

以调试“意图识别”节点为例：

步骤1 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。

步骤2 输入参数内容，单击“开始运行”。

图 4-42 编写输入参数内容



步骤3 在“运行结果”页面，查看当前节点的运行结果。

运行成功，节点处也将显示“运行成功”字样。

运行失败，需要根据提示调整节点参数。

版权所有 © 华为云计算技术有限公司

图 4-43 单节点调试运行成功示例



----结束

4.5.3 配置触发器

触发器是工作流中实现任务自动化的关键功能，它能够在特定条件下自动启动任务执行，无需人工干预。可以利用触发器灵活设置任务的启动条件，触发器能确保任务按时按需完成，从而提高工作流应用的自动化水平和响应速度。

说明

仅任务型工作流支持配置触发器，对话型工作流不支持。

触发方式

通过定时任务调用工作流，使其按照触发指令执行任务。

添加触发器

步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“工作流”页签，进入工作流应用管理界面。

步骤3 单击目标任务工作流，在工作流应用右上角单击触发器配置按钮 。

步骤4 在触发器配置页面右侧，单击+。

 版权所有 © 华为云计算技术有限公司

步骤5 配置触发器信息，参数配置说明请参考**表4-22**。

表 4-22 创建触发器参数说明

参数	说明
触发器名称	触发器的名称。 由2~20个字符组成，支持中英文、数字、下划线，仅支持中英文开头。
触发类型	• 周期触发：用户可设定每周、每日或每月固定时间触发。 • 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。
触发时间	按设置的时间触发智能体应用的执行。例如，设置触发时间为每1小时执行一次，则每隔1小时，重复执行一次会话。 • 周期触发 – 用户可自定义触发时间按每日执行。 – 用户可自定义触发时间按每周执行。 – 用户可自定义触发时间按每月执行。 • 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。 – 间隔“天”：取值范围1~31。 – 间隔“小时”：取值范围1~23。 – 间隔“分钟”：取值范围1~59。 – 间隔“秒”：取值范围1~59。
机器人提示	输入自然语言指令，触发时Agent遵循该指令定时执行。如：“发送月底发票报销提醒”，则机器人会在设定的时间发送提醒。
调用方式	• 同步：触发器触发后，会等待被调用的操作完全执行完毕并返回结果，才继续执行后续流程。 适用于需要立即获取执行结果或后续流程依赖本次执行结果的场景。 • 异步：触发器触发后，仅发起调用请求，不等待操作完成，直接继续执行后续流程。 适用于执行时间较长、无需等待结果或仅需触发执行的场景，可避免因等待超时而阻塞流程。

 版权所有 © 华为云计算技术有限公司

图 4-44 创建定时任务触发器



步骤6 单击“确定”，即完成触发器创建，可在触发器列表中对触发器进行查看，修改和删除等操作。

注意

删除触发器后，该自动化任务将立即停止，请确认业务影响后再操作。

----结束

常见问题

1. 如果工作流被删除，关联的触发器会如何处理?

   如果工作流被删除，关联的触发器将自动失效。

4.5.4 发布并部署工作流

工作流试运行成功后，可对其进行发布及部署，便于后续使用，如在单智能体应用中使用工作流，请参考**添加工作流**。

费用说明

工作流发布后通过API接口调用时，会产生tokens消耗。

- 如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

前提条件

工作流已调试完成，具体请参见**试运行工作流**。

版权所有 © 华为云计算技术有限公司

发布工作流

步骤1 在工作流编排页面，单击右上角“提交版本”，在弹出的提交版本弹窗中填写“版本名称”和“版本描述”，单击“确定”。

表 4-23 提交版本界面信息

配置项	功能描述
版本名称	系统根据时间戳自动生成，同时支持用户自定义。
版本描述	（可选）在此处输入版本的详细描述，以帮助团队成员和用户了解版本更新的具体内容和改进点。
部署至实例	默认勾选。 将该版本同步部署为API接口，便于后续接口调用。

图 4-45 提交版本



步骤2 单击“确定”。

提交版本后，在“工作流”页面的卡片上，显示“已提交”。

如果勾选“部署至实例”，进入“部署智能体”页面，具体部署操作请参考**部署智能体**。

步骤3 发布完成后可以单击“分享”按钮，也可通过左侧导航栏选择“开发中心 > 智能体管理”，单击目标应用，进入应用配置页面，选择上方“渠道管理”页签，可进入渠道管理页面。

版权所有 © 华为云计算技术有限公司

图 4-46 版本分享



步骤4 在**渠道管理页面**中的**调用方式**页签单击“查看API”即可查看API调用接口信息，详细操作可参考**通过API调用工作流**。

如果未部署过版本，则API详情为空。

图 4-47 调用 API



步骤5 在**渠道管理**的**分享渠道**页签，可以看到多平台发布状态。

- 在**分享渠道**页签，**云商店**发布渠道的**操作**列单击“发布”按钮后将出现“复制链接”文字按钮。

  单击“复制链接”，可直接复制网页URL嵌入或分享到其他应用场景。
- 在**渠道管理**页面，**网页**分享渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。

图 4-48 发布 Agent



图 4-49 应用发布成功后



- 立即访问：单击当前页面的“立即访问”按钮，可立即进入网页版应用。
- 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。

 版权所有 © 华为云计算技术有限公司

- 重新生成：重新生成应用链接。生成的应用链接有效时限为24小时（从发布为网页或“重新生成”算起），超过有效时限则无法正常访问链接，需重新单击网页分享渠道的**操作**列的“重新生成”。

说明

- 已发布的应用支持在Agent应用编辑页面选择“更新版本”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。
- 发布为网页的应用，每日单用户调用应用的次数不超过100次，且不超过“每日调用限额（次）”设置的限额。

----结束

对接 APIG 网关并提交商品发布

Agent应用分享到云商店渠道后，商家可将应用发布到华为云云商店。

APIG 网关类商品要求商家先在华为云 APIG 共享版网关上创建并发布 API，云商店发布商品时绑定 APIG 的 API 分组。

说明

在华为云云商店发布产品请先完成这些**准备工作**，否则无法正常管理和上架。

步骤1 获取智能体API 信息。

1. 在左侧导航栏选择“开发中心 > 智能体管理”，单击目标应用，选择上方“渠道管理”页签。
2. 单击API调用方式操作列的“查看API”，进入API详情页，可获取服务地址、请求Path、请求方式、凭证Key、请求参数。

说明

如果未部署过任何版本，“查看API”页面中的API详情中的URL为空，请先完成部署后再查看。具体部署操作请参考**部署智能体**。

图 4-50 服务地址



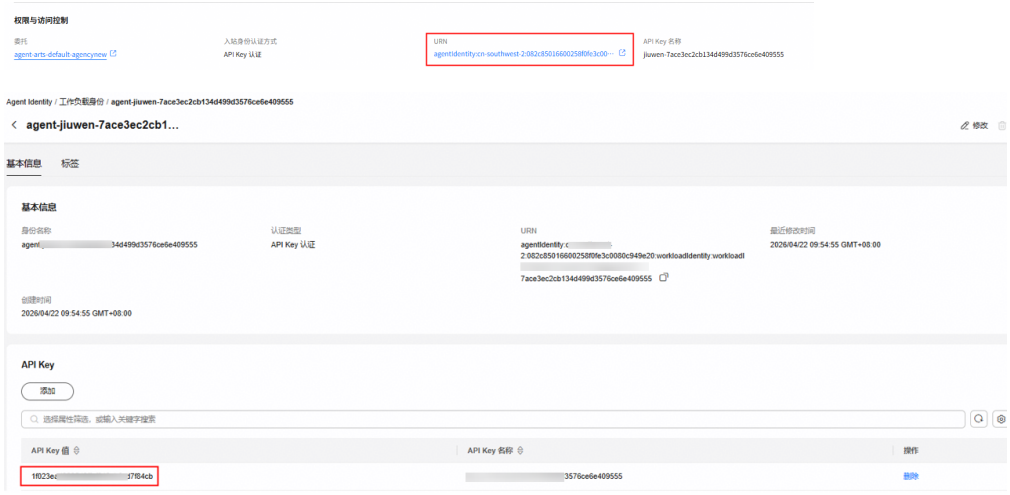
 版权所有 © 华为云计算技术有限公司

图 4-51 请求信息



3. 获取服务鉴权信息：凭证Value。

   在API详情页复制智能体ID，跳转至“部署运行 > 智能体运行时”进行搜索，在智能体运行详情页面打开RUN界面即可获取API Key。



步骤2 在 APIG 网关完成技术对接，详细操作请参见**APIG网关类商品技术对接方案**。

创建API填写样例：

 版权所有 © 华为云计算技术有限公司

图 4-52 定义 API 请求

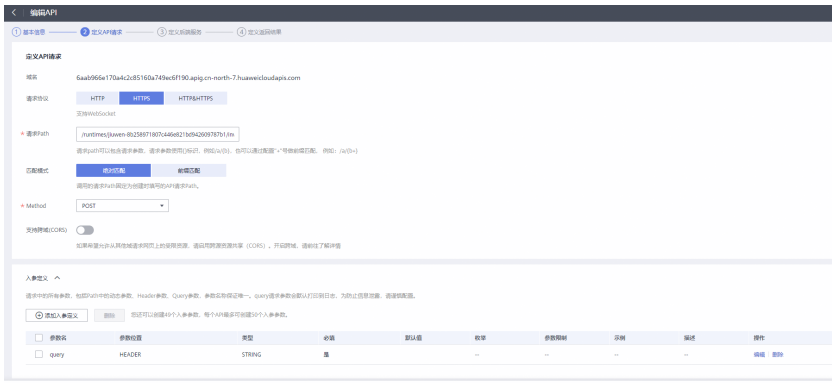
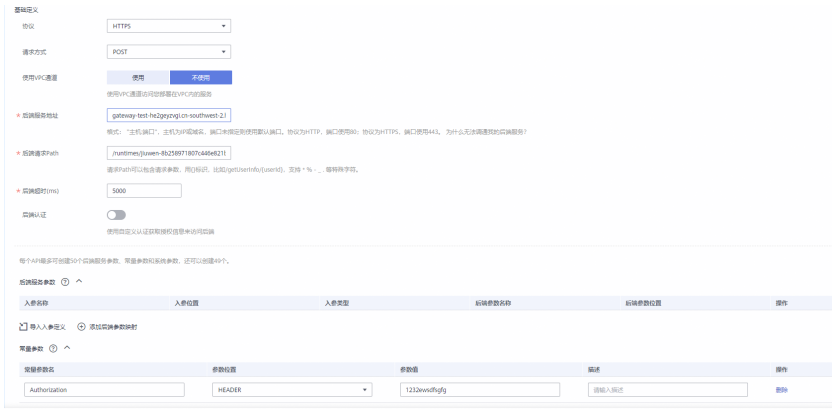


图 4-53 定义后端服务



步骤3 在卖家中心创建 APIG 网关类商品并提交审核，详细操作请参见**发布APIG网关类商品**。

----结束

准备工作

如需在华为云云商店发布产品请先完成以下准备，否则无法正常管理和上架：

步骤1 完成企业账号实名认证

华为云账号需完成企业实名认证，支持企业法人支付宝认证、对公账户打款认证等多种方式。详细操作请参见**企业账号实名认证流程**。

步骤2 加入华为云合作伙伴网络HCPN

完成实名认证后，申请加入 HCPN，获取合作伙伴基本权益，这是进入卖家中心的前置条件。详细操作请参见**加入华为云合作伙伴网络（HCPN）**。

步骤3 注册成为华为云商家

加入 HCPN 后，在卖家中心提交入驻申请，完成商业信息认证（含银行账号、税率等结算信息），审核通过后方可发布商品。详细操作请参见**注册成为商家**。

步骤4 开通华为支付（可选）

 版权所有 © 华为云计算技术有限公司

若需在云商店向买家收费，请提前完成华为支付开通。免费商品可跳过此步骤。详细操作请参见**开通华为支付**。

----结束

部署智能体

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航，选择“开发中心 > 智能体管理 ”。

步骤3 选择“工作流”页签，单击待部署的工作流应用卡片。

步骤4 在“工作流配置”页面右上角，单击  。

步骤5 在“部署智能体”页面，配置部署信息如**图4-54**所示，参数配置请参考**表4-24**。

图 4-54 部署智能体

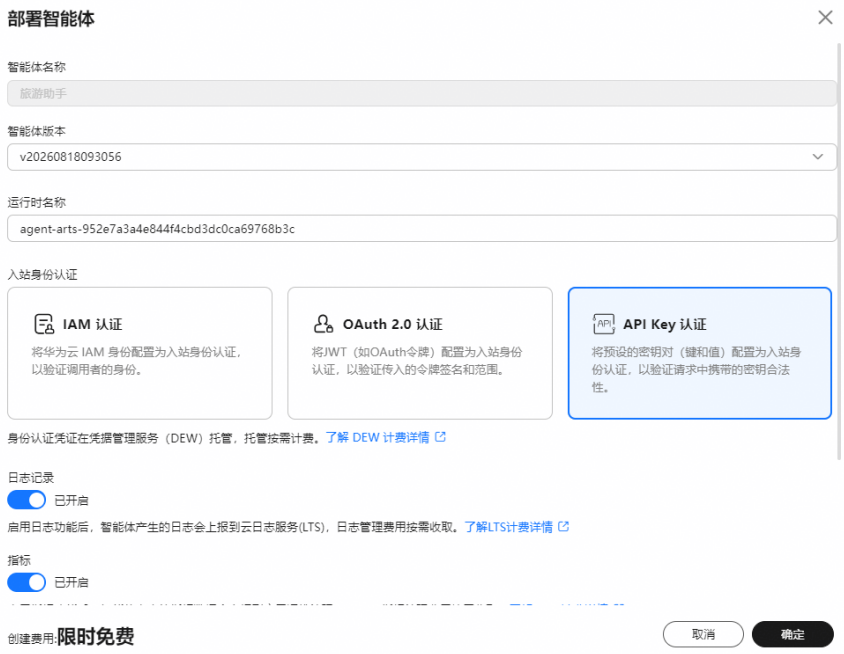


表 4-24 部署智能体

参数	说明
智能体名称	为**创建工作流**自定义的智能体名称，不可修改。
智能体版本	为**发布工作流**的版本名称。默认为最新的版本名称，可以在下拉框中切换。
运行时名称	运行时的名称。**仅第一次部署时支持自定义。** 由2~48个字符组成，包含小写字母、数字、中划线，以小写字母开头，以小写字母或数字结尾。

版权所有 © 华为云计算技术有限公司

参数	说明
入站身份认证	选择调用此API接口时，使用的认证鉴权方式，如何进行认证鉴权请参考**认证鉴权**。默认IAM认证，**仅第一次部署时支持选择。** • IAM认证：将华为云IAM身份配置为入站身份认证，以验证调用者的身份。适用于华为云内部系统或已接入华为云IAM的系统，集成复杂度低，安全等级高。 • OAuth 2.0认证：将JWT（如OAuth令牌）配置为入站身份认证，以验证传入的令牌签名和范围。适用于已有身份提供商（如Okta、Cognito等）的企业系统，集成复杂度高，安全等级高。 – Discovery URL：身份提供商（例如Okta、Cognito等）提供的Discovery URL，该URL通常可在该提供商的文档中找到。 – JWT授权配置：配置JWT令牌的访问控制规则。 ▪ 允许的受众：指定的受众是否在Agent Identity运行中指定的受众匹配或为其子集。 ▪ 允许的客户端：指定的客户端标识符是否被允许访问Agent Identity。 ▪ 允许的范围：仅当令牌包含此处配置的至少一个必须范围时，才允许访问自定义声明。 ▪ 自定义声明：仅当令牌中的特定声明与预定义字符串值相匹配时，才允许访问。 • API Key认证：将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。适用于内部测试或轻量级集成，集成复杂度低，安全等级中。 **注意** 身份认证凭证在凭据管理服务（DEW）托管，托管按需计费，使用DEW产生的费用请参考**DEW计费说明**。
日志记录	默认开启。 开启日志记录功能后，智能体产生的日志会上报到云日志服务 LTS，便于日志的集中管理和分析。日志管理费用按需收取，具体计费详情请参考**LTS计费说明**。 日志信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“日志”页签中。
指标	默认开启。 开启指标功能后，智能体产生的指标数据会上报到应用运维管理 AOM，便于实时监控和分析应用的运行状态。指标管理费用按需收取，具体计费详情请参考**AOM计费说明**。 指标信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“指标”页签中。

版权所有 © 华为云计算技术有限公司

参数	说明
调用链	默认开启。 开启调用链功能后，智能体产生的调用链数据会上报到应用性能管理 APM，帮助开发者追踪和分析应用的性能问题。调用链管理费用按需收取，具体计费详情请参考**APM计费说明**。 调用链信息展示在“运营运维 > 观测 > 调用链分析”的详情页的“链路信息”页签中。
智能体卫士	默认关闭。 开启后，实时检测智能体执行链路与操作轨迹，监控全量行为并识别安全异常。不同安全风险处理规则和检测记录可参考**企业主机安全 HSS**管理及查看。HSS的智能体卫士当前为公测阶段，暂不收费。

步骤6 单击“确定”。

部署成功后，可以在“部署运行 > 智能体运行时”页面，通过搜索运行时名称查看部署的智能体。

工作流部署后，支持通过API调用，具体请参考**通过API调用工作流**。

步骤7 （可选）在“发布部署”页面，单击“分享”，进入“渠道管理”页面。在“渠道管理”中单击“查看API”即可查看API调用接口信息。

----结束

查看发布历史

单击右上角  ，可查看当前工作流发布历史记录。

版权所有 © 华为云计算技术有限公司

图 4-55 发布历史



说明

发布历史支持版本对比、还原版本和删除操作：

- 版本对比：单击版本对比，并在弹窗中选择需要比对的版本，可查看比对信息。
- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前工作流至配置前状态，工作流配置信息将不再保留，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前工作流发布历史。

4.6 使用工作流

4.6.1 通过 API 调用工作流

工作流发布成功后，可以使用API调用该工作流。

当您需要将开发好的工作流集成到外部业务系统（如企业网站、公众号后台）或通过脚本进行批量自动化处理时，可以通过调用API的方式使用工作流。

费用说明

通过API调用工作流将消耗资源（如Token用量），可能产生相关费用。请在调用前确认账户额度充足，具体计费规则请参考**计费项**。

前提条件

- 须确保工作流已发布，详情可参考**发布并部署工作流**。

版权所有 © 华为云计算技术有限公司

获取调用路径

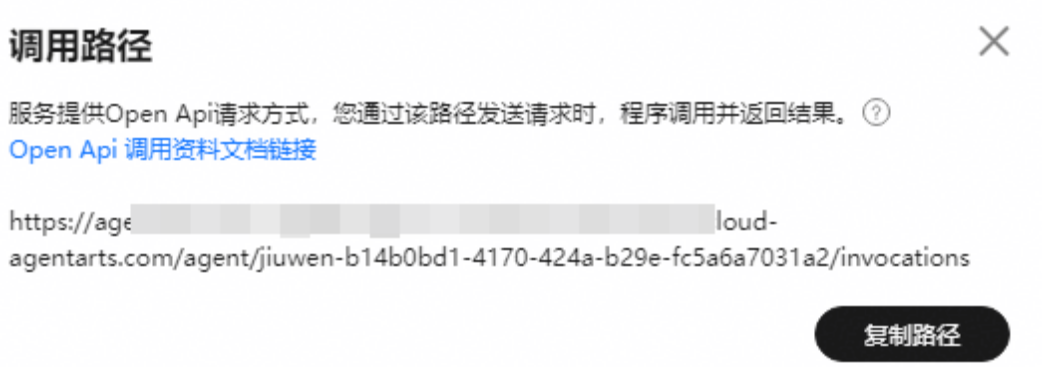
步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 智能体管理”，单击左上角“工作流”页签，进入工作流应用管理界面。

步骤3 单击选择“调用路径”。

步骤4 在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如图4-56所示。

图 4-56 获取应用调用路径



----结束

使用 API 调用工作流

工作流发布为API服务之后，可通过API调用工作流应用。详情请参考使用API调用工作流。

4.6.2 在单智能体应用中使用工作流

工作流功能是实现单智能体应用业务逻辑的核心部分。它定义了应用的输入和输出数据结构、数据接收与处理规则，以及决策流程。通过工作流，单智能体应用能够支持添加工作流技能，允许用户通过画布编排的方式，组合使用插件、大模型等不同节点，从而实现复杂且稳定的业务流程编排。

- **简单业务**：应用中至少有一个**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且由多个子任务组成，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。

费用说明

工作流运行过程中调用大模型或插件会产生相应的资源消耗与费用，请确保账户额度充足。

前提条件

- 须确保工作流已发布，详情可参考发布并部署工作流。

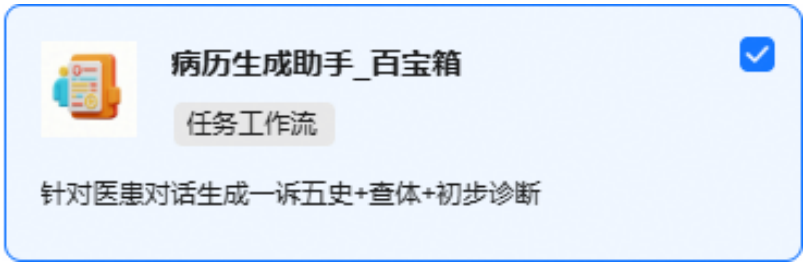
版权所有 © 华为云计算技术有限公司

在单智能体应用中配置工作流

步骤1 在“技能 > 工作流”模块，单击 + 。

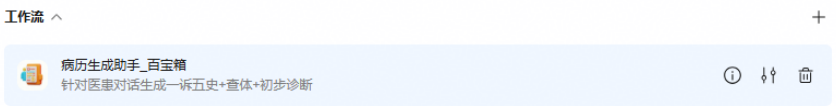
步骤2 在“添加工作流”窗口，单击 + 进行添加，再单击“确定”。

图 4-57 添加工作流



步骤3 添加工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 4-58 已添加工作流



----结束

常见问题

1. 用户在“添加工作流”窗口中可能找不到已发布的工作流?

   如果在添加窗口中找不到目标工作流，请检查该工作流是否已发布成功，且当前账号拥有该工作流的查看权限。

相关文档

单智能体应用中使用工作流示例，请参考**添加工作流**。

4.6.3 在多智能体应用中使用工作流

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。

前提条件

- 须确保工作流已发布，详情可参考**发布并部署工作流**。

在多智能体应用中配置工作流

步骤1 在“多Agent控制器”配置页面，支持添加意图识别、子智能体、全局意图、起始工作流、默认工作流和结束工作流，可根据业务需求选择对应的工作流。

版权所有 © 华为云计算技术有限公司

步骤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 4-59 添加工作流



说明

- 起始工作流：在会话开始时首先执行，常用于初始化参数或欢迎语引导。
- 意图识别与子智能体：根据用户输入识别意图，分流至对应的子工作流（子智能体）处理；例如多技能助手场景，需配置“意图识别”及多个“子智能体”，并建议配置“默认工作流”处理未知问题。
- 默认工作流：当无法识别用户意图时的兜底处理流程；例如简单问答场景，仅需配置“默认工作流”处理通用问答。
- 结束工作流：会话结束时的收尾处理，如数据归档或满意度调查。

步骤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 4-60 已添加工作流



 版权所有 © 华为云计算技术有限公司

步骤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”。

图 4-61 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 4-62 编排后包含工作流的多智能体应用



----结束

相关文档

多智能体应用中使用工作流示例，请参考**创建并配置多智能体应用**。

4.7 管理工作流

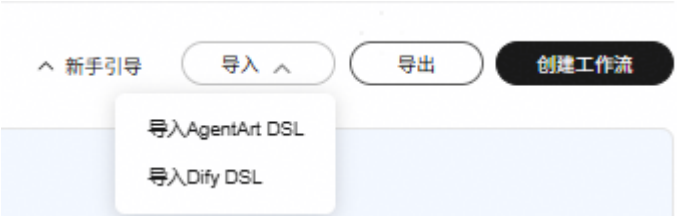
AgentArts平台致力于提供高效、灵活的工作流管理解决方案。除了支持**AgentArts工作流**的导入导出，为了提升用户体验，AgentArts现支持**Dify工作流**的导入功能。用户可以轻松将现有的Dify工作流文件导入平台，在导入工作流时，将同步导入与工作流关联的插件等配置，实现无缝对接和高效管理。无论是复杂业务流程还是简单任务自动化，AgentArts都能快速导入、配置和执行Dify工作流，显著提高工作效率和操作便捷性。

导入、导出工作流

步骤1 导入工作流。

- 导入AgentArts DSL文件
  a. 单击页面左上角“导入”下拉框，并选择“导入AgentArts工作流”。

  图 4-63 导入 AgentArts 工作流

  b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件，并单击“导入”。

版权所有 © 华为云计算技术有限公司

图 4-64 导入工作流



图 4-65 导入状态



说明

- 如果工作流关联的资源（如插件、MCP、知识库）导入失败，应用仍会正常导入，但这些资源将不会被导入。

  Dify不支持知识库和自定义插件导出：由于Dify DSL文件结构限制，不包含知识库向量数据及自定义插件源码，因此无法自动迁移，需用户在AgentArts平台手动重建。

- 如果导入失败，可在操作列查看原因。
- 如果导入的工作流中存在资源未配置鉴权，系统会标注“未配置鉴权”，单击操作列中的“配置鉴权”后可在配置鉴权窗口中进行操作。

c. 选择导入文件后，选择解析内容。

平台将自动解析jsonl文件。如果解析的文件已存在，勾选该文件将自动覆盖平台现有文件。

版权所有 © 华为云计算技术有限公司

说明

- 仅支持上传jsonl格式文件，工作流文件的最大导入大小为128MB。
- 勾选已存在的文件将覆盖平台现有工作流配置，该操作不可恢复，请谨慎操作。

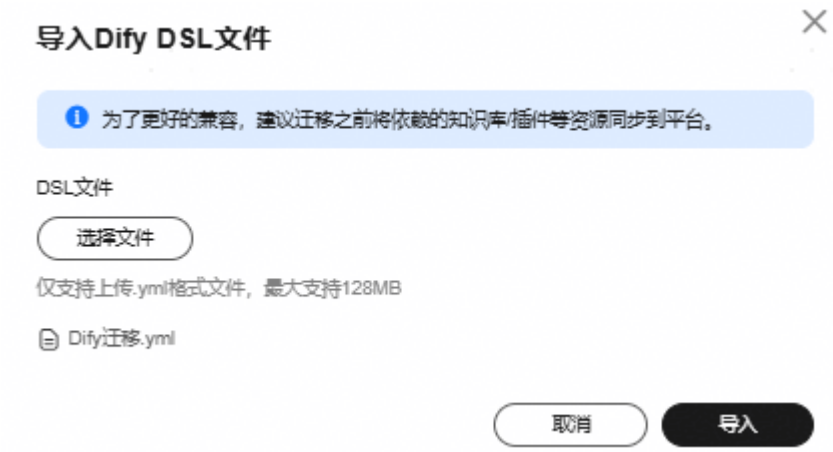
d. 单击“导入”，导入成功的工作流将在“开发中心 > 应用管理 > 工作流应用”页面中展示。

- 导入Dify DSL文件

a. 单击页面左上角“导入”下拉框，并选择“导入Dify DSL文件”。

b. 在“导入Dify DSL文件”页面，单击“选择文件”选择需要导入的yml文件，并单击“导入”按钮。

图 4-66 导入 Dify DSL 文件



说明

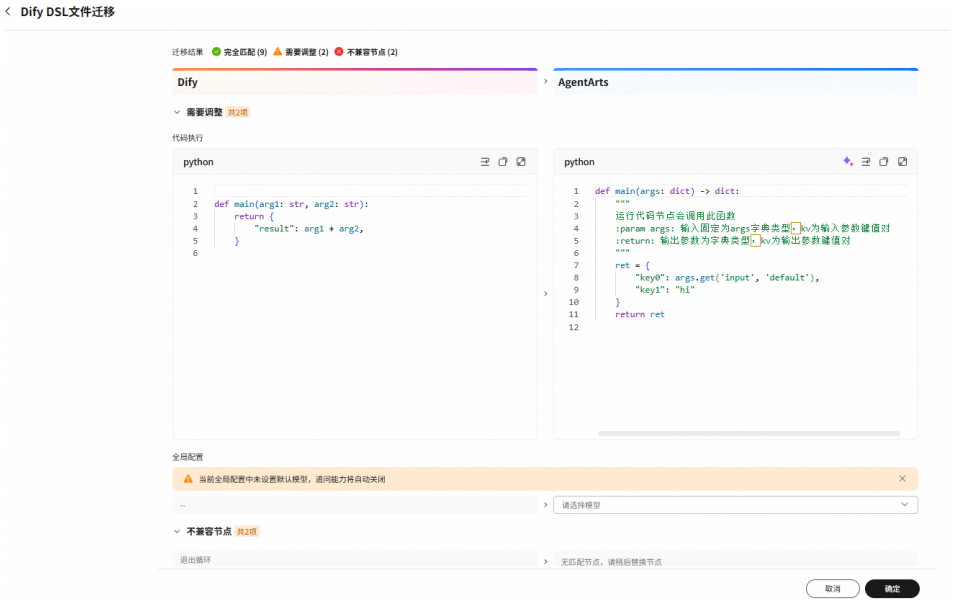
仅支持上传.yml格式文件，最大支持128MB。

为了更好地兼容，建议迁移之前将依赖的知识库/插件等资源同步到平台。

c. 在Dify DSL文件迁移界面确认迁移结果，单击“确定”后工作流成功导入，可在工作流配置界面对需要调整和不兼容的节点或模型等进行修改。

版权所有 © 华为云计算技术有限公司

图 4-67 Dify DSL 文件迁移



说明

- 对于不兼容的节点新增空白节点（保留Dify工作流中的输入和输出参数），可在同步数据完成后在工作流编排页面进行节点替换。
- 针对引用插件、MCP、模型等可根据需要重新配置鉴权。

步骤2 导出工作流。平台支持导出工作流。导出工作流时，将同步导出工作流关联的插件等配置。

1. 单击页面右上角“导出”，选择目标工作流，导出工作流。或鼠标移动到目标工作流，单击“更多 > 导出”。
2. 在“导出工作流”页面选择工作流，单击“导出”。单击“下载”，工作流将以一个jsonl格式的文件下载至本地。

图 4-68 导出工作流



版权所有 © 华为云计算技术有限公司

说明

- 如果应用关联的资源（如插件、MCP、知识库）导出失败，则导出的文件中将不包含对应资源。
- 如果导出失败，可在操作列查看原因。

----结束

注意事项

- Dify不支持知识库和自定义插件导出。如果Dify工作流中包含这些资源，用户需要手动在AgentArts中创建相应的知识库和自定义插件。详细操作请参见**创建知识库**和**创建插件**。
- Dify的迁移功能基于Dify DSL 0.6.0版本（产品版本1.12.0）开发。请注意以下几点：
  - **版本兼容性**：不同DSL版本之间可能存在较大的结构差异，这可能导致存量数据在迁移过程中出现不兼容问题。
  - **节点处理**：不在节点列表中的Dify节点将默认转换为空白节点。
  - **未来升级**：Dify未来的DSL版本升级可能会影响迁移功能的兼容性。如果发现严重的兼容性问题，将在后续版本中进行调整和优化。

表 4-25 Dify 工作流转换支持情况说明

Dify	AgentArts	AgentArts兼容性
系统变量	系统参数	以下系统参数在导入AgentArts时暂不支持： • **会话次数**：dialog_count。 • **应用ID**：app_id。 • **工作流ID**：workflow_id。 • **工作流运行ID**：workflow_run_id。 • **运行开始时间戳**：timestamp。
对话型工作流	全局配置	• **内容审查**：Dify对话型工作流导入AgentArts时，内容审查不支持API拓展。 • **语音转文字和文字转语音**：语音转文字和文字转语音功能在AgentArts中默认开启。 • **引用和归属配置**：AgentArts工作流不支持引用和归属配置。
任务型工作流触发器	开始节点+触发器	AgentArts任务型工作流不支持WebHook类型触发器。
用户输入或开始	开始	• **用户输入查询**：Dify的userinput.query对应AgentArts开始节点的query。 • **文件输入**：Dify的userinput.files在AgentArts的开始节点中新增一个sysFiles输入参数，类型为file/default。

版权所有 © 华为云计算技术有限公司

Dify	AgentArts	AgentArts兼容性
LLM	大模型	• **模型选择**：转换后，模型将被置空，需要重新选择。 • **视觉输入参数**：视觉输入参数需要重新配置。 • **异常场景返回**：AgentArts不支持在未开启结构化输出时配置异常场景的预设返回内容。 • **记忆提示词和记忆窗口**：AgentArts不支持配置记忆提示词和记忆窗口。 • **失败重试**：AgentArts不支持失败时的重试功能。
知识检索	知识检索	• **知识库迁移**：知识库不支持迁移，需要重新选择。 • **元数据过滤**：AgentArts不支持元数据过滤。 • **参数适配**： – 分段标题title对应AgentArts的subtitle。 – 分段链接url、分段图标icon、其他元数据metadata以及files召回的文件参数AgentArts不支持。
对话型-直接回复	消息节点 结束节点	基本兼容
Agent	Agent	AgentArts仅保留名称和描述，其余参数需要用户重新配置。
问题分类器	意图识别	• **模型选择**：转换后，模型将被置空，需要重新选择。 • **视觉输入参数**：AgentArts不支持视觉输入参数。 • **意图分类**：AgentArts意图分类不支持引用前置节点参数。
迭代	循环	• **错误响应方法**：AgentArts不支持错误响应方法。 • **扁平化输出**：AgentArts不支持扁平化输出。
循环	循环	• **终止条件**：AgentArts终止条件存在不兼容，不支持“开始是”和“结束是”。 • **退出循环节点**：AgentArts不支持退出循环节点。 • **退出条件**：AgentArts退出条件不支持选择循环内部节点。
代码执行	代码	**代码执行**：使用本地代码替换，用户需要根据提供的模板重新适配代码。
模板转换	无	AgentArts不支持模板转换，将转换为空白节点。

版权所有 © 华为云计算技术有限公司

Dify	AgentArts	AgentArts兼容性
变量聚合器	变量聚合	基本兼容
文档提取器	无	AgentArts不支持文档提取器转换，将转换为空白节点。 可使用插件市场的文档解析工具替换。
变量赋值	变量设置	**运算符**：AgentArts不支持 +=、-=、*= 、/=、追加、扩展和移除等运算符。
参数提取器	大模型节点	• **模型选择**：转换后，模型将被置空，需要重新选择。 • **推理模式**：AgentArts不支持推理模式。 • **记忆窗口**：AgentArts不支持记忆窗口。
HTTP请求	HTTP请求	• **SSL证书验证**：AgentArts不支持验证SSL证书配置。 • **超时设置**：AgentArts不支持超时设置。 • **输出变量**：AgentArts输出变量不支持files文件列表。 • **失败重试**：AgentArts不支持失败时重试。 • **异常分支**：AgentArts不支持配置异常分支。 • **请求体格式**：AgentArts中的body不支持form-data、x-www-form-urlencoded、raw和binary格式。 • **请求方法**：仅支持GET和POST请求方法。
列表操作	无	AgentArts不支持列表操作转换，将转换为空白节点。
工具	插件或MCP	AgentArts不支持工具转换，将转换为空白节点。可根据业务需要在AgentArts组件库中进行创建，详细步骤请参见**创建插件**或**创建MCP**。
注释	注释	基本兼容
条件分支	判断	• **判断条件子项**：AgentArts不支持判断条件子项进行判断。 • **条件判断**：AgentArts不支持条件判断中的“开始是”和“结束是”。
结束	结束	**多个结束节点**：当Dify中存在多个结束节点时，转换后只保留一个结束节点，该结束节点将包含所有结束节点的输出参数。

版权所有 © 华为云计算技术有限公司

表 4-26 规格约束

规格	Dify	AgentArts
输入/输出参数（名称、值、描述）	参数名称长度限制无，值长度限制无，描述限制无。	参数名称长度限制64字符，值长度限制100000字符，描述长度限制100000字符。
系统变量	• **user_id**：String，用户ID。 • **app_id**：String，应用ID。 • **workflow_id**：String，工作流ID。 • **workflow_run_id**：String，工作流运行ID。 • **timestamp**：Number，应用开始运行的时间戳。 • **conversation_id**：String，会话ID。 • **dialog_count**：Number，会话次数。	• **conversation_history**：Array<Object>，会话历史。 • **current_time**：String，当前系统时间。 • **user_id**：String，用户ID。 • **conversation_id**：String，会话ID。 • **dialogue_count**：integer，会话轮次。
注释的长度限制	无	100000字符
工作流节点的数量限制	无	无
节点中输入/输出参数的个数限制	无	无
判断节点的分支数量	无	无
节点名称	无	100000字符

工作流相关操作

AgentArts同时支持对工作流执行复制、获取工作流ID、调用路径、删除、渠道管理等操作。

表 4-27 工作流相关操作

类型	操作
复制	鼠标移动到目标工作流，单击“复制”，即可复制该工作流。

版权所有 © 华为云计算技术有限公司

类型	操作
获取工作流ID	工作流应用除了具有页面操作的能力之外，还具有Chat API调用能力，因此获取AppID十分必要。该ID为调用Agent应用接口的agent_id字段。 POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id} 鼠标移动到目标工作流，单击“复制ID”，可获取当前工作流ID。 弹出复制成功对话框，用于填写调用Agent应用接口的agent_id字段。 **图 4-69** 复制 ID 
调用路径	鼠标移动到目标工作流，单击“调用路径”，调用路径为工作流的API接口。详细API调用过程请参见**通过API调用工作流**。 **图 4-70** 获取调用路径 

 版权所有 © 华为云计算技术有限公司

类型	操作
删除	**注意** 如果工作流已被引用，删除后引用将被自动取消，可能会导致其他工作流或智能体无法运行，且该操作不可撤回，请谨慎操作。 鼠标移动到目标工作流，单击“更多 > 删除”，在弹出的对话框中单击“确定”。 • 如果工作流未被引用在弹窗中单击“确定”即可。 **图 4-71** 工作流未被引用  • 如果工作流被引用则删除后引用将自动取消，可能会导致工作流或应用无法运行，且该操作不可撤回。 **图 4-72** 工作流被引用 
渠道管理	鼠标移动到目标工作流，单击工作流下方的“渠道管理”。 跳转至“发布管理”页面。详细操作可参见**发布并部署工作流**节点。

版权所有 © 华为云计算技术有限公司

类型	操作
查看工作流运行时数据	• 在“智能体运行时”页面，通过搜索智能体ID搜索查看工作流运行时数据。 1. 登录**AgentArts智能体平台**。 2. 在左侧导航选择“部署运行 > 智能体运行时”，在搜索框中输入上方已获取的ID进行搜索。 3. 单击搜索到的工作流名称，进入详情页面查看工作流运行时数据。 • 在“运营运维 > 观测 > 智能体运行分析”页面，通过搜索智能体ID查看工作流运行时数据。 1. 登录**AgentArts智能体平台**。 2. 在左侧导航选择“运营运维 > 观测 > 智能体运行分析”，在搜索框中输入上方已获取的ID进行搜索。 3. 单击搜索到的工作流名称，进入详情页面查看工作流运行时数据。

4.8 基础节点

4.8.1 开始节点

开始节点用于触发一个工作流，结束节点用于返回一个工作流的最终结果。

开始节点

开始节点是工作流的“感知边界”与“数据入口”。它定义了工作流启动时需要接收哪些外部信息（是接收一段文字、一个文件，还是一个特定的参数）。

- 节点定位：它决定了工作流如何理解用户的意图。如果开始节点定义不清晰，后续的大模型节点或其他节点将无法正常传参。
- 默认参数：系统预置了query参数，表示用户在本轮对话中输入的原始内容。您也可以按需添加其他参数，用于下游节点的输入；同时开始节点也支持选择导入**对象模板**。

  说明

  开始节点支持的最大输入字符串长度为10万，超出此限制后系统将报错。如果搭建长文本处理、HTML清洗、数据清洗等工作流，请注意输入长度限制，防止工作流报错AgentArts.02001003。

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、Boolean、Object、File、Array多种类型的输入参数。对于复杂的业务参数（如用户信息），推荐使用Object（支持5层嵌套）以保持数据结构的整洁。

文件作为输入参数时，用户可增加“文件”或“文件数组”类型，并在试运行界面中上传文件。

版权所有 © 华为云计算技术有限公司

- 文件类型：“参数类型”选择“File”，并选择对应的文件类型，一个参数仅支持上传一个文件。
- 文件数组类型：“参数类型”选择“Array<File>”，并选择对应的文件类型，一个参数最多支持上传10个文件。
- 支持上传Default、Doc、Txt、Excel、PPT、Image、Audio、Video等格式的文件。

参数描述：

参数的描述信息。**极重要！**当工作流绑定到智能体时，大模型通过此描述决定如何从用户对话中自动提取信息。

是否必选：

参数是否必选。如果设为必选，缺失该参数将无法运行工作流。

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，然后再进行赋值。示例如下图所示。

图 4-73 开始节点 Object 类型参数嵌套



版权所有 © 华为云计算技术有限公司

图 4-74 开始节点 Object 类型参数



常见问题

- 开始节点输入长度有限制吗?

  开始节点支持的最大输入字符串长度为10万，超出此限制后系统将报错。如果搭建长文本处理、HTML清洗、数据清洗等工作流，请注意输入长度限制，防止工作流报错AgentArts.02001003。

- 为什么在后续节点无法引用自定义的变量?

  请检查变量类型是否一致，变量名称是否冲突，或者开始节点新增参数后是否单击了“确定”对参数进行保存。

4.8.2 结束节点

结束节点是工作流的“交互出口”。它决定了工作流执行完一系列复杂逻辑后，最终以何种形式将结果反馈给用户或下游系统。

- 对话型反馈：直接回复一段自然的文字。例如：为您找到的结果是{{result}}。
- 任务型反馈：返回结构化的数据（JSON），让工作流的返回结果从“自然语言对话”升级为“程序可调用的标准化数据格式”，满足系统对接、跨流程调用、数据复用等技术场景的需求。

结束节点

输入参数:

版权所有 © 华为云计算技术有限公司

用于接收上游节点传递过来的数据。支持引用上游节点的输出，或者输入一个常量值。

该参数无法直接返回，需在“指定回复”中以{{variable_name}}形式插入才会体现在最终结果里。

系统参数列举如下：

- “conversation_history”：存储当前对话的历史消息记录。
- “current_time”：获取当前的系统时间戳或时间字符串。
- “user_id”：识别当前用户的唯一标识符。
- “conversation_id”：当前工作流的唯一对话标识。
- “dialogue_count”：当前对话的轮次数计数。

输出参数：

结束节点的输出参数是专门用于对外以变量形式返回数据的配置项，同样支持引用上游数据和手动输入，工作流运行结束后所有输出参数会统一以JSON格式聚合返回，这类格式适配子工作流调用的场景，能让子工作流直接解析使用该数据。

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。

同时该参数可引用工作流全局配置的记忆变量，但有严格的使用限制，绝对不能在“指定回复”中插入或引用。

指定回复：

指定回复是工作流最终结果的可视化展示内容编辑区，用户可在该文本框中自定义撰写回复的文字内容，也能将已配置的输入参数以{{variable_name}}的形式插入到文本中，实现动态数据与固定文案的结合。

指定回复支持单击 $\{x\}$ 快速插入输入参数，或在指定回复输入框中键入“{{}}”快速引用参数。

指定回复中不能插入输出参数。

结构化信息：

功能开启时，可使用{{_NODE_OUTPUT}}引用“指定回复”的全部内容，是自然语言的结构化封装。

说明

结束节点的**输出参数**不允许引用大模型节点的raw_output参数。

版权所有 © 华为云计算技术有限公司

图 4-75 结束节点



相关操作

- 双击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

问题：报错“response_template类型错误、为空或引用异常”，怎么解决?

版权所有 © 华为云计算技术有限公司

图 4-76 结束节点报错



回答：检查变量引用是否有效（最常见原因）。如图4-75所示，“指定回复”中引用的参数没有在“输入参数”中定义，需将输入参数名称改为result，或将指定回复的引用参数改为{{accountName}}。

4.9 通用节点

4.9.1 大模型

大模型节点是工作流的“大脑”。它不仅能进行对话生成，还能基于上下文进行逻辑推理、任务拆解和非结构化数据提取。通过配置提示词和模型参数，可以控制其输出风格与格式，为下游节点提供决策依据或结构化数据。

核心能力：

- 内容生成：撰写文案、代码、翻译、创意写作。
- 逻辑决策：分析用户意图，判断下一步调用哪个插件。
- 信息提取：从自然语言中提取JSON格式的关键字段（如时间、地点、金额）。

约束与限制

- 如果在大模型节点后并行使用多个大模型，应将首个大模型节点配置为非流式输出。
- 大模型不支持失败时重试。

版权所有 © 华为云计算技术有限公司

图 4-77 示例



费用说明

大模型节点调用会消耗Token并产生费用。不同模型计费单价不同，具体计费规则请参考**计费项**。

配置大模型节点

配置大模型节点的详细参数说明如下：

模型配置

选择要使用的模型。此节点的输出内容质量很大程度上受模型能力的影响，建议根据实际业务场景选择模型。

您还可以单击图标，调整模型配置。模型配置的详细说明可参考**选择并配置模型**。

输入参数

输入参数定义了从前置节点传入的变量，在提示词中通过{{参数名}}引用。默认参数名为query。

当单击图标时，可新增输入参数，当单击图标时，可删除输入参数。

- 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：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置），以及系统参数。

    系统参数列举如下：

    - “conversation_history”：存储当前对话的历史消息记录。
    - “current_time”：获取当前的系统时间戳或时间字符串。
    - “user_id”：识别当前用户的唯一标识符。
    - “conversation_id”：当前工作流的唯一对话标识。
    - “dialogue_count”：当前对话的轮次数计数。

版权所有 © 华为云计算技术有限公司

- 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。

  说明

  当参数类型为引用时，设置参数值时，您可以在下拉框顶部的搜索栏中输入关键词，快速定位所需参数。

- 描述：（可选）用于说明该输入参数的具体含义。

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。

- **输出格式**：指定此节点输出的内容格式。输出格式支持设置为：
  - **文本**：纯文本格式。仅包含一个文本类型的输出变量，默认名称为raw_output（可自定义）。该变量直接承载模型生成的全部内容。适用于通用对话、简单文本生成场景。
  - **Markdown**：Markdown格式。仅包含一个文本类型的输出变量，默认名称为raw_output（可自定义）。该变量承载模型生成的Markdown格式文本。适用于需要富文本格式展示的场景。
  - **JSON**：标准JSON格式。支持定义多个结构化的输出变量（如price、location）。您需要预先定义好期望的输出参数，节点会自动尝试解析模型的回复，并将提取到的信息填充到对应的输出参数中。适用于下游需要结构化数据的场景（如字段提取、API入参构造）。
- **自定义参数**：
  - 参数名称：名称请命名清晰易理解，有助于大模型正确返回匹配的内容。例如，使用new_query表示改写后的问题，而不是简单的output1。

    说明

    大模型节点内置raw_output，表示原始输出，后续节点可直接引用，用户自定义输出参数不可与其同名。

  - 参数类型：参数类型请确保大模型实际输出的内容严格匹配。例如，如果类型定义为Integer（整数），模型却输出了“high”或3.14，下游程序可能会报错或执行异常。
  - 参数描述：清晰的描述就是给模型的明确指令，模型会参考此描述来理解每个字段应提取什么内容。请尽量写清楚含义。例如“用户的购买地点”而不是“地点”。
- **是否流式输出**：流式输出控制模型是逐字返回还是一次性返回结果。仅文本、Markdown格式支持流式输出。

提示词配置

- 短期记忆：用于控制大模型是否读取多轮对话的历史交互内容，开启时可确保多轮对话连贯性。支持通过单击右侧的开关按钮“启动”或“关闭”短期记忆功能，该功能默认关闭。
- 系统提示词：

  配置输入给大模型的提示词，系统级提示词，用于指导模型按要求进行回复。撰写提示词方式可参考**撰写提示词规范**。

  支持单击 {x} 快速插入当前节点的输入参数，或在输入框中键入“{{}}”快速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。例如：

 版权所有 © 华为云计算技术有限公司

需要让大模型能够根据用户输入的城市，预算以及旅行天数提供旅行建议的场景，输入参数分别定义为city，budget以及days，分别引用“开始”节点已定义好的city，budget、days参数，提示词中引用这些输入参数，如图4-78所示。

图 4-78 配置系统提示词



用户可单击 图标直接引用提示词模板；也可以自行撰写提示词，撰写后可单击 图标对系统提示词进行智能优化；也可单击 图标保存到模板，在“组件库>提示词”中可查看该模板。

- 用户提示词：

  用户直接输入给大模型的具体问题、指令或需求，例如，“你觉得未来的科技会怎样发展？”。支持单击 {x} 快速插入当前节点的输入参数，或在输入框中键入“{{}}”快速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。

  用户提示词的质量直接决定了大模型输出结果的好坏，提示词越清晰，AI的回答越准确。

 版权所有 © 华为云计算技术有限公司

当单击  图标时，系统会弹出“选择提示词”窗口，可在“提示词广场”或“组件库提示词”页签中进行选择；同时支持单击快速插入当前节点的输入参数，或在输入框中键入“{{}}”快速引用当前节点输入参数中已定义好的参数。

异常处理

支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。

- “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900秒。
- “重试次数”：大模型暂不支持失败时重试。
- “异常处理方式”：配置异常处理方式。
  - 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。
  - 返回设定内容：节点发生异常后，工作流不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。
  - 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点后新增节点，并为新增的异常分支配置相应的处理流程。

说明

- 当流式输出和异常处理功能开启时，异常处理参数默认为“不重试”和“中断流程”。
- 当流式输出功能关闭，输出格式为文本或Markdown时，异常处理方式仅支持“中断流程”。
- 当流式输出功能关闭，输出格式为JSON时，三种异常处理方式均支持。

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击  图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击  图标，可查看文档信息。

配置示例

例如通过大模型节点自动从用户输入的一段文本中提取关键信息，开始节点接收用户的文本输入，大模型节点将用户的输入提取出关键字段，例如购买地点、购买产品和价格，最后结束节点输出这些字段。工作流编排方式如**图2 工作流**所示。

版权所有 © 华为云计算技术有限公司

图 4-79 工作流



工作流核心节点如表4-28所示。

表 4-28 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点	此节点读取开始节点中用户提供的输入文本，大模型会根据提示词要求，在输出参数中提取相关字段。 • **模型配置**：选择合适的模型。 • **输入参数**：引用开始节点的query。 • **输出参数**：将需要提取的字段，在输出参数中定义，并在描述中说明提取的字段含义，便于大模型理解。 单击输出区域右上角的“ ”，选择输出格式为“JSON”。单击“ ”，添加参数。 – location：类型为“String”，描述为购买地点。 – product：类型为“String”，描述为购买产品。 – price：类型为“String”，描述为价格。 • **提示词配置**： – 系统提示词：定义一段系统提示词，如下所示： 角色/指令：你是一个信息提取器，请从文本中提取：购买地点、购买产品和价格。输出格式要求：请以json格式输出，如果原文未提及，值设为“未知”。 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
结束节点	此节点用于输出和展示提取的字段信息。 • 输入参数：单击“＋”，添加如下参数。 – location：引用大模型节点的location。 – product：引用大模型节点的product。 – price：引用大模型节点的price。 • 指定回复： 购买地点：{{location}} 购买产品：{{product}} 购买价格：{{price}}	

效果示例如下图所示。

图 4-80 效果示例

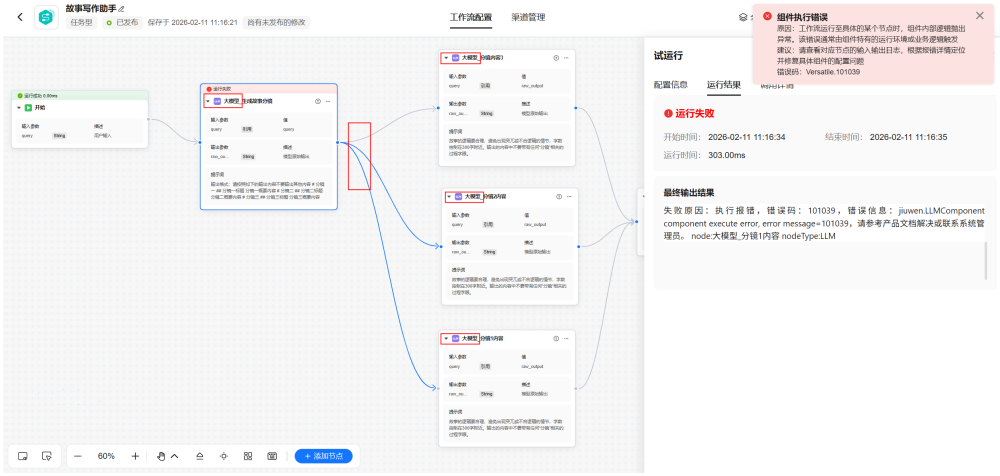


常见问题

- **将工作流中的第一个大模型节点设置为流式输出时，为什么会报错101039?**

版权所有 © 华为云计算技术有限公司

图 4-81 报错示例截图

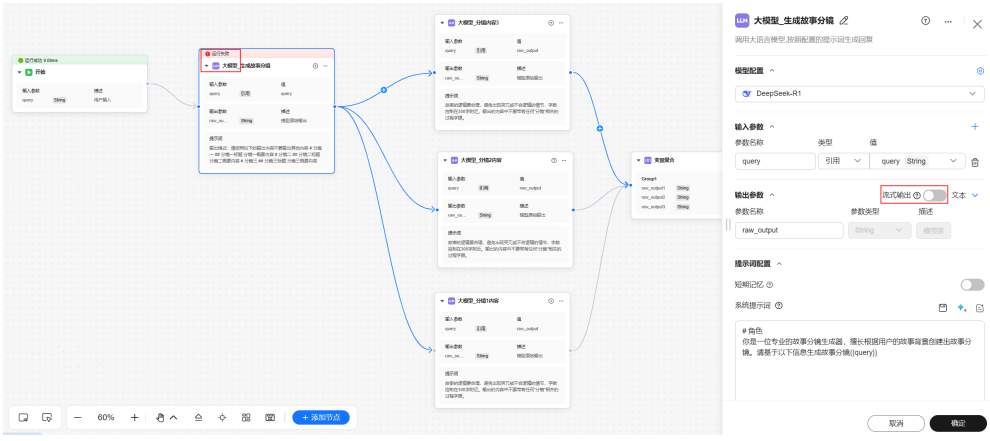


原因：如果在大模型节点后并行使用多个大模型，应将首个大模型节点配置为非流式输出，流式输出模式可能会导致数据流在并行处理时出现不一致或冲突，从而引发错误101039。详见**约束与限制**。

解决办法：

- 将第一个大模型节点的输出模式设置为非流式输出。确保配置文件中的“output_mode”参数设置为“batch”或其他非流式模式。
- 确认后续并行使用的大模型节点是否正确配置。确保每个节点的输入和输出格式一致，避免数据流冲突。
- 在测试环境中重新运行工作流，确保修改后的配置能够正常工作。逐步验证每个节点的输出，确保数据流的正确性和一致性。
- 查看模型和框架的官方文档，确认是否有关于流式输出和并行处理的特定要求或限制。
- 查看详细的错误日志，寻找更多关于错误101039的上下文信息，以便进一步诊断问题。

图 4-82 正确配置示例



- **为什么模型输出了JSON但后续节点无法识别？**
  - 解决方式一：确认节点“输出格式”是否已选为JSON。如果是“文本”模式，下游只能拿到字符串，无法取Key值。

版权所有 © 华为云计算技术有限公司

- 解决方式二：检查提示词是否严格约束了“不要使用Markdown代码块”。某些模型习惯输出```json {...}```，这会导致JSON解析失败。建议在提示词中加入“仅返回原始JSON”。

- **JSON输出解析失败，下游拿到的是字符串而非结构化对象**

检查项	操作
输出格式是否为JSON	确认已在输出区域选择“JSON”而非“文本”。
流式输出是否关闭	JSON格式要求关闭流式输出。
Prompt是否约束了格式	在系统提示词中明确写入不要包含```json标记，直接返回JSON。
字段定义是否完整	确认输出参数中的字段名与Prompt中要求的字段名一致。

4.9.2 工作流

在复杂的业务编排中，**工作流节点**允许您将另一个已发布的工作流作为“子程序”嵌入当前流程。

通过嵌套，您可以将常用的、标准化的任务处理流程（如“用户鉴权”、“发票OCR识别”）封装为不同的子工作流，并在主工作流中调用这些子工作流。通过工作流嵌套，可以实现复杂任务的模块化拆分和处理，从而使工作流编排逻辑更加灵活、清晰和易于管理。

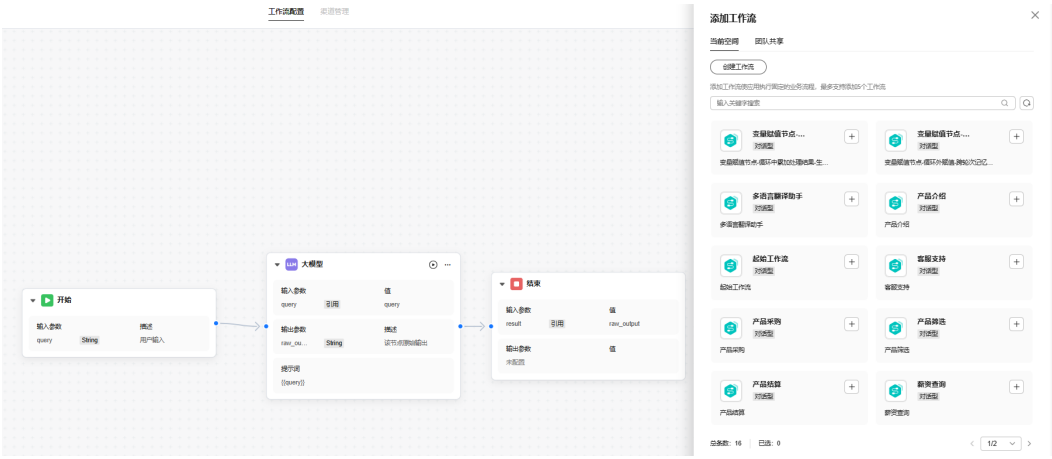
费用说明

运行工作流节点将触发子工作流的完整执行，会产生相应的资源消耗和费用（如Token消耗、并发实例占用）。请合理配置超时时间以避免资源浪费。

添加节点

在工作流编辑页面，单击下方“添加节点”，选择“工作流”节点，添加所需的已发布工作流。

图 4-83 添加工作流节点



版权所有 © 华为云计算技术有限公司

配置工作流节点

配置工作流节点的详细参数说明如下：

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，如需调整，请前往子工作流的开始节点进行修改并重新发布。
- 在工作流节点中您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，如需调整，请前往子工作流的结束节点进行修改并重新发布。
- 每个工作流节点都会自动包含一个名为response_content的固定输出参数，是子工作流结束节点的默认输出内容。

> **注意**
>
> 如果子工作流的输入或输出定义发生变更（如新增必填项），必须在主工作流中**刷新**该节点并补充配置，否则运行会报错。

图 4-84 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）



异常处理

支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。

“超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。

“重试次数”：工作流节点不支持重试。

“异常处理方式”：配置异常处理方式。

版权所有 © 华为云计算技术有限公司

- 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。
- 返回设定内容：节点发生异常后，工作流不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。
- 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点后新增节点，并为新增的异常分支配置相应的处理流程。

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可查看子工作流详情，重命名节点名称，复制或删除节点；也可以将节点连接到结束节点。
- 单击图标，可跳转至子工作流，对嵌入的子工作流的当前版本进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

将“智能天气助手”封装为一个子工作流，即用户询问城市天气信息，系统自动查询该城市的天气信息，并生成一段穿搭建议。该工作流未来可以在多个主工作流中复用（例如旅行助手、日程规划助手都可能调用这个子流程）。

以一个简单的工作流嵌套的示例为例，旨在熟悉工作流节点的使用。

- 主工作流如图4-85所示

图 4-85 主工作流全景图



主工作流核心节点如表1 节点配置说明所示。

表 4-29 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	

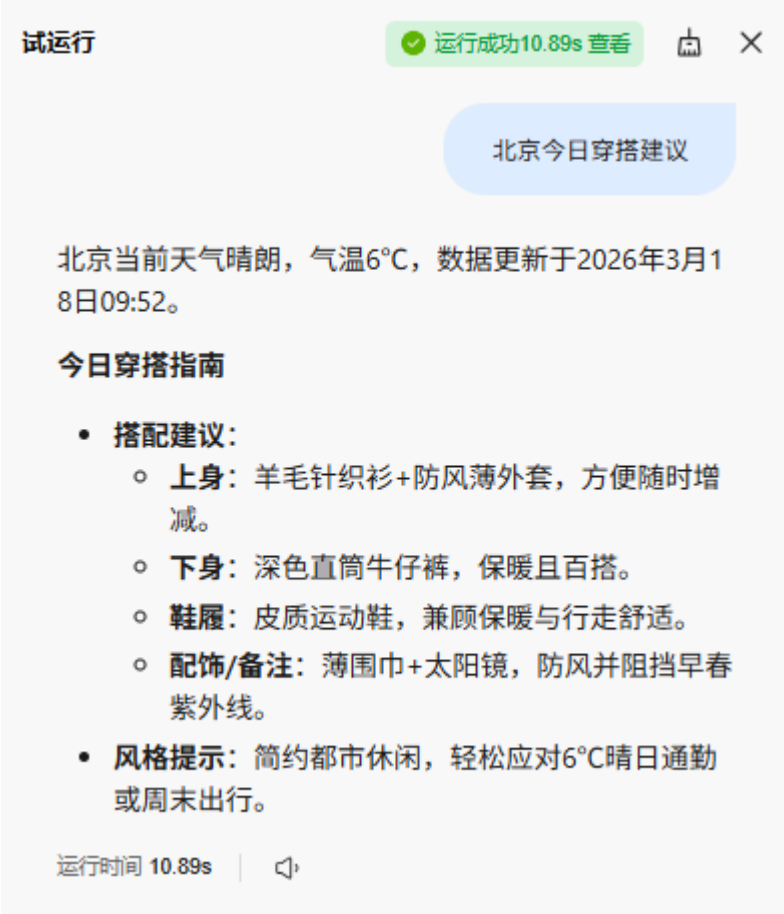
版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
工作流节点-智能天气助手	此节点用于根据用户的输入，运行子工作流，查询对应城市的天气，并给出穿搭建议。 **输入参数：** query，引用开始节点的query。	
结束节点	此节点用于输出“智能天气助手”工作流节点的最终结果。 **输入参数：** result：引用“智能天气助手”工作流节点的response_content。	

工作流运行的效果如图4-86所示。

版权所有 © 华为云计算技术有限公司

图 4-86 效果示例



常见问题

- **子工作流修改后，主工作流会自动更新吗?**

  不会。为了保证稳定性，主工作流锁定的是子工作流的**特定版本**。如需应用最新逻辑，请确保如下：

  a. 子工作流的最新修改已发布，详情请参见**发布并部署工作流**。

  b. 在主工作流的工作流编排界面，按“F5”刷新页面，单击工作流节点名称旁边的  ，可升级子工作流至最新版本。

- **为什么工作流节点执行超时?**

  子工作流内部逻辑复杂，或调用的外部API响应慢。请增大超时时间（最大900秒），或优化子工作流内部逻辑。

4.10 逻辑节点

4.10.1 判断

判断节点是工作流中的逻辑路由器。它基于前置节点的输出变量，通过预设的条件表达式（IF/ELSE），将工作流导向不同的执行分支。

版权所有 © 华为云计算技术有限公司

- 需要根据结构化数据做精确判断（如：订单金额 > 1000、状态 == 成功），使用**判断节点**。
- 要理解用户自然语言的意图（如：我想退款）→ 使用**意图识别节点**。

配置判断节点

当向该节点输入参数时，节点会依次检查各分支的条件，并执行第一个满足条件的分支。节点会判断是否符合**IF**区域的条件，符合则执行**IF**对应的工作流分支，否则执行**ELSE**分支。

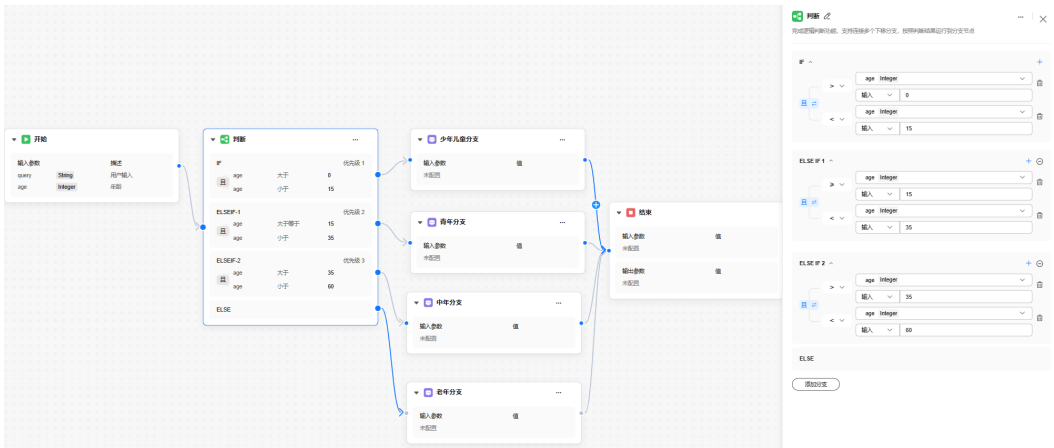
每个分支条件支持添加多个判断条件（且/或），可单击“添加分支”添加多个分支，添加多个分支后的执行顺序：先执行**IF**分支，不满足则执行**ELSE IF1**分支，再不满足则执行**ELSE IF2**分支，以此类推，都不满足则执行**ELSE**分支。

如图4-87所示，这是一个根据年龄判断是哪种群体的工作流。

表 4-30 配置判断节点

分支	条件	执行路径
IF	年龄 ≥ 0 且 < 15	“少年儿童”消息节点
ELSE IF 1	年龄 ≥ 15 且 < 35	“青年”消息节点
ELSE IF 2	年龄 ≥ 35 且 < 60	“中年”消息节点
ELSE	以上条件均不满足，即大于60	“老年”消息节点

图 4-87 配置判断节点



各参数类型支持的比较运算符：

版权所有 © 华为云计算技术有限公司

表 4-31 运算符说明

参数类型	支持的比较运算符
String	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
Integer / Number	大于、大于等于、小于、小于等于、等于、不等于
Boolean	为true、为false
Object	为空、不为空
Array	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

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

注意：删除节点后，该节点所有分支（IF / ELSE IF / ELSE）与下游节点的连接线将同时删除，需重新配置受影响节点的连线。建议删除前先记录各分支的连接关系。

配置示例

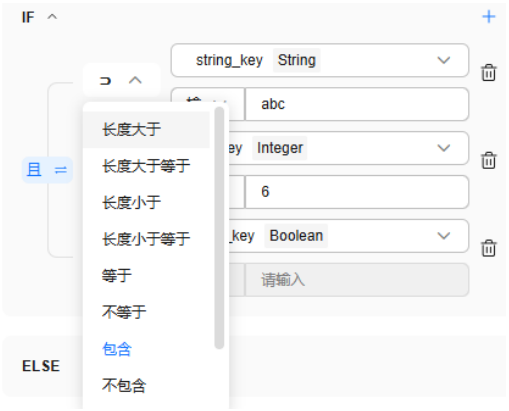
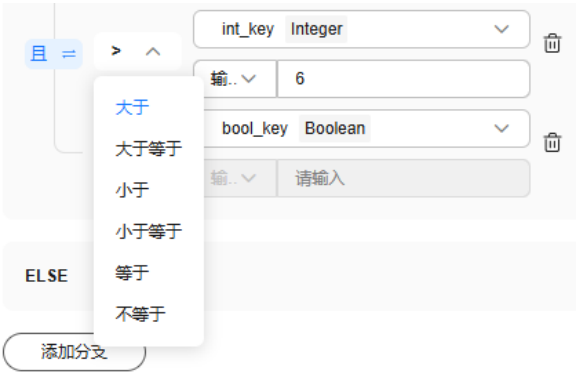
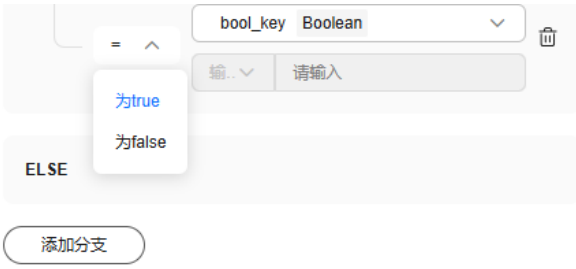
以常见的String、Integer和Boolean类型条件判断为例，在开始节点中定义三种类型的参数，模拟判断节点的输入参数，实现对于不同类型参数在不同条件下的逻辑判断。

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。

版权所有 © 华为云计算技术有限公司

表 4-32 判断节点配置示例

参数类型	参数名称	配置示例
String类型	IF	例如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	IF	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	IF	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，输入以下测试值并运行：

版权所有 © 华为云计算技术有限公司

表 4-33 试运行参数

参数	测试输入
string_key	abcd
int_key	7
bool_key	true

预期结果：上述三个条件均满足IF分支（abcd包含 "abc"、7 大于 6、bool_key为true），工作流应执行IF分支。节点右侧IF分支连线高亮，则配置正确。

常见问题

- **判断节点和意图识别节点有什么区别？如何选择？**
  这两个节点虽然都用于实现分支逻辑，但它们的判断依据完全不同：

  表 4-34 节点差异

维度	判断节点	意图识别节点
判断依据	结构化数据（数字、变量、字符串）	用户自然语言输入
判断方式	预设逻辑条件（>、==、包含等）	大模型语义理解
典型场景	订单金额是否 > 1000、状态是否为"成功"	用户意图识别：“我想退款”
特点	精确、可预期、无AI不确定性	灵活、可处理模糊表达

- **为什么我的ELSE IF分支永远不执行？**
  检查您的IF分支条件是否过于宽泛（例如使用了“不为空”作为第一个条件）。一旦满足IF条件，就不会继续检查ELSE IF，请将最具体的条件放在最上方。
  请将最具体、最严格的条件放在最上方，将兜底条件（如“不为空”）放在靠后的分支。
- **引用的前置节点变量为空（null）时，判断条件会如何执行**
  - "为空" 条件：判断结果为true
  - "不为空" 条件：判断结果为false
  - 其他比较运算符（如大于、包含、长度大于）：判断结果均为false，不会报错，但会继续向下检查下一个分支

  建议在工作流中为可能为空的变量添加“为空”的兜底分支，避免所有条件均不满足时意外进入ELSE分支。

4.10.2 代码

代码节点是工作流中用于执行自定义代码逻辑的节点。当工作流中需要进行**数据处理、格式转换、数学计算、逻辑判断**等操作，而这些操作用其他节点难以实现或用大

版权所有 © 华为云计算技术有限公司

模型处理不够准确时，可以通过代码节点编写代码来完成。代码执行的结果是确定的，相同的输入必然得到相同的输出。对于计算、格式转换等任务，代码比大模型更准确、更快、且不消耗Token。

代码节点在平台托管的沙箱环境中执行，代码运行时受到超时时间限制（默认900秒，最小0.1秒），且无法访问本地文件系统。支持的库范围见下方**Python 3.9代码要求**。

数据流转过程：上游节点输出 → 输入参数映射 → args字典传入main函数 → 代码处理 → return字典 → 输出参数映射 → 下游节点获取。

代码节点的典型使用场景如下：

表 4-35 代码节点使用场景

场景	示例
数据格式转换	将上游节点返回的JSON数据提取、重组为下游节点需要的格式。
数学计算	价格计算、折扣计算、数据汇总统计等。
文本处理	正则匹配、字符串拼接、内容截取、编码转换等。
逻辑判断	复杂的条件判断、数据校验、规则匹配等。
数据清洗	去重、过滤、排序、字段映射等。

不推荐使用代码节点的场景：

- 需要安装使用不支持的外部第三方库（详见**Python 3.9代码要求**）。
- 需要长时间运行（超过900秒最大超时限制）。
- 需要持久化存储状态或访问本地文件系统。

节点选型对比：

当工作流中存在多种节点可实现类似功能时，可参考以下对比进行选型：

表 4-36 节点选型对比

决策维度	代码节点	大模型节点	条件判断节点
执行确定性	确定执行，相同输入相同输出	可能产生不同输出	确定执行
适合任务类型	数据处理、数学计算、格式转换	文本生成、语义理解、内容创作	简单条件分支
Token消耗	不消耗Token	消耗Token	不消耗Token
灵活性	高（可编写任意代码逻辑）	中（依赖Prompt设计）	低（仅支持条件分支）

选型建议：

版权所有 © 华为云计算技术有限公司

- 如果您的场景是精确的数学计算或数据格式转换，推荐使用代码节点。
- 如果您的场景是需要语义理解或内容生成，推荐使用大模型节点。
- 如果您的场景是简单的条件判断（如A>B），推荐使用条件判断节点；如果条件逻辑复杂（如多条件组合、正则匹配），推荐使用代码节点。

配置代码节点

代码节点的配置参数说明如下：

表 4-37 代码节点配置参数

参数	说明
输入参数	输入参数用于声明代码中所使用的输入变量，每个输入参数需要和代码中定义的参数保持一致。 参数值可以直接填写一个固定的值，也可以选择引用工作流中上游节点的输出和全局配置的记忆变量。 描述：（可选）用于说明该输入参数的具体含义。
输出参数	输出参数用于定义代码执行完成后向下游传递哪些数据，输出参数的名称和类型必须与代码中返回的键值对严格对应，否则下游节点将无法获取到对应的值。 描述：（可选）用于说明该输出参数的具体含义。
代码	编写代码时不要更改第一行函数定义。 编写Python代码片段，仅支持Python系统库；代码配置示例请参考**Python 3.9代码要求**。 在代码编辑器的顶部支持选择大模型，智能优化代码。同时支持自动折行、复制和全屏显示、实时代码低错检测（如缺少冒号、多余逗号、缩进等）等功能。
异常处理	支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。 “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。 “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。 “异常处理方式”：配置异常处理方式。 • 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。 • 返回设定内容：节点发生异常后，工作流不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。 • 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点后新增节点，并为新增的异常分支配置相应的处理流程。

版权所有 © 华为云计算技术有限公司

Python 3.9 代码要求

支持Python 3.9版本及其标准库，并支持如下的非标准库，可直接在Python函数代码中声明使用。

表 4-38 Python 运行时集成的非标准库

模块	功能	版本号
dateutil	日期/时间处理	2.6.0
httplib2	httpclient	0.10.3
numpy	数学计算	1.13.1
redis	redis客户端	2.10.5
obsclient	OBS客户端	-
smnsdk	访问SMN服务	1.0.1

示例如下：

```
def main(args: dict) -> dict:
    """
    运行代码节点会调用此函数，请勿对下面的函数名做修改
    :param args: 输入固定为args字典类型，kv为输入参数键值对
    :return: 输出参数为字典类型，kv为输出参数键值对
    """
    ret = {
        "key0": args.get('input', 'default'),
        "key1": "hi"
    }
    return ret
```

- 平台通过固定的main函数触发代码执行，函数名不可修改。
- 输入参数args是一个字典（dict），包含了在代码节点界面配置的所有输入参数，以参数名为key、参数值为value的键值对形式传入。通过args.get('参数名', '默认值') 读取，建议始终设置默认值，避免参数缺失时报错。
- 返回值是一个字典（dict），直接return即可，无需做JSON转换。返回字典中的key必须与代码节点界面配置的“输出参数”名称完全一致（区分大小写），否则下游节点将无法获取到对应的值。

  在代码节点界面的“输出参数”区域配置的参数名，必须与代码return字典中的key对应。
- 编写约束：

  ```
  入口函数名必须是main，接收一个dict参数，返回一个dict
  直接返回字典即可，无需json.dumps()转换
  即使只有一个输出字段，也必须以字典形式返回
   ×  return "hello"
   √  return {"message": "hello"}
  ```
- **不支持的命令：**

  系统命令执行：system、popen、subprocess、startfile、execl、spawn。

  网络请求命令：requests、urllib、httpx、aiohttp、socket、socks。

  动态代码执行：eval、exec、compile、importlib、__import__、execfile。

  序列化/反序列化：pickle、marshal、yaml、shelve。

 版权所有 © 华为云计算技术有限公司

文件操作：open、remove、rename、shutil、listdir、tempfile、glob、unlink。
其他：torch、environ、functools、ast。

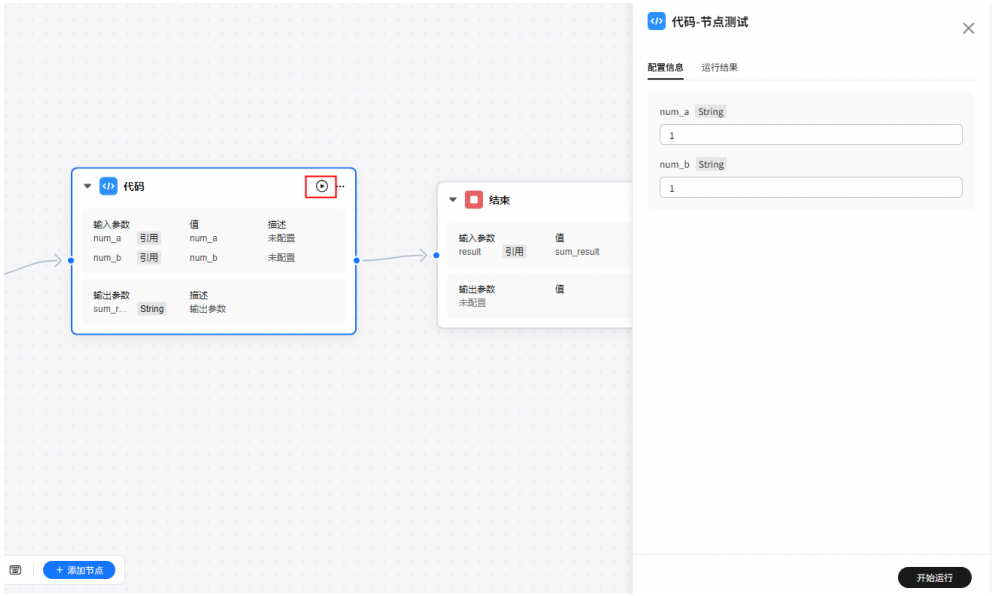
验证代码节点配置

代码节点配置完成后，建议通过以下步骤验证配置是否正确：

单击节点中的测试图标，在测试面板中为每个输入参数填写测试值，单击“开始运行”。等待输出结果，确认各输出参数的值和类型与配置一致。

验证通过的标准：输出参数的名称、类型和值均与预期一致，且无语法错误或运行时异常。

图 4-88 验证代码节点配置



相关操作

- 单击节点名称，可修改节点名称。
- 单击···图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

- **代码执行后报SyntaxError**

  Python代码的常见语法问题包括：缩进不统一（混用了空格和Tab）、括号或引号未成对闭合、中英文符号混用（如使用了中文冒号： 或中文括号（ ））。

- **import第三方库报ModuleNotFoundError**

  默认情况下仅支持内置标准库，不支持安装和使用第三方库。

版权所有 © 华为云计算技术有限公司

- **下游节点获取不到代码节点的输出值**

  最常见的原因是输出区域配置的参数名与代码return中的key不一致。两者必须完全匹配，包括大小写。例如代码返回了totalPrice，输出区域配置成totalprice就会取不到值。

- **下游节点获取到的值类型与预期不符**

  检查输出区域配置的参数类型是否与代码实际返回的数据类型一致。例如代码返回了数字100，但输出参数类型配置为String，下游节点拿到的就是字符串"100"而非数字。

4.10.3 循环

循环节点用于在工作流中重复执行一系列任务，直到满足特定条件为止。当你需要批量处理数据、逐一遍历列表元素或按固定次数重复执行某个操作时，可以使用循环节点来实现。

循环是一种常见的工作流控制机制，它让一组操作能够被重复执行，而不需要手动复制粘贴相同的节点。在实际业务场景中，循环节点可以帮助您：

- **逐条处理列表数据：**对一组数据中的每一条逐一执行相同的操作，例如依次对多篇文章进行摘要提取。
- **重复执行固定次数的任务：**例如让大模型连续生成多轮创意方案，每轮基于上一轮结果优化迭代。
- **满足条件时自动终止：**配合终止条件，在循环过程中自动判断是否需要提前退出，避免无效执行。

约束限制

在使用循环节点之前，请了解以下限制：

- 循环节点内不支持配置问答节点、提问器节点、输入节点或带有这些节点的子工作流。
- 循环体内的流程必须从“循环输入”开始，经过具体任务节点，最终连线到“循环输出”，否则循环将无法正常执行。
- 在使用“数组循环”模式时，传入的变量值必须为标准的数组（Array）格式。

循环体中的节点将在每轮循环中重复执行。如果循环体包含大模型节点等消耗资源或费用的节点，循环次数越多，资源消耗和费用越高。建议始终配置合理的终止条件，首次调试时将循环次数设为较小值（如1~3次），确认逻辑无误后再调大。

循环类型说明

循环节点的运行模式由循环类型决定，不同的循环类型适用于不同的场景。目前支持以下两种循环类型：

表 4-39 循环类型说明

循环类型	适用场景	循环次数	终止方式
使用数组循环	遍历已知列表中的每个元素	取决于数组长度	数组遍历完毕后自动结束，或满足终止条件时提前退出

版权所有 © 华为云计算技术有限公司

循环类型	适用场景	循环次数	终止方式
指定循环次数	需要重复执行固定次数的操作	手动指定次数	达到指定次数后自动结束，或满足终止条件时提前退出

循环类型选型建议：

- 如果你已有一组待处理的数据（如列表、数组），推荐使用“数组循环”，循环次数由数据量自动决定，无需手动计算。
- 如果你需要重复执行某个操作N次，且不依赖外部数据，推荐使用“指定循环次数”，直接设定次数，配置更简单。
- 如果你需要获取当前循环轮次信息，两种模式均可通过内置变量index获取。

使用数组循环

数组循环类似编程语言中的for语法结构，用于遍历一个已知的序列（数组），对序列中的每个元素执行完全相同的步骤。

- 运行机制：循环的次数取决于您传入的数组长度（例如传入包含5个元素的数组，就循环5次）。您需要将上游节点的数组输出引用给循环节点。
- 内置变量：在数组循环中，系统会自动提供两个仅限循环体内部使用的内置变量：
  - item（元素）：当前这一轮循环正在处理的具体数据内容。
  - index（索引）：当前循环的轮次标识（从0开始计算。0代表第1次循环，1代表第2次循环，以此类推）。
- 典型场景：长文档分段总结、批量翻译多国语言、根据一组核心词批量生成多篇营销文案。

指定循环次数

指定循环次数时，不需要依赖外部输入的数组，而是直接设定一个固定的循环次数。

典型场景：多轮次剧情推理、固定部署的数据清洗与重试。

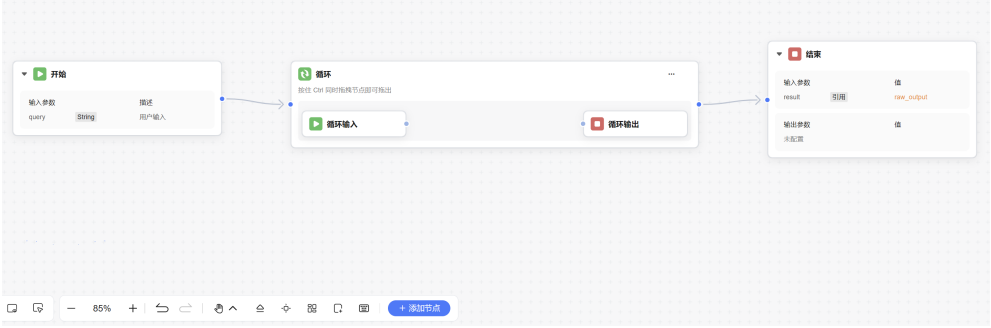
配置循环节点

配置循环节点的详细参数说明如下：

添加循环节点

在工作流编辑页面，单击下方“添加节点”，选择“循环”节点。循环节点中执行具体任务的节点，可以添加后，拖拽至循环节点内部。循环体内的流程必须从“循环输入”开始连线，经过具体任务节点，最终连线到“循环输出”。

版权所有 © 华为云计算技术有限公司



循环类型

- 如果选择“使用数组循环”，需配置arr_loop_var变量。arr_loop_var是数组循环模式下的固定入参，用于接收需要遍历的数组。循环节点会依次取出该数组中的每个元素，逐轮传入循环体执行处理。它支持引用上游节点的输出，且**值必须为数组格式**，数组的长度即为循环的执行次数。
- 如果选择“指定循环次数”，直接输入设定的次数，循环次数支持设置为1~1000次。

循环节点本身不消耗大模型Token，但若循环体内部部署了大模型节点，每运行一轮循环都会消耗对应大模型的Token。请合理配置循环次数和终止条件以控制运行成本。关于大模型的费用，请参考**模型计费说明**。

中间变量

中间变量是循环节点中用于在每一轮循环之间**传递和更新数据**的参数。中间变量可以输入一个固定值，或者引用前置节点的输出变量和全局配置中的记忆变量。

- 中间变量可以将不需要遍历的背景信息（如产品介绍、全局要求、用户设定的固定条件）一次性传入循环体，让每一轮循环都能直接读取和使用，而无需在节点内反复手写。例如，上游节点通过插件查询到了“今天的天气是晴天”，您可以将天气结果通过中间变量传给循环体，让大模型在后续的每一轮循环创作中都知道今天是晴天。
- 除了作为常量使用外，在循环体中通过结合“变量赋值”节点，将本轮处理后的最新结果覆盖更新到该中间变量中。下一轮循环启动时，节点会自动读取更新后的值，从而实现中间变量的动态更新。

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。

终止条件

通常情况下，循环节点会一直执行，直到“数组遍历完”或“达到指定的循环次数”。但在实际业务中，往往需要根据**运行过程中的动态结果**来随时停止循环（例如：AI已经给出了满意的答案、或者调用的插件连续报错）。

终止条件并非实时监控，而是在每次循环体执行完成后进行判定。终止条件通常需要与中间变量搭配使用，这是最常见的使用模式。原因在于：终止条件需要一个"判断依据"，而这个依据往往是循环过程中不断更新的状态值（即中间变量）。

终止条件由一个或多个条件表达式组成，每个条件表达式的结构为：[判断参数] [比较条件] [比较参数]。

版权所有 © 华为云计算技术有限公司

- 判断参数：选择来自循环体内前序节点的输出参数或中间变量。这是终止条件的“监控对象”。
- 比较条件：系统会根据你选择的判断参数类型，自动提供对应的判断规则。包括但不限于：
  - 文本类：等于/不等于、包含/不包含、为空/不为空等。
  - 数值/数组类：长度大于等于 (>=) / 长度小于等于 (<=) 等。
- 比较参数：用于对比的目标值，支持两种设置方式：引用（选择前置节点输出变量或全局配置中的记忆变量）和输入（直接填写具体值）。

配置示例

本示例创建一个产品文案写作助手，使用数组循环方式，通过循环体生成["小红书", "微信朋友圈", "微博"]风格的宣传文案。

步骤1 创建一个任务型工作流，并添加循环节点、大模型节点，并将大模型节点放置在循环体内部，连接各节点。

图 4-89 创建任务型工作流

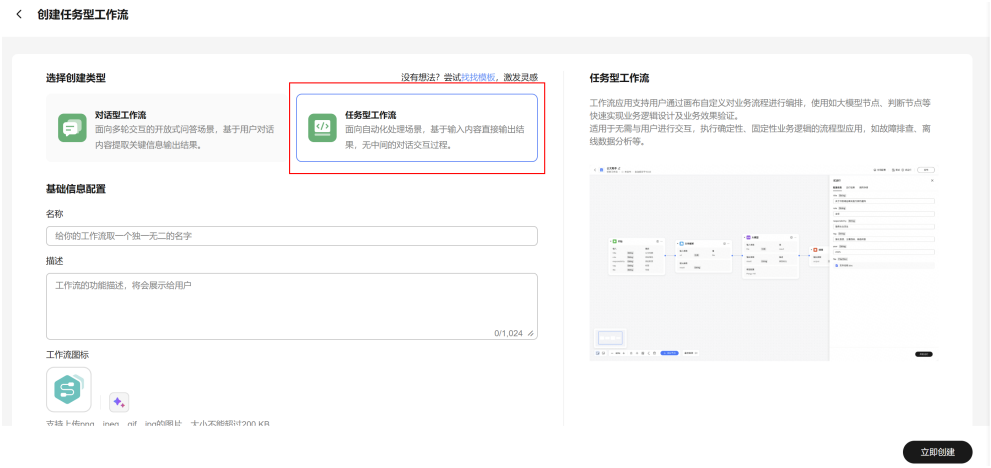


图 4-90 编排循环节点



步骤2 配置开始节点，开始节点共计两个参数，将新增的item参数设置为Array<String>格式，并单击 ••• 输入默认值["小红书", "微信朋友圈", "微博"]。

版权所有 © 华为云计算技术有限公司

图 4-91 配置开始节点



步骤3 配置循环节点。单击循环节点，参考下图完成配置。

循环类型：使用数组循环。

循环数组：引用开始节点的item参数。

输出参数：设置为output，引用大模型的输出结果参数raw_output。

图 4-92 配置循环节点



步骤4 配置循环体中的大模型节点，单击大模型节点，参考下图完成配置。

模型配置：选择任一可用模型。

输入参数：新增query参数，引用开始节点的query参数；新增item参数，引用循环节点的item参数。

输出参数：使用默认。

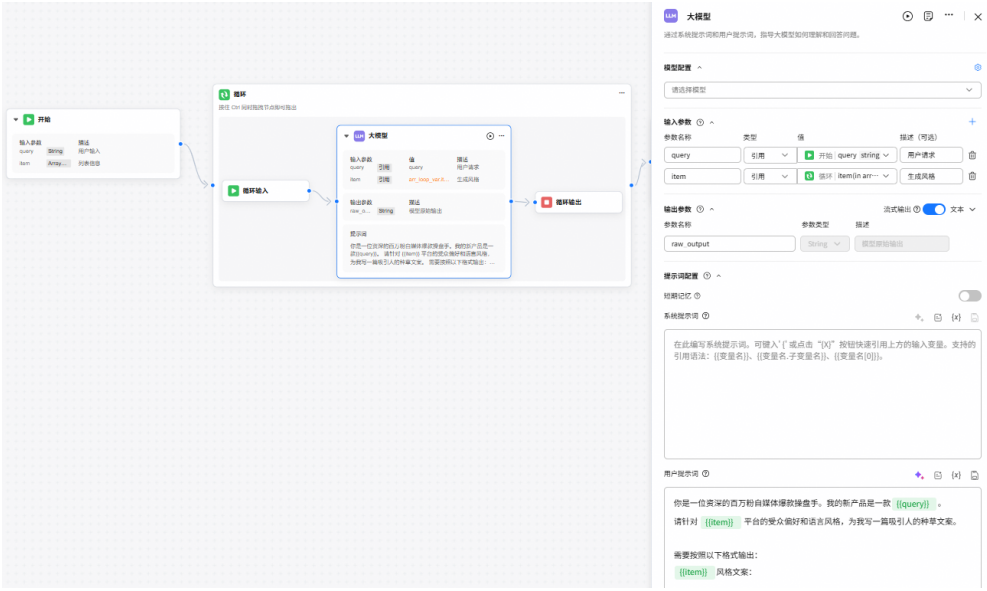
用户提示词：

```
你是一位资深的百万粉自媒体爆款操盘手。我的新产品是一款{{query}}。
请针对 {{item}} 平台的受众偏好和语言风格，为我写一篇吸引人的种草文案。

需要按照以下格式输出：
{{item}}风格文案：
```


版权所有 © 华为云计算技术有限公司

图 4-93 配置大模型节点



步骤5 配置结束节点，result参数引用循环节点的输出参数output。

图 4-94 配置结束节点



步骤6 单击“试运行”，参考下图输入测试问题后，运行工作流。

版权所有 © 华为云计算技术有限公司

图 4-95 试运行工作流



步骤7 检查输出结果是否符合预期。

可查看工作流已经按照要求输出对应的文案。单击“调试”可查看工作流的调用链信息，由于本案例中["小红书", "微信朋友圈", "微博"]共有3组数据，因此可以看到大模型通过循环的方式执行了3次。

图 4-96 检查输出结果

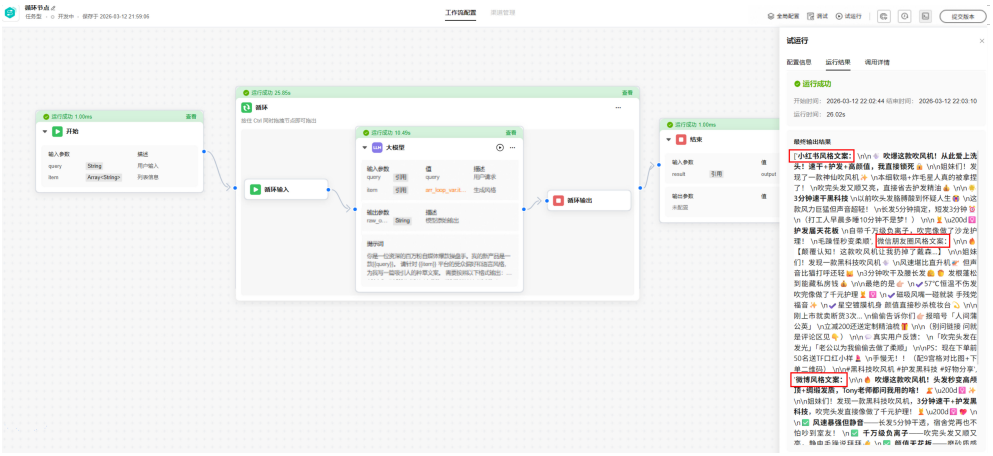


图 4-97 查看调用链



----结束

常见问题

- **不设置终止条件会怎样**

  不设置终止条件时，循环将完整执行。数组循环会遍历完所有元素，指定循环次数会跑满设定轮次。如果您的场景需要每条数据都处理，可以不设置终止条件。

版权所有 © 华为云计算技术有限公司

- **循环次数设置为1是否有意义**

  有。设置为1次本质上就是执行一次循环体中的逻辑。在某些场景下，你可能希望通过循环节点的结构来组织一段独立的子流程，或者先将循环次数设为1进行单轮调试，确认逻辑正确后再调大轮数。

- 为什么循环体不支持Agent节点或问答节点

  循环节点的设计初衷是处理“批量任务”。Agent和问答节点属于强交互式任务，包含复杂的上下文交互和人工干预逻辑，目前在循环结构中尚未开放，以避免死循环或响应超时。

4.10.4 意图识别

意图识别节点用于识别用户自然语言表达的意图或目的，适用于需要对用户问题进行分类，或者综合类功能需要不同分支处理的场景。

通过对用户输入进行推理分析，匹配预定义的意图类别，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，当意图数量小于或等于20时，选择意图识别节点。
- 高级模式：适用于对大量可归类意图进行分类的场景，当意图数量大于20时，选择**高级意图识别**。

意图识别节点和判断节点的区别：

表 4-40 意图识别节点和判断节点的区别：

对比维度	意图识别节点	判断节点
输入类型	自然语言文本	确定性条件表达式
分类方式	大模型语义理解	条件逻辑判断（if/else）
适用场景	用户问题分类、意图路由	参数值判断、业务逻辑分支
结果确定性	存在概率性，可调生成随机性进行控制	完全确定性
示例	输入“我想订机票”→预订机票意图	age >= 18→成人流程

说明

当分类依据是自然语言语义时选用意图识别节点；当分类依据是明确的参数值条件时选用判断节点。两者可配合使用。

计费说明

意图识别节点每次执行都会调用所选大模型，产生相应的Token消耗费用。

- 使用平台预置模型：
  - 试运行阶段：系统优先使用平台赠送的2,000,000Tokens，用完即止。

版权所有 © 华为云计算技术有限公司

  - 赠送额度用尽或工作流发布后： 需开通模型并设置鉴权，开通操作请参见**接入预置的付费模型**，计费规则请参见**MaaS模型即服务价格**
- 使用自定义接入的第三方模型：计费规则请参考该第三方模型的官方说明。

配置意图识别节点

配置意图识别节点的详细参数说明如下：

模型配置

选择要使用的模型。意图识别属于逻辑推理任务，建议选择推理能力强的模型，而非创意生成型模型。

您还可以单击图标，调整模型配置。模型配置的详细说明可参考**选择并配置模型**。

注意

意图分类要求输出具有高度确定性，避免发散性结果，建议将“生成随机性”配置项设置为0。如需处理语义边界模糊的场景，可适当调至0.1~0.2。

输入参数

输入参数固定为1个，参数名称为input，且不可修改，类型是字符串，表示待进行意图识别的用户文本。

设置参数值时，支持“引用”和“输入”两种方式。

- 引用：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置）以及系统参数。

  说明

  - 最常见的配置：引用“开始”节点的输入参数“query”，即直接对用户输入进行意图分类。
  - 进阶用法：引用“大模型”节点的输出参数“raw_output”，先用大模型节点对用户输入进行预处理或标准化，再进行意图识别。

  系统参数列举如下：
  - “conversation_history”：存储当前对话的历史消息记录。
  - “current_time”：获取当前的系统时间戳或时间字符串。
  - “user_id”：识别当前用户的唯一标识符。
  - “conversation_id”：当前工作流的唯一对话标识。
  - “dialogue_count”：当前对话的轮次数计数。
- 输入：支持输入指定内容，适用于调试测试场景。

意图配置

用户意图的分类选项，支持设置多个分类。默认提供2个意图分支，用户可以单击“添加意图”添加多个意图分支，意图数量最多为20个。

版权所有 © 华为云计算技术有限公司

⚠ 注意

意图数量增多可能降低分类准确率，建议单个意图识别节点不超过15个意图分类。超过20个上限时需考虑拆分为多级意图识别工作流。

当所有预设意图均不匹配时会命中“其他意图”分支，这是兜底分支，通常连接到通用大模型节点，处理闲聊或未定义问题。

意图描述：

意图输入框中的描述内容，是意图识别节点**最关键**的配置，决定了模型能否准确理解用户意图。配置质量越高，识别准确率越高。

好的意图描述应清晰定义边界并包含关键要素，意图描述示例如表4-41所示。

表 4-41 意图描述示例

分类	描述内容	理由
好的意图描述	用户询问某个城市或地区的天气情况，包括当前天气、未来天气预报、气温、下雨等。	列出具体场景和常见关键词。
一般的意图描述	用户询问天气相关的问题。	有基本方向但缺少关键词和场景细化。
差的意图描述	查天气。	过于简短，大模型无法理解该意图的完整范围。

高级配置

高级配置可使用默认的提示词。当意图识别效果不佳时，可自定义修改提示词优化效果。具体说明如下：

- **提示词：**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。提示词的撰写影响到意图识别节点的准确性。当意图识别效果没有达到预期时，可以调整提示词优化效果。例如可以在提示词中补充“用户提问飞机时，识别为预订机票功能”，提升“预订机票”意图识别的成功率。同时支持单击 {x} 快速插入输入参数，或在输入框中键入' {' 快速引用参数。
- **历史对话轮次：**选择是否打开历史对话引用功能，取值范围为0~30。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检索匹配进行意图识别，提升意图识别节点的分类能力。
  - 意图样例知识库：开启辅助识别后，用户需要先创建分类样例知识库，向知识库上传意图FAQ，并选择配置该知识库。
  - 过滤标签：可填写意图样例知识库上传FAQ时标注的标签值，表示在该标签范围内进行FAQ检索匹配。如果不填写，则默认在整个知识库范围下做FAQ检索匹配。
  - 匹配阈值：当分类样例的匹配度低于设置阈值时，会采用所配置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

输出参数

节点的输出参数，可作为变量被后续节点引用。

表 4-42 输出参数说明

参数名	类型	说明	示例值
classification_id	Integer	意图ID。根据意图配置中的顺序，从上到下依次排序，第一个意图ID为1；未命中任何已配置意图时ID为0，执行**其他意图**分支。	1
name	String	意图描述名称。	文本翻译

相关操作

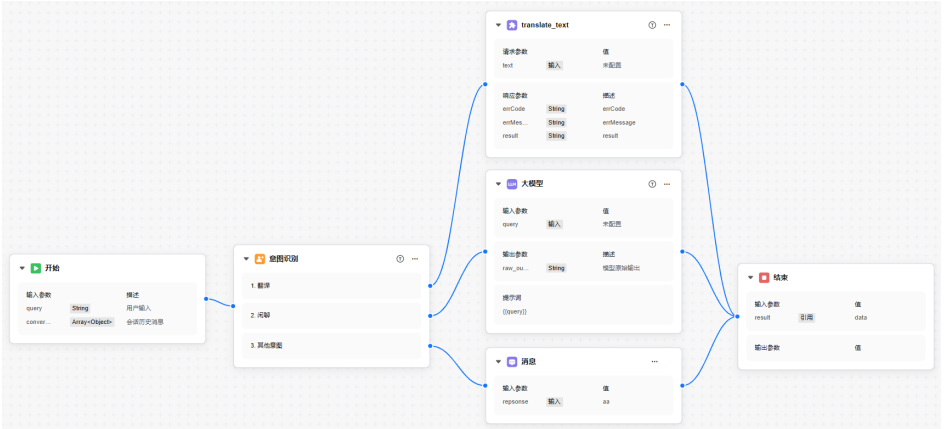
- 单击节点名称，可修改节点名称。
- 单击⋯图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击▷图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

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。

例如，提供翻译功能的工作流，节点配置如下：

图 4-98 配置示例



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”，该意图分支连接翻译插件节点，实现翻译功能。

版权所有 © 华为云计算技术有限公司

意图2的类别描述为“用户闲聊”，该意图分支连接大模型节点，实现闲聊功能。

默认的其他意图分支连接消息节点，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。

常见问题

表 4-43 常见问题

问题现象	可能原因	解决方案
用户表达明确，但分到了错误的意图	1. 目标意图的描述不够具体。 2. 另一个意图的描述覆盖面过宽。	1. 在正确意图的描述中添加该表达的关键词。 2. 在错误意图的描述中明确排除（“不包含XX场景”）。
同一输入多次运行结果不同	生成随机性设置过高。	将生成随机性降至0（推荐）或0.1-0.2。
始终走“其他意图”	1. 意图描述与用户表达差异过大。 2. 输入参数未正确引用。	1. 优化意图描述，增加同义表达。 2. 检查input是否正确引用了开始节点的输入参数{{query}}。
辅助识别未生效	1. 知识库未绑定。 2. FAQ的“答案”与意图名称不一致。 3. 匹配阈值设置过高。	1. 确认已选择知识库。 2. 确保FAQ中的答案字段与对应的意图名称完全一致（含大小写）。 3. 适当降低阈值。
输入参数下拉列表为空	节点未连接前置节点。	确保意图识别节点已连线到开始节点或其他上游节点。
意图数量超过20个无法添加	系统上限为20个。	考虑**高级意图识别节点**，高级意图识别支持20个以上意图。

4.10.5 高级意图识别

高级意图识别节点是普通**意图识别节点**的**增强版本**，专为需要管理**20个以上意图**的复杂场景设计。它通过**意图包**集中管理大量意图分类，并使用**子工作流**为每个意图分支提供独立的处理逻辑，实现主工作流的模块化编排。

表 4-44 意图识别和高级意图识别节点差异

特性	意图识别节点	高级意图识别节点
适用规模	<=20个轻量级意图	>20个企业级海量意图
维护方式	节点内直接编辑	意图包统一管理
执行逻辑	节点后直接连线分支	映射子工作流

版权所有 © 华为云计算技术有限公司

特性	意图识别节点	高级意图识别节点
典型场景	简单问答、指令跳转（如3~5个意图的客服引导）	海量意图分类（如银行客服50+意图、综合政务100+意图、全功能电商助手80+意图）

说明

当意图数量在20个左右时，如未来有扩展需求，建议直接选用高级意图识别节点，避免后续迁移成本。

运作机制

高级意图识别节点的执行流程如下：

1. **接收输入**：接收用户输入文本，通过input参数传入。
2. **模型推理**：调用所配置的大模型，基于绑定的意图包中所有意图分类进行推理。
3. **输出分类**：输出分类结果（classification_id和name）。
4. **调度子工作流**：意图动作配置节点根据分类结果，调度对应的子工作流执行。
5. **结果回传**：子工作流执行完成后，结果回传至主工作流继续执行。

说明

意图动作配置是高级意图识别节点的配套节点，在高级意图识别节点保存后自动生成。它负责根据分类结果将请求路由到对应的子工作流，是主工作流与子工作流之间的桥梁。

计费说明

- 高级意图识别节点每次执行都会调用大模型进行推理，产生Token消耗费用。意图数量越多，单次推理消耗的Token越高。
- 意图动作配置节点调度子工作流时，子工作流中的节点也会产生相应的执行费用。

创建意图包

在配置高级意图识别节点前，必须先创建意图包。意图包是意图分类的“集中管理容器”，独立于工作流存在，可被多个工作流复用。创建意图包的操作请参见**创建意图包**。

配置高级意图识别节点

配置高级意图识别节点的详细参数说明如下：

模型配置

选择要使用的模型。意图识别属于逻辑推理任务，建议选择推理能力强的模型，而非创意生成型模型。

您还可以单击图标，调整模型配置。模型配置的详细说明可参考**选择并配置模型**。

版权所有 © 华为云计算技术有限公司

⚠ 注意

意图分类要求输出具有高度确定性，避免发散性结果，建议将“生成随机性”配置项设置为0。如需处理语义边界模糊的场景，可适当调至0.1~0.2。

输入参数

输入参数固定为1个，参数名称为input，且不可修改，类型是字符串，表示待进行意图识别的用户文本。

设置参数值时，支持“引用”和“输入”两种方式。

- 引用：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置），以及系统参数。

  说明

  - 最常见的配置：引用“开始”节点的输入参数“query”，即直接对用户输入进行意图分类。
  - 进阶用法：引用“大模型”节点的输出参数“raw_output”，先用大模型节点对用户输入进行预处理/标准化，再进行意图识别。

  系统参数列举如下：
  - “conversation_history”：存储当前对话的历史消息记录。
  - “current_time”：获取当前的系统时间戳或时间字符串。
  - “user_id”：识别当前用户的唯一标识符。
  - “conversation_id”：当前工作流的唯一对话标识。
  - “dialogue_count”：当前对话的轮次数计数。
- 输入：支持输入指定内容，适用于调试测试场景。

意图配置

配置意图包后，节点将使用该意图包中的所有意图分类作为大模型推理的候选项。

⚠ 注意

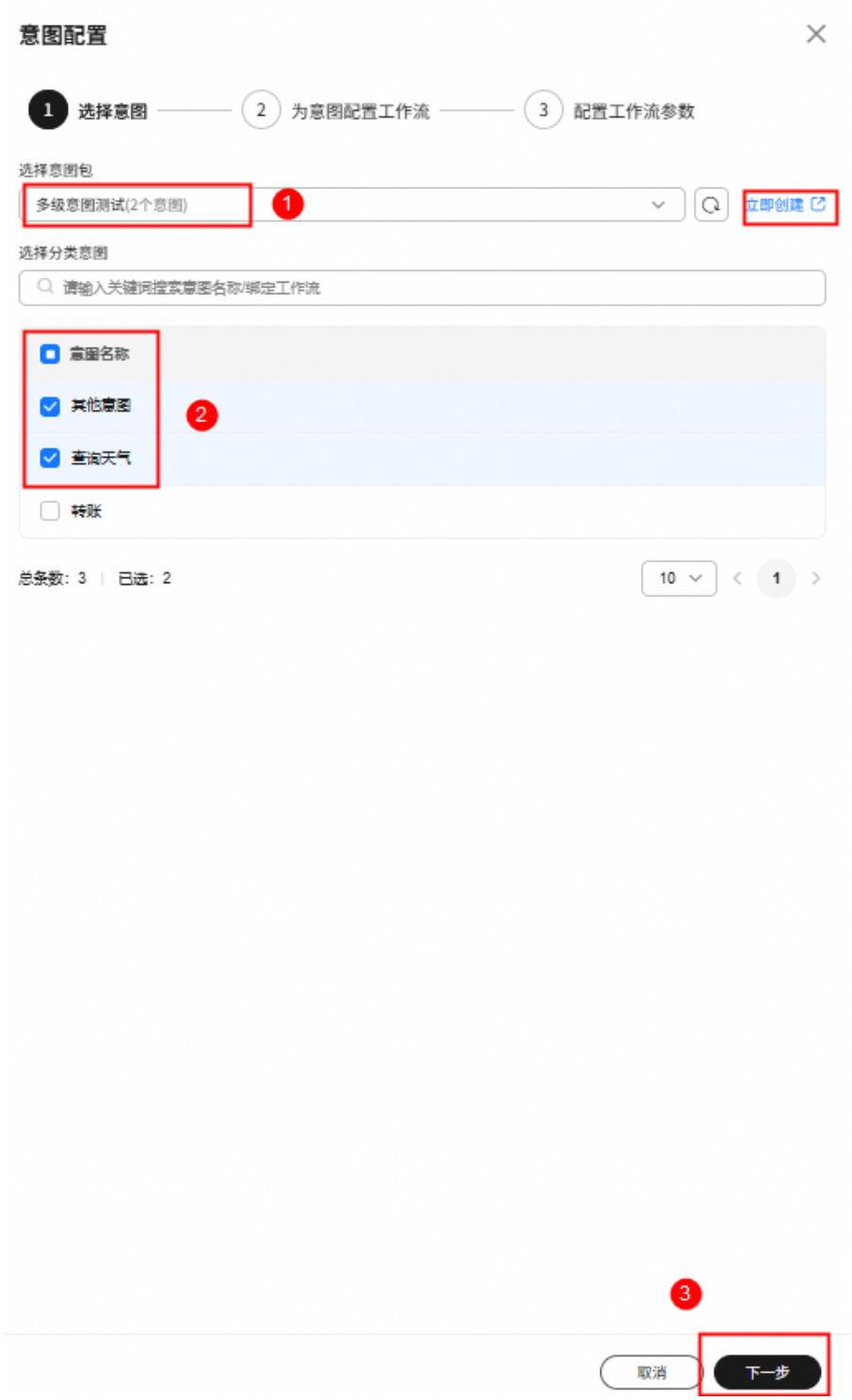
一个高级意图识别节点只能绑定1个意图包。若需管理更多意图，请在意图包中增加分类，或拆分为多级意图识别架构。

意图配置步骤如下：

步骤1 单击“请选择意图包进行配置”。您可以选择已配置的意图包，或单击“立即创建”创建一个新的意图包，可参见**意图管理**。

版权所有 © 华为云计算技术有限公司

图 4-99 选择意图



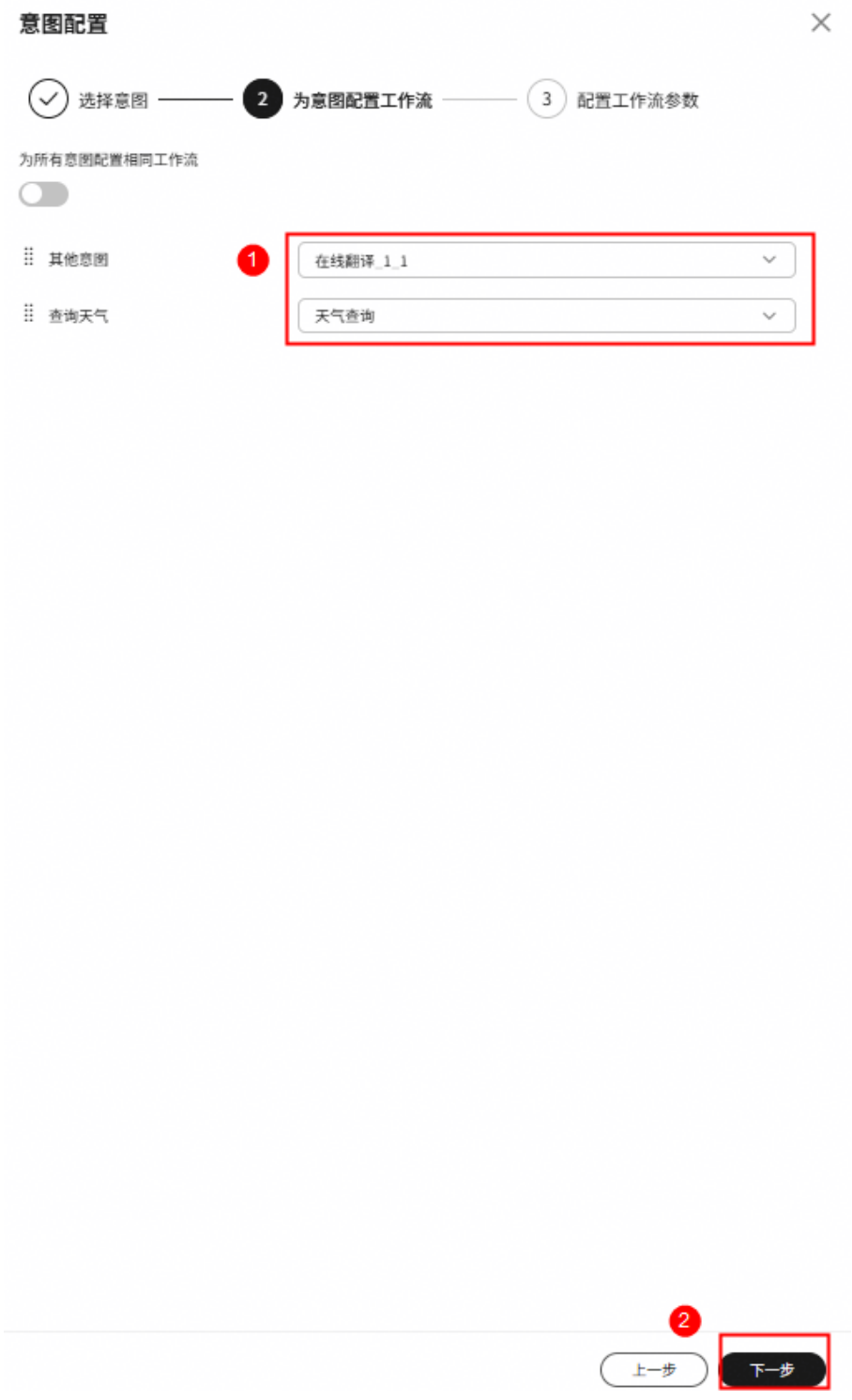
步骤2 **为意图配置工作流**。为每个意图分支配置对应的处理逻辑。配置完成后单击“下一步”。

说明

如需为所有意图配置相同工作流，可打开“为所有意图配置相同工作流”开关。

版权所有 © 华为云计算技术有限公司

图 4-100 配置处理逻辑



步骤3　配置工作流的输入参数。配置完成后，单击“确定”。

版权所有 © 华为云计算技术有限公司

图 4-101 配置输入参数



----结束

高级配置

高级配置默认使用系统提示词，也可自定义修改以优化意图识别效果。

- **提示词：**如果不配置将会使用系统默认值。提示词的撰写直接影响意图识别节点的准确性。当意图识别效果没有达到预期时，可以调整提示词优化效果。例如可

版权所有 © 华为云计算技术有限公司

以在提示词中补充“用户提问飞机时，识别为预订机票功能。”，提升“预订机票”意图识别成功率。同时支持单击 {x} 快速插入输入参数，或在输入框中键入“{{}}” 快速引用参数。

- **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。

输出参数

节点的输出参数，可作为变量被后续节点引用。

表 4-45 输出参数说明

参数名	类型	说明	示例值
classification_id	Integer	意图ID。根据意图配置中的顺序，从上到下依次排序，第一个意图ID为1；未命中任何已配置意图时ID为0，执行**其他意图**分支。	1
name	String	意图描述名称。	文本翻译

配置意图动作配置节点

当高级意图识别节点绑定意图包并保存后，系统会自动生成一个意图动作配置节点。该节点为意图包中的每个意图分类创建一个出口分支，您需要为每个分支配置对应的子工作流。

步骤1 在画布中找到自动生成的“意图动作配置”节点。

步骤2 **配置子工作流**。单击意图动作配置节点，为每个意图配置分支对应的处理逻辑。配置完成后单击“下一步”。

图 4-102 配置处理逻辑



步骤3 **配置子工作流的输入参数**。配置完成后，单击“确定”。

版权所有 © 华为云计算技术有限公司

图 4-103 配置输入参数



步骤4 连接意图动作配置节点和其他节点。

----结束

相关操作

- 单击节点名称，可修改节点名称。
- 单击⋯图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

表 4-46 常见问题

问题现象	可能原因	解决方案
意图包下拉列表为空	未在“配置管理>意图管理”中创建意图包。	先创建意图包并添加意图分类，然后回到节点配置中刷新选择。
意图动作配置节点未自动生成	高级意图识别节点未绑定意图包或未保存。	确认已选择意图包，并单击“确定”保存节点配置。
用户输入被分到错误意图	1. 目标意图的示例不够丰富。 2. 另一个意图的示例覆盖面过宽。 3. 生成随机性过高。	1. 为正确意图添加更多样化的示例。 2. 缩小错误意图的示例范围。 3. 将生成随机性降至0.1或更低。
始终走“其他意图”兜底	1. 意图示例与用户实际表达差异过大。 2. 输入参数未正确引用。	1. 用真实用户表达更新意图示例。 2. 确认input已引用开始节点的输入参数{{query}}。

版权所有 © 华为云计算技术有限公司

问题现象	可能原因	解决方案
子工作流收不到主工作流的变量	意图动作配置节点未配置输入参数映射	在意图动作配置节点中为每个分支显式配置参数映射。
子工作流执行完但主工作流无输出	子工作流的结束节点未定义输出参数。	在子工作流的结束节点中配置输出参数。
同一输入多次测试结果不同	生成随机性设置过高（大量意图场景更敏感）。	将生成随机性降至0.05-0.1。
意图数量很多时响应变慢	大模型需要处理大量候选意图的描述。	1. 考虑分层分类减少单次推理的候选项。 2. 选择推理速度更快的模型。

4.11 工具节点

4.11.1 插件

插件节点是工作流中实现第三方能力调用的核心组件之一，允许工作流通过调用插件来执行特定功能任务。

尽管大语言模型（LLM）具备强大的自然语言处理与推理能力，但存在两大局限：信息滞后（无法获取训练截止日期之后的实时信息）和行动力缺失（无法直接与外部系统交互）。插件节点正是突破这些局限的关键桥梁，每个插件本质上是经过标准化封装的API或者函数功能的工具集合，提供即插即用的模块化服务，让工作流具备连接外部世界的能力。

插件与工具的关系：插件是集成了一个或多个工具的功能集合，定义了这些工具共用的基础属性（如服务域名、鉴权方式）；工具是插件内具体的执行单元，每个工具对应一个独立的功能。工作流中添加插件节点时，实际调用的是插件中某一个特定工具。

插件与工具层级关系：插件 → [工具1, 工具2, ...]（每个工具对应一个独立功能）

插件节点的执行原理：插件节点执行时，平台服务器根据插件定义中的API地址和请求方式，将输入参数组装为HTTP请求发送给目标服务，等待响应后将返回数据映射为输出参数传递给下游节点。

数据流转过程：上游节点输出 → 输入参数映射 → 平台发起API请求 → 接收响应 → 输出参数映射 → 下游节点获取

相关文档：

示例：创建联网搜索插件

基于API创建插件

添加节点

工作流中支持添加两种类型的插件。

版权所有 © 华为云计算技术有限公司

表 4-47 插件类型

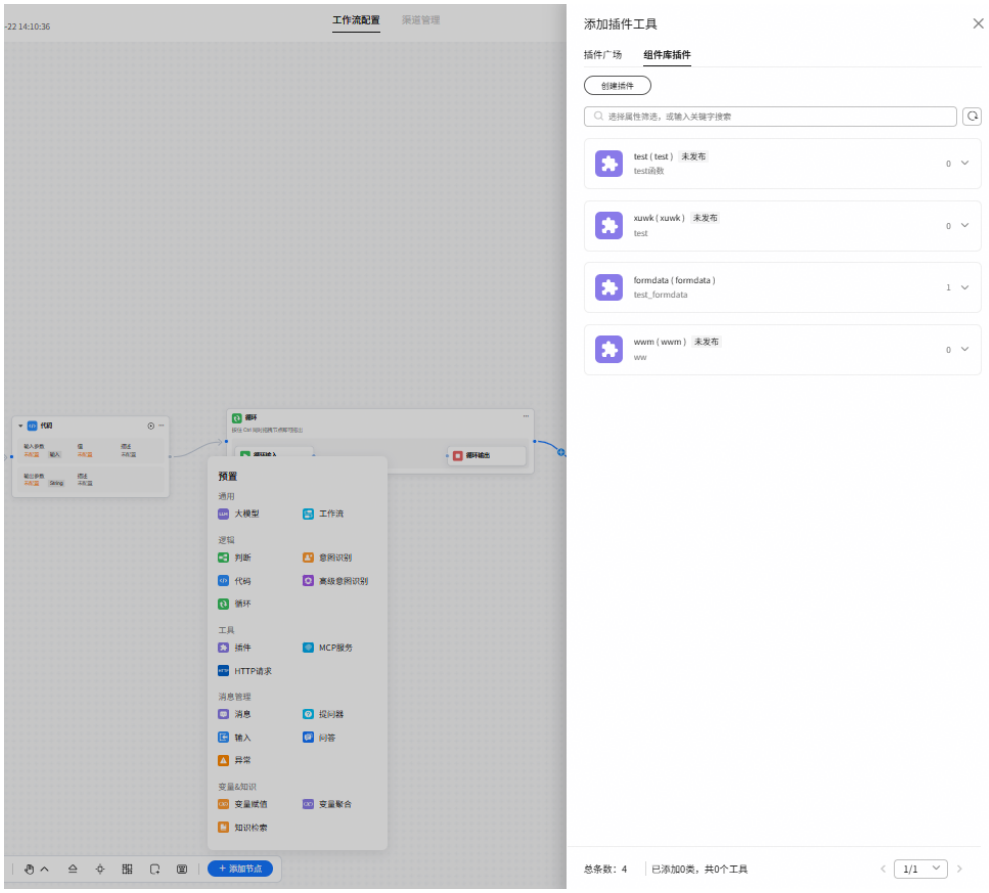
插件分类	说明	获取方式
插件广场	由AgentArts平台预置至“资产广场”的官方插件，由平台提供技术支持和维护。	来自“资产广场 > 插件广场”的插件，注意部分插件需要配置鉴权或收费。
组件库插件	开发者自行创建的插件，支持API类型（封装外部RESTful API）和函数类型（在平台沙箱中运行代码脚本）。	通过“组件库 > 插件”入口自行配置创建。

如果是需要调用通用的第三方成熟公共服务（如地图查询、天气预报、通用翻译等），推荐优先选择“插件广场”中的官方预置插件，即插即用。

如果是需要对接企业内部专有的业务系统、未公开的微服务或执行自研的代码脚本，推荐前往“组件库 > 插件”页面自行配置创建“组件库插件”。

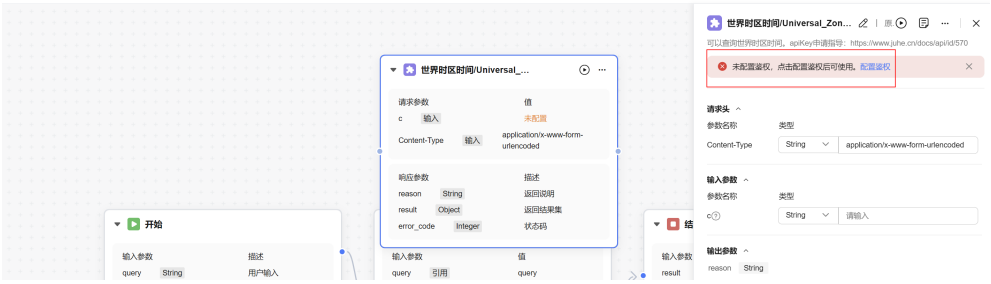
在工作流编辑页面，单击下方“添加节点”，选择“插件”节点，可按类型筛选，添加所需的插件。

图 4-104 添加插件节点



版权所有 © 华为云计算技术有限公司

图 4-105 插件鉴权



说明

- 使用官方插件时，系统会提示需要鉴权的插件，请根据页面指引完成相应配置鉴权操作后方可使用。
- 使用插件调用外部服务时，可能产生第三方服务的调用费用，该费用由第三方服务方收取。

配置插件节点

参数只读限制：插件节点的输入、输出参数结构均由插件创建时的定义（元信息）决定，在工作流节点内为只读状态，无法直接修改、添加或删除。若需要调整参数结构，必须回到“组件库 > 插件”页面修改插件定义。

也就是说，您在创建插件时（或插件开发者在定义API接口时）配置的请求参数会自动映射为插件节点的输入参数，API返回的响应字段会自动映射为插件节点的输出参数。

如果需要调整插件节点的参数结构（如新增、删除或修改某个字段），需要回到“组件库 > 插件”页面中修改插件本身的定义，而非在工作流节点中直接修改。配置插件节点的详细参数说明如下：

请求头

平台向插件发起请求时会自动附加自定义的HTTP头部信息，配置的请求参数会自动映射为插件节点的请求头。详细配置请参见**基于API创建插件**。

请求头的值支持手动输入固定值，或引用上游节点输出及全局配置的记忆变量。

输入参数

需要为插件节点的每个输入参数指定数据来源，支持以下两种取值方式：

- 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。前置节点为开始节点时，可以引用其输入参数。
- 输入：支持用户输入固定值。

无论是引用还是输入，均需要保证参数类型正确。

同时支持用户自定义输入参数的具体含义。

说明

注意：手动输入Array（数组）或Object（对象）类型时，必须使用标准的JSON格式，否则节点运行时会报格式错误。详细格式要求请参考**常见问题**。

版权所有 © 华为云计算技术有限公司

输出参数

输出参数的所有信息从插件元信息中自动导入，无需手动修改。插件节点运行时会调用工具处理输入参数，并返回处理后的数据供下游节点使用。

异常处理

支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。

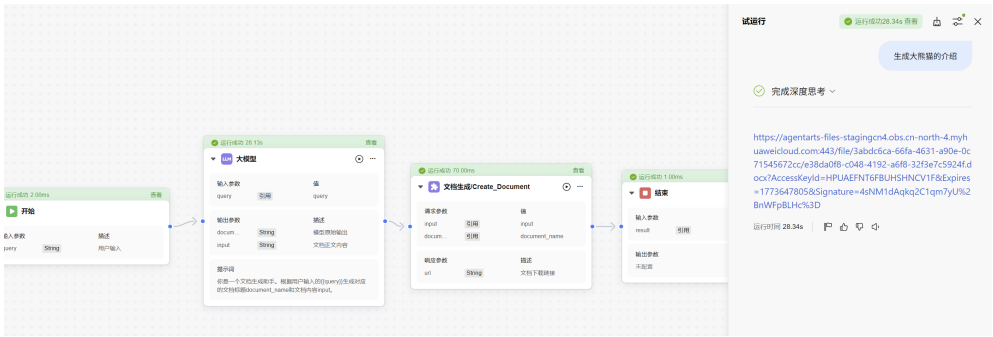
- “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。
- “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。
- “异常处理方式”：配置异常处理方式。
  - 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。
  - 返回设定内容：节点发生异常后，工作流不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。
  - 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点后新增节点，并为新增的异常分支配置相应的处理流程。

配置示例

本示例使用预置的“文档生成”插件并结合大模型生成文档内容。用户输入待生成的内容后，大模型会自动给出文档标题、文档内容，并由后续的插件节点将内容合并，生成一个可供下载的文档链接。

运行成功后，结束节点将返回包含文档下载链接的输出，用户可通过该链接下载生成的文档。

图 4-106 示例效果



- **添加插件**

  新建一个工作流，添加插件节点，选择来自插件广场的“文档生成”插件，该插件会将用户输入的标题、内容合并生成一份文档。

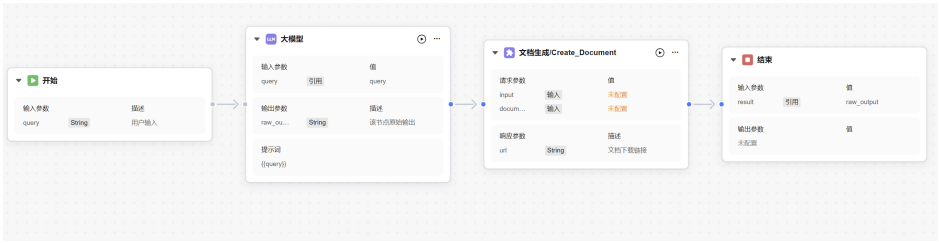
版权所有 © 华为云计算技术有限公司

图 4-107 添加插件



依次连接开始节点、大模型节点、插件节点、结束节点。

图 4-108 连接各节点



- **变量引用**
  - 所有节点的输入参数均使用“引用”关系，引用前置节点的内容作为本节点的输入。
  - 由于示例使用的“文档生成”插件需要输入文档标题、文档内容，该内容可以由前面的大模型节点生成。配置大模型节点“用户提示词”如下。配置提示词时需要考虑后续插件节点的参数输入形式，生成与插件节点参数相同的内容。

    ```
    你是一个文档生成助手。根据用户输入的{{query}}生成对应的文档标题document_name和文档内容input。
    ```

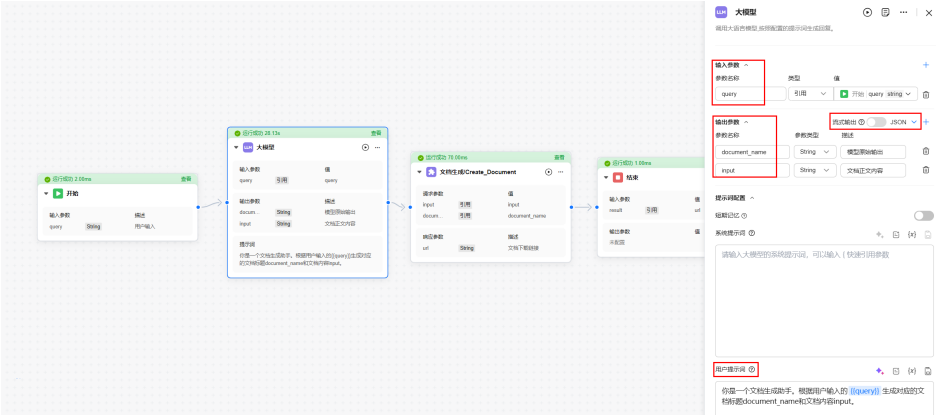
  - 大模型节点除了通过提示词的方式将输出格式与插件节点对齐外，还需要将输出参数改为一致。关闭输出参数的流式输出模式，改为JSON输出。添加document_name、input参数，分别表示文档标题、文档正文内容。

    说明

    本示例中大模型节点的输出参数名称和类型必须与插件节点的输入参数完全对齐，否则变量引用时将无法匹配，导致插件调用失败。

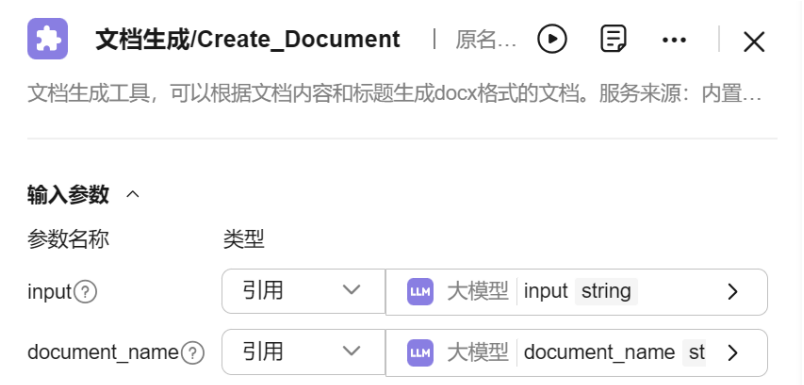
版权所有 © 华为云计算技术有限公司

图 4-109 配置大模型节点



- 文档生成插件节点的输入参数引用大模型节点的输出参数。

图 4-110 配置插件节点



- 结束节点引用插件节点的输出参数。

图 4-111 配置结束节点



相关操作

- 单击节点名称，可修改节点名称。
- 单击⋯图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击▷图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。

版权所有 © 华为云计算技术有限公司

- 单击图标，可查看文档信息。

常见问题

- **引用上游节点变量后，节点运行提示“参数类型不匹配”**

  检查上游输出类型与插件要求类型是否一致，例如插件要求Array，但上游输出的是一个String类型的参数。

- **必填参数已配置，但仍提示“缺少必填参数”**

  引用的上游变量在运行时为空值（null或空字符串），虽然配置了引用，但上游节点在某些分支下未产出该变量。请检查上游节点的输出是否正常。

  引用了错误的变量，变量名看起来相似但实际指向了另一个字段，请仔细核对引用路径。

- **手动输入固定值时，Array或者Object类型报错格式错误**

  Array和Object类型的固定值必须使用合法的JSON格式。

  表 4-48 JSON 格式示例

参数类型	正确写法	常见错误写法
Array	["地点1", "地点2"]	地点1, 地点2
Array（数字）	[1, 2, 3]	1, 2, 3
Object	{"city": "地点1", "count": 5}	city: 地点1, count: 5
嵌套结构	{"list": [1, 2], "info": {"key": "val"}}	缺少引号或括号不配对

- **插件调用超时如何处理**

  检查网络连通性和目标服务的响应速度。如果目标服务确实需要较长响应时间，可在异常处理中将超时时间调大（最大900秒）。如果超时频繁发生，建议检查目标API是否存在性能问题。

- **插件鉴权配置失败如何排查**

  a. 确认鉴权信息（如API Key、Token）是否正确且未过期；

  b. 确认鉴权信息的格式是否符合插件要求（如是否需要加Bearer前缀）。

- **插件返回数据结构与输出参数定义不符怎么办**

  输出参数从插件元信息自动导入，如果API实际返回的字段与定义不一致，需要回到“组件库 > 插件”修改插件的响应参数定义，使其与API实际返回结构对齐。

- 插件节点是否收费

  执行插件节点执行本身不额外收费，但以下情况可能产生费用：

  a. 部分插件广场插件由提供方收费；

  b. API类型插件调用外部服务产生第三方费用；

  c. 循环或重试场景下费用将按次数累加。

 版权所有 © 华为云计算技术有限公司

4.11.2 MCP 服务

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，允许工作流通过调用MCP服务来执行特定功能任务。

MCP服务与插件的关系：两者都是为工作流引入外部能力的方式。插件采用平台定义的封装规范，开发者可将外部API或自定义函数封装为插件使用；MCP服务则基于业界标准协议，能够接入更广泛的第三方工具生态。在功能上两者互补，开发者可根据实际需求选择使用。

表 4-49 MCP 与插件节点对比

维度	MCP节点	插件节点
协议标准	基于MCP开放协议标准（业界通用）	基于平台自定义封装规范
生态范围	可接入开源社区、第三方市场的广泛MCP工具生态	平台预置插件 + 开发者自建插件
配置方式	安装MCP服务后直接使用，工具参数自动导入	需配置输入输出参数映射和请求头
适用场景	需要快速接入开源社区MCP工具、标准化协议对接	需要封装企业内部API

选型建议：

- 如果您需要使用开源社区的现成MCP工具（如联网搜索、文件操作等），推荐使用MCP服务。
- 如果您需要封装企业内部RESTful API或自定义函数逻辑，推荐使用插件。
- 如果MCP和插件都能满足需求，优先选择MCP服务，可复用更广泛的生态资源。

MCP服务节点的执行原理：MCP节点执行时，平台会根据MCP服务的协议配置，将输入参数按MCP协议规范组装请求并发送给MCP服务端，等待MCP服务端处理完成后，将返回数据映射为输出参数传递给下游节点。

数据流转过程：上游节点输出 → 输入参数映射 → 平台按MCP协议发起调用 → MCP服务端处理并返回 → 输出参数映射 → 下游节点获取

添加节点

平台支持三类MCP创建方式，均支持在平台内便捷配置调用，详细MCP介绍请参考**MCP介绍**：

版权所有 © 华为云计算技术有限公司

表 4-50 MCP 类型

接入方式	说明	获取方式	选型建议
平台精选MCP	平台官方上架到资产广场的标准化服务。	资产广场一键获取，免去自主托管与底层代码开发。接入方法请参考**基于模板创建MCP**。	若需调用通用成熟工具（如Bing联网搜索等），推荐优先选择平台精选MCP，极速接入。
自定义接入MCP	支持自主接入开源社区或自行开发的标准服务（自定义创建）。	在平台内通过提供MCP服务端运行地址完成接入，接入方法请参考**自定义接入MCP**。	当精选库无法满足业务，且您拥有自研或开源社区中基于MCP协议托管的工具时推荐使用。

- **自定义接入MCP**（空白创建）：适用于接入开源社区的MCP或自主开发的MCP服务。当平台精选与第三方MCP服务无法满足业务需求时，可通过自定义MCP服务接入。接入方法请参考**自定义接入MCP**。
- **平台精选MCP**：平台官方上架到资产广场的标准化MCP服务，用户可直接添加使用（部分MCP需要配置鉴权参数后才可使用），无需额外配置开发，实现快速接入调用。接入方法请参考**基于模板创建MCP**。

鉴权区分：在安装MCP时，请务必先区分该服务是否需要配置鉴权参数。您可参考下图所示的配置项进行区分，避免因鉴权缺失导致调用报错。

图 4-112 MCP 鉴权区分方式



说明

使用MCP服务调用外部工具时，可能产生第三方服务的调用费用，该费用由第三方服务方收取，与平台费用无关。特别是从第三方安装的MCP服务（如开源社区的MCP服务），可能存在订阅费用或按调用次数计费，请在安装前确认费用规则。

配置 MCP 服务节点

参数只读限制：MCP服务节点的输入/输出参数完全由所选工具元信息动态生成，在工作流节点内为只读状态，不支持直接在工作流中修改参数名或添加参数项。

配置MCP节点的详细参数说明如下：

版权所有 © 华为云计算技术有限公司

工具

如果MCP内包含多个工具，可以选择所需的工具进行使用。

如果当前接入的MCP服务内包含多个子工具，可以在下拉列表中选择本次节点需要执行的具体工具。选择不同的工具后，输入参数会自动切换为该工具定义的参数结构。如果之前已配置了输入参数的引用或固定值，切换工具后需要重新配置，否则可能导致参数缺失或类型不匹配。

输入参数

为MCP所需的入参指定数据来源。输入参数的名称和类型从MCP服务元信息中自动导入，无需手动添加。选择不同的工具后，输入参数会自动切换为该工具定义的参数结构。

每个输入参数需要指定数据来源，支持两种方式：

- 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。前置节点为开始节点时，可以引用其输入参数。
- 输入：支持用户输入固定值。

当输入参数选择“引用”方式时，所引用的上游节点输出变量的类型必须与MCP服务定义的输入参数类型一致或兼容，否则可能导致节点运行报错或数据解析异常。

输出参数

定义节点运行完毕后向下游传递的数据。输出参数的所有信息从MCP服务元信息中自动导入，无需手动修改。MCP节点运行时会调用工具处理输入参数，并返回处理后的数据供下游节点使用。

异常处理

支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。

- “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。
- “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。
- “异常处理方式”：配置异常处理方式。
  - 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。
  - 返回设定内容：节点发生异常后，工作流不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。
  - 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点后新增节点，并为新增的异常分支配置相应的处理流程。

配置示例

本示例使用预置的Bing MCP搭建具备联网搜索功能的工作流。用户提问问题后，会先经过Bing MCP进行搜索，再由大模型进行答案总结。

版权所有 © 华为云计算技术有限公司

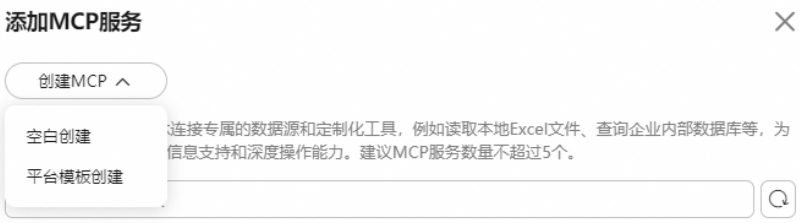
图 4-113 效果示例



- **添加MCP服务**

  在工作流编辑页面添加MCP服务节点，如果之前已经安装过Bing MCP，可以直接选择该MCP。如果先前未安装，选择“空白创建”，按照页面提示创建Bing MCP，该MCP无需鉴权，可以直接安装使用。

  图 4-114 添加 MCP 服务

- **变量引用**
  - 依次连接开始节点、Bing MCP节点、大模型节点、结束节点。
  - Bing MCP节点引用开始节点的输入，大模型节点的输入参数引用Bing MCP的输出参数，结束节点引用大模型节点的输出。

  图 4-115 配置节点变量引用

  - 试运行并验证

    单击“试运行”，输入测试问题（如“最近有什么新闻”），运行工作流。

    验证要点：

    - **Bing MCP节点成功调用，返回了搜索结果。**
    - **大模型节点基于搜索结果生成了回答，而非凭空编造。**
    - **结束节点输出包含搜索来源的总结内容。**

版权所有 © 华为云计算技术有限公司

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

- **MCP安装失败**

  如果MCP安装失败，多数原因是MCP本身的问题，如开源社区的MCP维护不及时或后端服务失效。遇到安装失败的情况，请尝试更换MCP来源。

- **引用上游变量后，节点运行报错“参数类型不匹配”**

  上游输出类型与MCP工具要求类型不一致；MCP中如果包含多个工具，选择工具后参数类型会发生变化，需要重新配置参数变量。

- **MCP鉴权配置失败如何排查**

  1. 确认鉴权参数（如API Key、Token）是否正确且未过期；2. 确认鉴权参数的格式是否符合MCP服务要求（如是否需要加Bearer前缀）；3. 确认MCP服务端是否正常运行且网络可达；4. 如果是从第三方安装的MCP，确认订阅状态是否有效。

- **MCP服务调用超时如何处理**

  检查网络连通性和MCP服务端的响应速度。如果MCP服务确实需要较长响应时间，可在异常处理中将超时时间调大（最大900秒）。如果超时频繁发生，建议检查MCP服务端是否存在性能问题，或考虑更换MCP来源。

4.11.3 HTTP 请求

HTTP请求节点用于在工作流中调用外部服务的API接口，支持POST、GET请求。通过该节点，可以向指定的服务器地址发送请求，获取数据或提交信息，从而将工作流与外部系统打通，实现查询第三方平台的信息、向业务系统写入处理结果、调用外部模型服务等功能。

HTTP请求节点执行时，平台服务器根据配置的URL和请求方式，将输入参数拼装为HTTP请求并发送给目标服务器，等待响应后将返回数据自动映射为输出参数传递给下游节点。请求发出后等待响应，最长等待时间为50秒，超时将返回错误。

此外，HTTP请求节点还提供了异常忽略功能保证API服务异常情况下工作流正常执行。

HTTP请求节点的核心能力：

- 调用外部API：向支持HTTP协议的服务发送请求，获取或提交数据。
- 动态参数拼接：请求路径、请求参数、请求体中的内容均可以引用上游节点的输出，实现根据实际数据动态构建请求。
- 安全鉴权：支持API Key认证方式，密钥经加密存储，保障接口调用安全。
- 容错处理：提供异常忽略功能，即使请求失败也不会中断整个工作流。

在工作流中，HTTP请求节点、插件节点与MCP服务节点均可用于引入外部系统能力，您可以根据以下维度进行选型：

版权所有 © 华为云计算技术有限公司

表 4-51 选型说明

维度	HTTP请求节点	插件节点	MCP节点
适用场景	快速调用单个API接口，配置灵活。	将多个API封装为标准化插件，参数自动导入。	接入基于MCP协议的开源工具与生态。
配置方式	手动配置URL、参数结构、鉴权。	参数从插件元信息中自动导入。	参数从MCP服务定义中自动导入。
鉴权支持	仅支持API Key认证。	支持多种鉴权（API Key、OAuth等）。	支持标准的MCP协议鉴权。
灵活性	高（可自由配置任意第三方API）。	中（需先在组件库中创建并定义插件）。	中（需先安装并配置对应的MCP服务）。
复用性	低（每次使用均需重新在节点内配置）。	高（插件可在多处工作流中重复调用）。	高（MCP服务可在多处工作流中重复调用）。

选型建议：

- 如果您需要快速调用一个外部API接口且无需复用，推荐使用HTTP请求节点。
- 如果您需要将接口封装为供团队内多处引用的标准化工具，或需要复杂的鉴权方式（如OAuth），推荐使用插件节点。
- 如果您需要接入MCP协议的开源社区工具，推荐使用MCP服务节点。

约束限制

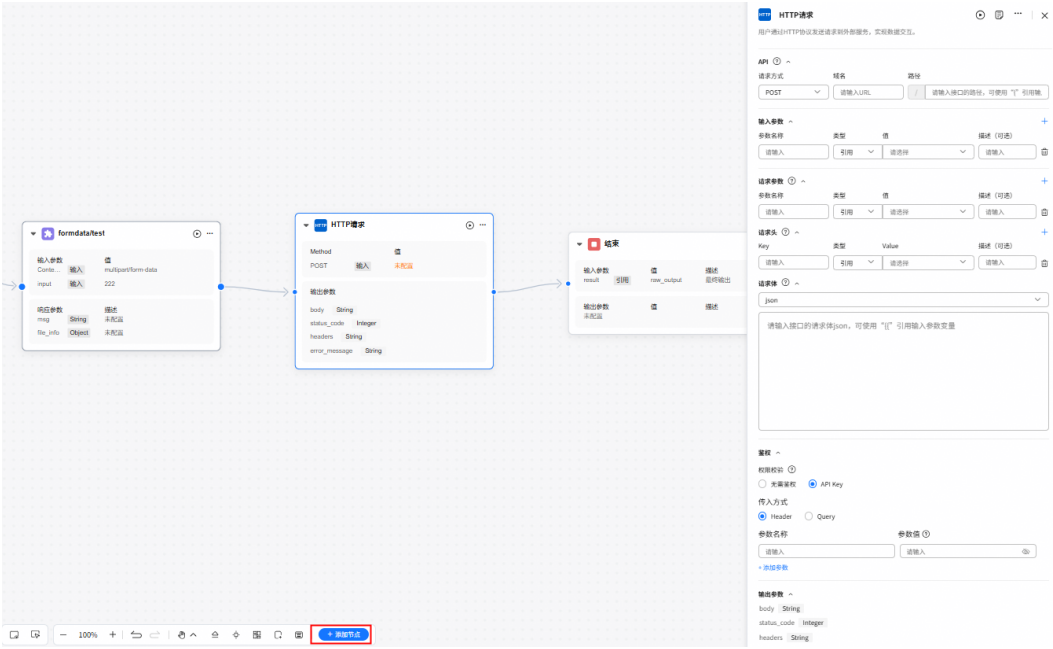
- 当前仅支持POST、GET两种请求方法。
- 鉴权方式仅支持API Key认证，暂不支持OAuth等其他鉴权协议。
- 不支持流式接口。
- 域名参数需填写为固定值，不支持动态引用变量。详情请参考**常见问题**。
- HTTP请求节点调用的API响应时间不能超过50秒。

添加节点

在工作流编辑页面，单击下方“添加节点”，选择“HTTP请求”节点。

版权所有 © 华为云计算技术有限公司

图 4-116 添加 HTTP 请求节点



配置 HTTP 请求节点

配置HTTP请求节点的详细参数说明如下：

API

用于指定请求的目标地址和方法，决定“向哪里发送什么类型的请求”。API配置包含请求方式、域名和路径：

- 请求方式：选择HTTP请求的方法。GET：从外部服务获取数据，适用于查询类操作。POST：向外部服务提交数据，适用于创建、更新类操作。
- 域名：请求目标服务器的域名地址（如https://api.example.com）。此参数为固定值，不支持动态引用变量。
- 路径：请求的具体接口路径（如v1/chat/completions）。支持动态引用变量——通过 {{变量名}} 的方式嵌入输入参数，在运行时自动替换为实际值。

输入参数

输入参数用于定义可在路径或请求体中动态引用的变量，实现请求内容随上游数据变化而变化。

输入参数支持引用和输入两种传参方式。

- 引用：选择工作流中前置节点的输出值或全局配置中的记忆变量。适用于参数值需要随工作流运行动态变化的场景。
- 输入：自定义填写固定值。适用于参数值确定不变、但希望集中管理便于后续修改的场景。
- 描述：（可选）用于说明该输入参数的具体含义。

版权所有 © 华为云计算技术有限公司

请求参数

请求参数是附加在URL末尾的键值对（如 ?keyword=手机&page=1），用于向服务器传递查询条件或筛选信息。

请求参数支持引用和输入两种传参方式。

- 引用：选择前置节点的输出值或全局配置中的记忆变量。适用于参数值由上游节点动态生成的场景。
- 输入：自定义填写固定值。适用于参数值固定不变的场景。
- 描述：（可选）用于说明该请求参数的具体含义。

请求头

请求头用于携带请求的附加元信息，例如指定数据格式（Content-Type）、标识客户端类型（User-Agent）等。大多数API都要求在请求头中传递特定的字段。

请求头支持引用和输入两种传参方式。

- 引用：选择前置节点的输出值或全局配置中的记忆变量。
- 输入：自定义填写固定值。
- 描述：（可选）用于说明该请求头的具体含义。

请求体

请求体是HTTP请求中携带的数据内容，通常在API请求中用于向服务器提交结构化数据。支持none和json两种配置模式：

- none：不携带请求体。通常用于GET请求，或不需要提交数据的场景。
- json：以json格式编写请求体。可通过{{变量名}}引用输入参数实现动态拼装。引用字符串类型的变量时，无需在变量外额外添加引号，系统会自动处理类型信息。

鉴权

鉴权方式用于配置调用外部API时的身份认证信息，确保请求能够通过目标服务的安全验证。

- 无需鉴权：目标API不要求身份认证时选择此项，无需配置额外信息。
- API Key：通过API密钥进行身份认证。发起请求时，系统会自动将密钥附加到请求头或URL参数中。选择API Key鉴权时，需要配置以下参数：
  - 传入方式：指定密钥附加的位置。Header：将密钥放入请求头中传递。Query：将密钥作为URL查询参数传递。
  - 参数名称：API Key对应的字段名称（如Authorization、api-key等），由目标API指定。
  - 参数值：API Key的具体密钥值。

关于密钥加密：鼠标移动至“参数值”右侧的 上，可查看密钥的加密方式。平台默认使用**KMS加密 + 默认密钥**方式对认证密钥进行加密存储，防止信息泄露。

- 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。

版权所有 © 华为云计算技术有限公司

- 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。

输出方式

HTTP请求执行完成后，系统会自动生成以下输出参数，供下游节点引用：

- body：HTTP响应的主体内容，即接口返回的实际数据。
- status_code：HTTP响应状态码，表示请求的执行结果。常见值：200表示成功，400表示请求格式错误，401表示认证失败，404表示接口地址不存在，500表示服务器内部错误。
- headers：HTTP响应的头信息，包含响应的元数据（如内容类型、编码方式等）。
- error_message：请求失败时返回的错误描述信息，帮助定位失败原因。请求成功时此字段为空。

异常配置

在实际使用中，外部API可能因网络波动、服务不可用等原因偶尔请求失败。如果您不希望单个请求的失败导致整条工作流中断，可以开启异常忽略功能。

开启后，当该节点执行失败时，工作流不会中断，而是跳过此节点继续运行后续下游节点。如果下游节点引用了此节点的输出，则使用预先配置的默认输出内容代替实际返回值。

说明

开启异常忽略时，建议同时为输出参数配置合理的默认值，避免下游节点因接收到空值而产生二次异常。

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

- **域名可以动态设置吗**

  不可以。域名参数仅支持固定值输入，不支持引用变量。如果您的业务需要根据不同条件调用不同域名的服务，建议通过条件分支节点拆分为多个HTTP请求节点，每个节点配置不同的固定域名。

- **请求路径中如何使用动态参数**

  在路径字段中使用 {{变量名}} 语法引用输入参数。例如，需要根据用户ID查询详情：

  a. 添加输入参数user_id，引用上游节点输出的用户ID。

版权所有 © 华为云计算技术有限公司

b. 路径设置为 /api/v1/users/{{user_id}}/profile。

运行时，{{user_id}} 会被替换为实际的用户ID值。

- **如何判断请求是否成功**

  通过输出参数status_code判断。您可以在HTTP请求节点的下游添加条件分支节点，判断status_code是否等于200：

  - 等于200：走正常处理分支，引用body中的返回数据。
  - 不等于200：走异常处理分支，引用error_message查看具体错误原因。

- 接口要求的鉴权方式不是API Key怎么办

  当前HTTP请求节点仅支持API Key认证。如果目标接口要求OAuth、Bearer Token等其他鉴权方式，可以尝试以下变通方案：

  - 手动添加请求头：部分鉴权方式（如Bearer Token）本质上是在请求头中传递固定格式的令牌。您可以在请求头配置中手动添加对应字段，例如参数名称设为Authorization，值设为具体的鉴权信息。
  - 使用插件或者MCP，如果目标API已经封装为插件或者MCP服务，可在工作流中添加插件、MCP。

4.12 消息管理节点

4.12.1 消息

消息节点是工作流中的**即时反馈机制**。与“结束节点”不同，消息节点允许工作流在执行过程中，向用户界面主动推送中间结果，而无需等待整个流程结束。

核心价值

- 降低等待焦虑：在执行耗时任务（如联网搜索、生成长文）前，先发送“正在搜索中...”的提示。
- 分步输出：将复杂的推理过程拆分为多个消息气泡，提升阅读体验。
- 调试与监控：在开发阶段输出变量的中间状态，用于逻辑排查。

消息节点使用场景：

表 4-52 消息节点使用场景

场景	示例	为什么不用结束节点
进度提示	工作流需要调用多个API，先告诉用户“正在查询中，请稍候…”	结束节点只能在最后输出，用户会长时间看到空白
流式输出	大模型生成长文本时，逐字逐句展示给用户（打字机效果）	结束节点等大模型完整输出后才一次性展示
兜底回复	意图识别走到“其他意图”分支时，返回一段默认话术	该分支无需复杂处理，消息节点比大模型节点更轻量灵活。
调试与日志	开发阶段输出中间变量值，验证工作流逻辑是否正确	调试信息不应出现在最终结果中

版权所有 © 华为云计算技术有限公司

场景	示例	为什么不用结束节点
分段回复	先回复文字摘要，工作流继续执行后续操作（如生成图表）	用户先看到文字，不必等全部结果就绪

配置消息节点

配置消息节点的详细参数说明如下：

输入参数

输入参数定义了从前置节点传入的变量，在指定回复中通过 {{参数名}} 引用。当选择回复内容为纯文本时，可以不配置输入参数。

当单击+图标时，可新增输入参数。

- 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：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置），以及系统参数。

    系统参数列举如下：

    - “conversation_history”：存储当前对话的历史消息记录。
    - “current_time”：获取当前的系统时间戳或时间字符串。
    - “user_id”：识别当前用户的唯一标识符。
    - “conversation_id”：当前工作流的唯一对话标识。
    - “dialogue_count”：当前对话的轮次数计数。
  - 输入：支持用户自定义取值。
- 描述：（可选）用于说明该输入参数的具体含义。

指定回复

撰写发送给用户的消息内容，支持纯文本或引用{{参数名}}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。

插入变量的方式如图4-117所示，想要消息节点输出检索到的知识切片的内容，则添加输入参数"rep_out"，引用前置知识检索节点的输出参数"content"，并在指定回复中引用{{rep_out}}即可；也可以单击 {x} 快速插入输入参数，或在输入框中键入' {' 快速引用参数。

会话写入历史控制消息节点的交互内容是否记录到对话历史中：

版权所有 © 华为云计算技术有限公司

选项	行为	适用场景
写入（默认）	消息节点输出都会写入会话历史。	一般用于记录用户与机器人的对话历史，以便模型能更好理解用户意图。
不写入	输出不写入会话历史。	一般机器与机器的交互历史无需记录，提升历史对话信息准确性，减少历史长度。

图 4-117 设置消息节点



如果开启下面的“结构化信息”，可在“结构化信息”中使用{{_NODE_OUTPUT}}引用指定回复。

结构化信息

回复消息为普通文本时无需开启。当需要前端应用（如Web SDK、小程序）接收JSON数据或渲染UI卡片或下游系统需要解析消息的结构化字段，开启“结构化信息”开关。开启后，节点将不再只发送文本，而是将“指定回复”的内容封装在JSON对象中返回。功能开启时，可使用{{_NODE_OUTPUT}}引用“指定回复”中的信息实现结构化输出。

示例：

假设前端不仅显示文本，还能获取原始数据用于图表渲染。

- **指定回复**：已为您生成分析报告，共包含{{count}}条数据。
- **结构化信息**(JSON格式)：

```
{
  "msg_type": "report_card",
  "display_text": "{{_NODE_OUTPUT}}",
  "data":
  {
      "report_id": "{{report_id}}",
      "timestamp": 1679000000
```


版权所有 © 华为云计算技术有限公司

```
 }
}
```

- **最终前端收到的数据包：**

```
{
  "msg_type": "report_card",
  "display_text": "已为您生成分析报告，共包含 12 条数据。",
  "data": { ... }
}
```

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以工作流调用外部API查询数据，耗时较长。在API调用前消息节点先向用户发送等待提示为例。

图 4-118 工作流



工作流核心节点如表4-53所示。

表 4-53 节点配置说明

节点类型	说明	示例
开始节点	添加一个必选的输入参数，名为ip，即用户需要查询区域的IP地址。	

 版权所有 © 华为云计算技术有限公司

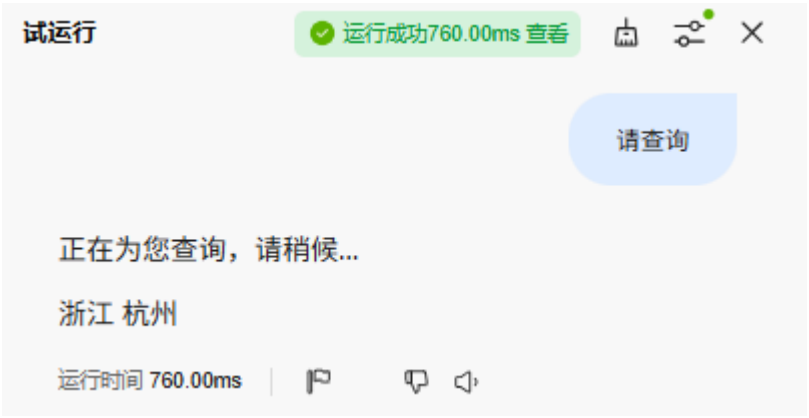
节点类型	说明	示例
消息节点	此节点用于向用户发送进度提醒消息。 **指定回复：** 正在为您查询，请稍候…	
插件节点-IP查询	此节点用于查询IP地址的归属地。 **输入参数：** ip：引用“开始”节点的输入参数ip。	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
结束节点	此节点用于输出“IP查询”插件节点的IP归属地查询结果。 **输入参数：** result：引用“IP查询”插件节点的输出参数area。	

在试运行配置界面输入IP地址，例如“10.0.0.0”，工作流运行的效果如图4-119所示，消息节点在调用API之前，给用户发送进度提醒“正在为您查询，请稍候…”。

图 4-119 效果示例



常见问题

- **消息节点和结束节点的区别是什么？**

  表 4-54 消息节点和结束节点区别

维度	消息节点	结束节点
触发时机	工作流执行到该节点时**立即发送**	工作流**全部执行完毕**后输出

版权所有 © 华为云计算技术有限公司

维度	消息节点	结束节点
出现次数	一个工作流中可以有**多个**消息节点	通常只有**一个**结束节点
是否阻断流程	**不阻断**，消息发送后工作流继续执行后续节点	**阻断**，输出后工作流结束
流式输出	支持引用流式参数（逐字展示）	不支持引用流式输出参数
典型用途	中间消息、进度提示、流式内容	最终结果、完整回复

- **为什么消息内容显示 {{order_id}} 原文而非变量值?**
  参数名拼写错误，或输入参数中未定义该变量。请检查“指定回复”中的 {{参数名}} 与“输入参数”中的名称是否完全一致（区分大小写）。
- **为什么消息内容为空?**
  引用的上游节点未输出结果（如API调用失败），请在上游节点中检查输出，确认变量有值。可在消息模板中加入固定文本兜底。
- **为什么结构化信息中 {{_NODE_OUTPUT}} 为空?**
  “指定回复”未填写内容，由于{{_NODE_OUTPUT}} 引用的是“指定回复”的渲染结果，需先填写指定回复内容。

4.12.2 输入

输入节点用于在工作流**执行过程中暂停流程**，向用户收集额外信息。当下游节点需要某些参数，而这些参数在之前的流程中尚未获取时，您可以插入一个输入节点来主动追问用户。

输入节点使用场景：

表 4-55 输入节点使用场景

场景	示例	为什么不在开始节点收集
缺失信息追问	用户说“帮我查天气”但没说城市，插件节点需要city参数	用户的初始输入中不一定包含所有必要信息
条件性信息收集	仅当用户选择“退款”时才需要收集订单号和退款原因	该信息仅在特定分支才需要，不应在开始节点强制收集
多步表单	贷款申请需依次收集：身份信息→收入信息→贷款金额	一次性收集太多信息会增加用户放弃率
二次确认	执行敏感操作前（如删除数据、提交订单），要求用户确认	确认动作必须发生在操作之前的特定位置
人工补充	大模型分析后发现信息不足，需要用户补充更多细节	信息缺口是在工作流执行过程中才发现的

版权所有 © 华为云计算技术有限公司

配置输入节点

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。

当单击+图标时，可新增输入参数。

- 参数名称：变量标识符，下游节点通过此名称引用，只允许输入字母、数字、下划线，且不能以数字开头。建议用业务语义命名：city、order_id、refund_reason。
- 参数类型：可选String、Integer、Number、Boolean、Object、Array、File类型。

表 4-56 参数类型说明

类型	说明	用户输入形式	典型用途	示例值
String	字符串文本	自由文本输入	城市名、姓名、描述、原因	"北京"
Integer	整数	数字输入	数量、年龄、ID	3
Number	数值（含小数）	数字输入	金额、比率、坐标	99.5
Boolean	布尔值 取值逻辑：空输入=false，非空输入=true；描述须明确true、false对应的具体业务场景。	是/否选择	确认操作、开关选项	true
Object	JSON对象	结构化输入	复合信息（地址、配置）	{"city": "北京", "district": "海淀"}
Array	数组/列表	多值输入	多选标签、批量ID	["标签A", "标签B"]
File	文件	文件上传	图片、文档、表格	上传的文件对象

- 描述：向用户展示的提示文字，引导用户理解需要输入什么。参数描述质量直接决定用户是否能正确输入。描述示例如**表4-57**所示，输入节点配置如**图4-120**所示。需要用户输入时，可查看输入参数描述，如**图4-121**所示。

版权所有 © 华为云计算技术有限公司

表 4-57 描述示例

分类	参数名	描述内容	理由
好的描述	city	用户询问某个城市或地区的天气情况，包括当前天气、未来天气预报、气温、下雨等	列出具体场景和常见关键词
	order_id	请输入您的订单号（8位数字，可在"我的订单"页面查看）	说明了格式+告诉用户去哪找
差的描述	param1	输入参数	用户完全不知道该输入什么
	city	城市	缺少引导——查天气的城市？发货的城市？用户所在城市？

图 4-120 配置输入节点



图 4-121 查看输入参数说明



- 必填：是否强制用户填写。建议核心参数设为必填，补充信息设为非必填。

版权所有 © 华为云计算技术有限公司

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以天气查询为例，用户需要查天气，但没提供需要查询的城市，因天气插件需要city参数，因此需要输入节点向用户收集需要查询的城市名。

图 4-122 工作流



工作流核心节点如表4-58所示。

表 4-58 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
输入节点	此节点用于向用户收集需要查询的城市名。 **输入参数：** 单击+图标，新增输入参数。 ● 参数名称：city ● 参数类型：String ● 描述：用户询问某个城市的天气情况，包括当前天气、气温等 ● 必填：勾选	
插件节点-高德天气	此节点根据用户输入的城市名，查询当前天气情况。 **输入参数：** city：引用输入节点的“city”。	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
结束节点	此节点根据“高德天气”插件节点查询的结果，生成一段包含城市天气和气温的信息给用户。 **输入参数：** • city：引用“高德天气”插件节点的city。 • weather：引用“高德天气”插件节点的weather。 • temperature：引用“高德天气”插件节点的temperature。 **指定回复：** {{city}}当前天气为{{weather}}，温度为{{temperature}}°C	

效果示例如图4-123所示。

图 4-123 效果示例



版权所有 © 华为云计算技术有限公司

常见问题

- **输入节点和开始节点有什么区别?**

  表 4-59 输入节点和开始节点区别

维度	开始节点	输入节点
位置	工作流**最前端**，仅有一个	工作流**任意中间位置**，可有多个
触发时机	用户发起对话时**自动触发**	工作流执行到该位置时**主动暂停并追问**
信息类型	用户的**初始输入**（如自由文本query）	工作流**中途需要**的额外信息（如订单号、城市名）
是否阻断	不阻断（工作流从这里开始）	**阻断**工作流执行，直到用户输入完毕
适用场景	收集每次对话都需要的基础信息	收集特定分支/条件下才需要的补充信息

- **其他常见问题**

  表 4-60 其他常见问题

问题现象	可能原因	解决方案
工作流暂停后用户看不到追问提示	参数的描述字段为空	为每个参数填写清晰的描述文字，这是用户看到的追问内容
用户输入后工作流未恢复	存在未填写的**必填**参数	检查是否所有必填参数都已提交。适当减少必填参数数量
下游节点引用输入参数时值为空	工作流通过条件分支**绕过**了输入节点	确保输入节点在当前执行路径上。或在下游判断节点中增加空值处理逻辑
用户输入的格式不符合预期（如Integer参数输入了文字）	参数类型与用户实际输入不匹配	①在描述中明确告知格式要求 ②考虑使用String类型接收，在后续节点中做格式校验
参数名称报错	名称以数字开头或包含特殊字符	仅允许字母、数字、下划线，必须以字母或下划线开头
多个输入节点的参数同名，下游引用混乱	不同输入节点使用了相同的参数名（如都叫input）	使用不同的节点名称+参数名称组合，下游通过{{节点名.参数名}}区分

版权所有 © 华为云计算技术有限公司

问题现象	可能原因	解决方案
工作流一直暂停，用户无响应	用户离开了对话或不理解追问	①优化参数描述让追问更清晰 ②考虑在描述中提供默认值建议 ③评估是否可以将该参数改为非必填

4.12.3 提问器

提问器节点在工作流执行过程中**主动向用户发起提问**，从自然语言对话中提取完成任务所需的结构化参数。它会循环追问，直到收集到所有必需信息或达到最大交互轮数。

运作机制

提问器的执行流程如下：

1. 分析用户对话内容，提取已提供的参数；
2. 判断是否所有必需参数均已提取；
3. 如果有缺失，生成追问继续收集；
4. 重复1~3，直到所有参数提取完成或达到最大回复轮数。

约束与限制

提问器节点仅适用于对话型工作流。

配置提问器节点

配置提问器节点的详细参数说明如下：

模式偏好选择

可选择效果优先或速度优先，参数说明如表4-61所示。

版权所有 © 华为云计算技术有限公司

表 4-61 模式偏好选择

模式	适用场景	特性
效果优先	适合高精度场景，例如预订系统、表单填报、审批流程等参数准确性要求高的场景	• **开启时间增强**：将自然语言时间转换为标准格式YYYY-MM-DD HH:MM:SS，例如用户输入：明天下午两点半，则提取为2026-03-09 14:30:00。 • **开启反思**：模型对提取结果进行二次校验，发现错误时自动纠正或继续追问。例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。 • 延迟较高（因为模型多次推理）。
速度优先	适合低延迟场景，例如闲聊助手、简单信息收集等对速度敏感、容错率高的场景	无时间增强，无反思，响应极快。

模型配置

选择用于参数提取和反思纠正的模型。建议选择具备良好指令遵循能力的模型。

您还可以单击图标，调整模型配置。模型配置的详细说明可参考**选择并配置模型**。

输入参数

输入参数用于向提问话术中注入**动态数据**，通常引用前置节点的输出。

单击图标时，可新增输入参数，单击图标时，可删除输入参数。

- 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：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置），以及系统参数。

    系统参数列举如下：
    - “conversation_history”：存储当前对话的历史消息记录。
    - “current_time”：获取当前的系统时间戳或时间字符串。
    - “user_id”：识别当前用户的唯一标识符。
    - “conversation_id”：当前工作流的唯一对话标识。
    - “dialogue_count”：当前对话的轮次数计数。
  - 输入：手动填写固定值，选择"输入"类型后，填写的固定值将直接传递给大模型，前置节点的输出内容不会被引用。
- 描述：（可选）用于说明该输入参数的具体含义。

版权所有 © 华为云计算技术有限公司

示例：以会议室预订场景为例，假设前置节点“查询空闲会议室”输出了可用会议室列表，提问器节点定义输入参数“available_rooms”，引用前置节点“查询空闲会议室”的输出变量值{{room_list}}。“问题”配置为：有以下几个会议室供您选择：{{available_rooms}}，请选择您想预订的会议室。

输出参数

输出参数定义了提问器需要从用户对话中提取的结构化数据。

- 默认输出参数：每个提问器节点自动包含以下系统参数，如表4-62所示。

表 4-62 默认输出参数

参数名	类型	说明
USER_RESPONSE	string	用户的原始回复文本（未经处理）
STATUS	integer	提取状态码，用于下游节点判断提取结果。状态码的含义如下： ● 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（仅“效果优先”模式可能出现）。

- 自定义提取参数：开启“参数提取”功能后，可定义需要从用户回复中提取的具体参数。

表 4-63 参数说明

配置项	说明	示例
参数名称	只允许输入字母、数字、下划线、短横线。	meeting_room
展示名称	用于追问话术中展示（不可为空）	会议室
类型	输出参数的数据类型，可选String、Integer、Number、Boolean。	String
默认值	所有追问结束但仍未提取到参数时，使用的兜底值	A101

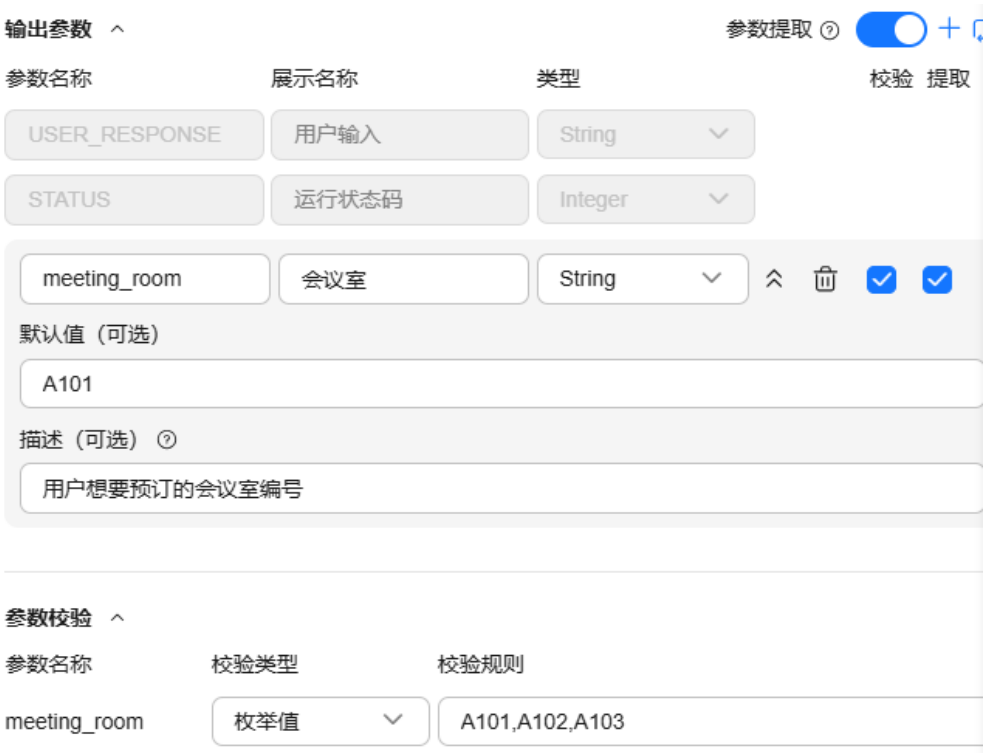
版权所有 © 华为云计算技术有限公司

配置项	说明	示例
描述	对参数的详细描述，帮助大模型理解参数含义，提升提取准确率	用户想要预订的会议室编号
校验	限定参数的可选值范围。 ● 勾选：可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。校验不通过时，提问器会自动追问用户重新输入。 ● 不勾选：对输出参数不进行校验。	例如，会议室预订场景，用户可以选择预订的会议室有A101、A102和A103。 则校验类型选择“枚举值”，校验规则填写为“A101,A102,A103”。
提取	● 勾选：该参数为必填项（必须提取到或使用默认值）。 ● 不勾选：允许为空。	开启

示例：以会议室预订场景为例，需要询问用户预订哪个会议室，可选会议室有A101、A102和A103。如果用户反馈预订A102，则meeting_room参数值提取为“A102”；如果用户反馈B203，由于不满足校验规则，提问器节点会继续追问用户需要预订哪个会议室；经过多轮追问后，如果用户没有反馈有效的会议室，则meeting_room参数值根据配置的默认值提取为“A101”，输出参数配置如图4-124所示。

版权所有 © 华为云计算技术有限公司

图 4-124 配置输出参数



- 引用插件参数：如果您提取参数与已有插件的请求参数一致，可以单击按钮，从“插件广场”或“我的插件”中选择插件，一键导入参数定义和校验规则。

问题配置

配置提问器向用户展示的问题文本，也可以单击 {x} 快速插入输入参数，或在输入框中键入“{{}}”快速引用参数，支持引用多个参数。

版权所有 © 华为云计算技术有限公司

表 4-64 问题配置

配置项	说明
问题	首次提问时展示给用户的文本。 • 配置问题文本：问题文本原样展示给用户，可通过Jinja语法在问题中使用输入参数。最大支持输入3000个字符。 示例1：请问你的名字是什么？ 示例2：请问你是哪个班级的，可选班级有{{classes}}（classes先在输入参数配置好）。  • 未配置问题文本：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。
最大回复轮数	与用户的最大交互次数。超过后提问器退出，未提取参数使用默认值或置空。

高级配置

- **允许用户退出交互**：开启后，如果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。
- **输出参数确认**：开启后，提问器在参数提取完毕后会请求用户进行确认。
- **提取约束**：
  为大模型提供额外的约束信息，提升参数提取的准确性，例如指定被提取参数的格式要求。
  示例如下：
  - 用户希望提取电话号码tel_number，提取约束可以写：tel_number必须是11位数字，以1开头。
  - 用户希望提取会议日期meeting_date，提取约束可以写：meeting_date格式为YYYY-MM-DD，不接受过去的日期。

  用户可单击图标直接引用提示词模板；也可以自行撰写提取约束，撰写后可单击图标对当前约束进行语义优化，提升提取效果；也可单击图标保存到模板，在“组件库>提示词”中可查看该模板；也可以单击快速插入输入参数，或在输入框中键入“{{}}”快速引用参数，支持引用多个参数。
- **追问模式**：
  当用户回复未能提供所有必需参数时，提问器会发起追问。三种模式区别如下：

版权所有 © 华为云计算技术有限公司

  - 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：（推荐）由大模型根据缺失字段自动生成自然的追问，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}是系统占位符，不可修改或删除，系统会自动替换为尚未提取的参数中文名称。每次追问内容相同。

  示例：以要提取“姓名”和“年龄”，用户仅提供了名字为例，三种模式的追问如下：

  - 默认：请您提供年龄相关的信息。
  - 智能追问：哎呀，您还没告诉我年龄呢。（模型生成，内容不固定。）
  - 自定义追问：请补充以下信息：年龄。（按模板生成）

- **追问显示枚举值**：

  开启后，如果参数设置了枚举值校验，追问时自动展示参数的可选枚举值，帮助用户快速选择。

  例如，输出参数meeting_room配置了枚举值校验A101,A102,A103，开启后追问为：请选择会议室，可选项为：A101,A102,A103。

- **示例配置**：

  给大模型一段参数提取的示例，增强大模型对参数提取场景的理解。

  示例格式：

  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"}

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以根据输入的文档内容和文档标题，自动生成一个可以下载的word格式的文档为例。

图 4-125 工作流



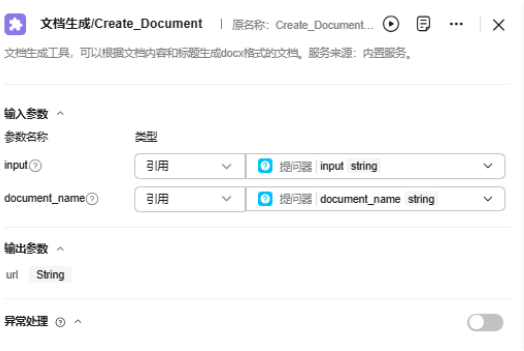
工作流核心节点如表4-65所示。

 版权所有 © 华为云计算技术有限公司

表 4-65 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	
提问器节点	此节点用于向用户提问，需要用户提供文档内容和文档标题。 **模型配置：** DeepSeek-V4-Flash **输入参数：** input：引用“开始”节点的query。 **输出参数：** 单击按钮，从“插件广场”选择“文档生成”插件，一键导入参数定义和校验规则。 **问题配置：** 您好，为了能更好地协助您处理相关文档，可以告诉我您想处理的文档名称和文档内容吗? **高级配置：** 追问模式：勾选“智能追问”。	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
插件节点-文档生成	此节点基于用户提供的文档内容和文档名称，生成一个可以下载的word格式的文档。 **输入参数：** • input：引用“提问器”节点的输出参数input。 • document_name：引用“提问器”节点的输出参数document_name。	
结束节点	此节点用于输出文件解析的结果。 **输入参数：** result：引用“文档生成”插件节点的输出参数URL。	

工作流运行的效果如**图4 效果示例**所示，“您好，为了能更好地协助您处理相关文档，可以告诉我您想处理的文档名称和文档内容吗？”，用户发送文档名称后，提问器追问，需要用户提供文档内容。

版权所有 © 华为云计算技术有限公司

图 4-126 效果示例



常见问题

- **什么时候用“提问器”，什么时候用“问答节点”配合“循环”？**
  - 提问器：适合非线性的信息收集。用户可以说“订明天的票，两个人”，一次性收集time和people两个参数。
  - 问答+循环：适合线性的问卷调查。必须按顺序问完问题1再问问题2，不允许跳跃。
- **模型总是提取错误的参数值，怎么办？**

  按优先级尝试以下方法：

  a. 切换为**效果优先**模式，开启反思功能。
  b. 在**提取约束**中增加明确的格式要求和取值范围。
  c. 在**示例配置**中添加1–2个典型的提取示例。
  d. 优化输出参数的**描述**，让大模型更准确理解参数含义。
  e. 尝试更换能力更强的模型。
- **用户被困在提问循环中无法退出？**

  在高级配置中开启**允许用户退出交互**。开启后，用户表达中止意图时提问器自动结束（STATUS=100）。
- **如何让用户在提交前确认提取结果？**

  在高级配置中开启**输出参数确认**。所有参数提取完毕后，提问器会展示结果摘要并请求用户确认。
- **下游节点如何判断提取是否成功？**

  在提问器下游添加**条件分支节点**，根据提问器节点的输出参数{{STATUS}}路由到不同处理逻辑：
  - 如果STATUS=0或10，则正常处理。
  - 如果STATUS=100或101，则部分结果兜底处理。
  - 如果STATUS>=201，则错误处理或重试。

版权所有 © 华为云计算技术有限公司

- **STATUS为100/101时，未提取到的参数值是什么?**
  - 如果配置了默认值，则使用默认值。
  - 如果未配置默认值，则参数值为空（需下游节点做判空处理）。
- **反思模块报错（STATUS=202）如何处理?**

  “STATUS=202”仅在“效果优先”模式下可能出现，表示反思校验环节异常。此时提取结果可能不准确，建议：
  - 在下游条件分支中将202路由到降级处理逻辑。
  - 或切换为“速度优先”模式绕过反思环节（牺牲准确率）。

4.12.4 问答

问答节点用于在工作流执行过程中向用户提出预设问题。根据配置模式的不同，它既可以暂停工作流等待用户回答（需要回答模式），也可以仅发送问题后继续执行（无需回答模式）。

问答节点使用场景：

表 4-66 问答节点使用场景

场景	模式	示例
引导式信息收集	需要回答	向用户提问“请问您的预算范围是？”，收集回答后传递给推荐节点
多轮对话追问	需要回答	大模型节点判断信息不足后，通过问答节点追问用户缺失的关键信息
随机互动	需要回答	从多个预设问题中随机选一个向用户提问（如知识问答、用户调研）
主动消息推送	无需回答	向用户展示一条提示信息或引导语，不等待回答直接继续
分步引导	无需回答	在复杂流程的每个阶段向用户展示当前步骤的说明

配置问答节点

配置问答节点的详细参数说明如下：

问题配置

- 问答策略：
  - 需要回答：工作流执行到问答节点时**暂停**，向用户展示问题，等待用户回答后将回答内容作为输出参数传递给下游节点。
  - 无需回答：工作流执行到问答节点时**不暂停**，将问题文本本身作为输出参数传递给下游节点，同时向用户展示该问题（类似消息节点的行为）。

场景	说明
引导语/提示	在复杂流程的每个阶段向用户展示当前步骤说明。

版权所有 © 华为云计算技术有限公司

场景	说明
问题文本传递	将问题文本作为输出变量传给下游大模型节点，让大模型基于该问题生成更详细的引导
结构化信息输出	配合结构化信息功能，向前端推送带格式的问题卡片

- 问题选择策略：

  当在问答节点中配置了**多条问题**时，需要选择问题的选择策略：

  - 随机选择

    每次执行时，从所有预设问题中**随机**选取一条展示给用户。

适用场景	示例
用户调研	随机抽取一道调研问题（避免每次问同一个问题）
知识问答游戏	从题库中随机出题
多样化开场白	随机选择不同的欢迎语或引导语，提升对话新鲜感

  - 指定index

    通过输入参数index精确指定展示第几条问题。

配置项	说明
index起始值	从0开始（第一条问题index=0，第二条=1，以此类推）。
index来源	可引用上游节点的输出变量，也可手动输入固定值。

适用场景	示例
分步引导	根据当前步骤编号展示对应的引导问题。
条件性提问	上游判断节点输出分支编号，问答节点根据编号展示对应问题。
顺序提问	配合循环或计数变量，按顺序展示问题。

- 会话写入历史：

  控制问答节点的交互内容是否记录到对话历史中：

选项	行为	适用场景
写入（推荐默认）	问题和用户回答都会写入会话历史，下游大模型节点可通过历史对话轮次引用。	适用于用户对话类场景，大模型需要了解之前问了什么、用户答了什么。

版权所有 © 华为云计算技术有限公司

选项	行为	适用场景
不写入	问题和回答不写入会话历史。	适用于系统内部流转场景，问答内容仅用于参数传递，不需要被大模型作为上下文参考。例如：内部流程控制的确认步骤、调试用的临时追问。

问题内容

在此处预先编写问答节点要展示的问题，也可以在输入框中键入“{{}}”快速引用参数。支持配置多条问题（添加一条问题后，会自动生成第二条问题的文本框）。例如，调研用户对产品的意见，每次对话随机问一道调研题，问题编写示例如图4-127所示。

版权所有 © 华为云计算技术有限公司

图 4-127 配置问答节点



 版权所有 © 华为云计算技术有限公司

输入参数

固定为index，仅在“指定index”策略下生效。用于指定展示第几条问题（从0开始计数）。

设置参数值时，支持“引用”和“输入”两种方式。

- 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：手动填写固定值，例如填写0，则始终展示第一条问题。

同时支持用户自定义输入参数的具体含义。

输出参数：

节点的输出参数，可作为变量被后续节点引用。

response：在“需要回答”模式下，response的值为用户输入的回答文本；在“无需回答”模式下，response的值为被选中的问题文本本身。

结构化信息：

默认关闭。当集成的前端应用需要按消息类型做不同渲染（如问题卡片、选项按钮等）时，开启“结构化信息”开关。功能开启时，可使用{{_NODE_OUTPUT}}引用“问题内容”中的信息实现结构化输出。

说明

{{_NODE_OUTPUT}}是节点输出变量的引用语法，代表当前节点“问题内容”字段的值。

示例：结构化信息如下，前端应用可解析options数组渲染为快捷选项按钮，提升用户回答效率。

```
{
  "type": "question",
  "content": "{{_NODE_OUTPUT}}",
  "options": ["5000以下", "5000-10000", "10000以上"]
}
```

相关操作

- 单击节点名称，可修改节点名称。
- 单击···图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以电脑推荐场景为例，工作流通过问答节点询问用户的预算，再根据用户的预算推荐性价比高的笔记本电脑。

工作流核心节点如表4-67所示。

版权所有 © 华为云计算技术有限公司

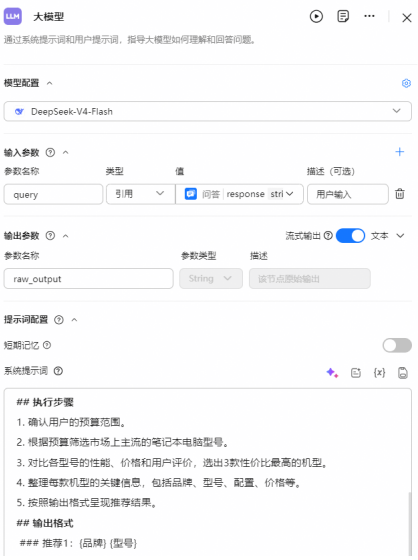
图 4-128 工作流



表 4-67 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	
问答节点	此节点询问用户购买笔记本电脑的预算范围。 • **问答策略：**需要回答 • **问题选择策略：**随机选择 • **会话历史写入：**写入 • **问题内容：**请问您购买笔记本电脑的预算范围是多少？例如5000元左右	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点	此节点用于根据用户输入的预算，推荐性价比高的笔记本。 • **模型配置：** DeepSeek-V4-Flash • **输入参数：** 引用“问答”节点的response。 • **系统提示词：** ## 人设 - 角色：笔记本电脑选购顾问 - 特长：熟悉各类笔记本电脑的性能、价格、品牌特点，能够根据用户预算和需求推荐合适的机型。 ## 任务描述 根据用户提供的预算范围，推荐3款符合预算且性价比高的笔记本电脑，帮助用户在预算内做出最佳选择。 ## 约束条件 1. 推荐的笔记本电脑价格必须在用户预算范围内。 2. 每款推荐需包含品牌、型号、关键配置（如CPU、内存、硬盘、显卡等）、价格和购买链接（可选）。 3. 推荐需基于当前市场行情和主流品牌的热门机型。 4. 按照<输出格式>输出。 5. 按照<执行步骤>一步一步执行。 ## 执行步骤 1. 确认用户的预算范围。 2. 根据预算筛选市场上主流的笔记本电脑型号。 3. 对比各型号的性能、价格和用户评价，选出3款性价比最高的机型。 4. 整理每款机型的关键信息，包括品牌、型号、配置、价格等。 5. 按照输出格式呈现推荐结果。 ## 输出格式 ### 推荐1：{品牌} {型号} - **推荐理由**： - **价格**：{价格} - **关键配置**： - CPU：{CPU型号} - 内存：{内存容量} - 硬盘：{硬盘容量} - 显卡：{显卡型号} - **购买链接**：{链接}（可选） ### 推荐2：{品牌} {型号} - **推荐理由**： - **价格**：{价格} - **关键配置**：	

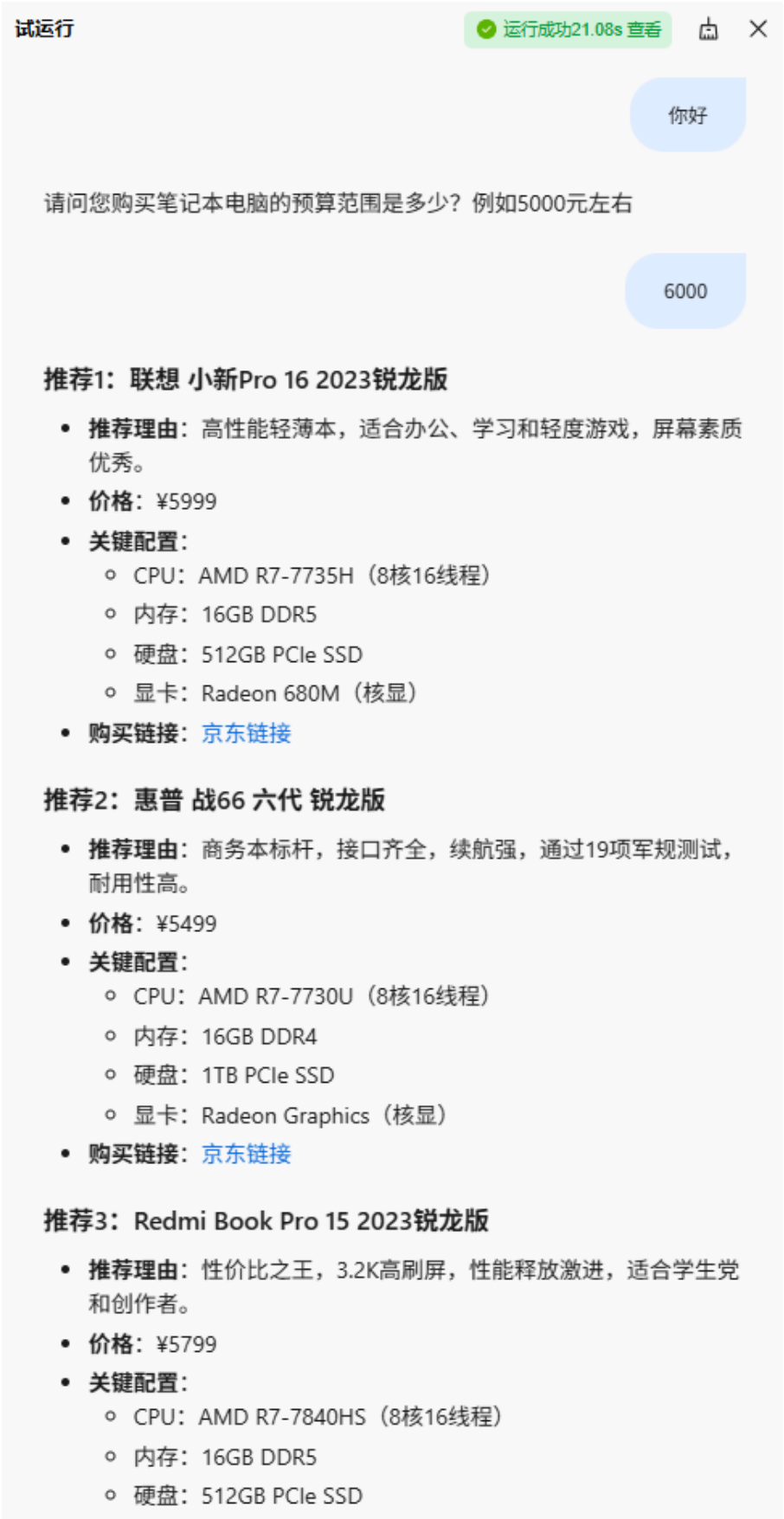
版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
	- CPU：{CPU型号} - 内存：{内存容量} - 硬盘：{硬盘容量} - 显卡：{显卡型号} - **购买链接**：{链接}（可选） ### 推荐3：{品牌} {型号} - **推荐理由**： - **价格**：{价格} - **关键配置**： - CPU：{CPU型号} - 内存：{内存容量} - 硬盘：{硬盘容量} - 显卡：{显卡型号} - **购买链接**：{链接}（可选）	
结束节点	此节点用于展示大模型节点的推荐结果。 • **输入参数：** result：引用大模型节点的raw_output。 • **指定回复：** {{result}}	

效果示例如下图所示。

版权所有 © 华为云计算技术有限公司

图 4-129 效果示例



 版权所有 © 华为云计算技术有限公司

常见问题

- **问答节点、输入节点和消息节点有什么区别?**
  这三个节点都涉及"与用户交互"，但定位和行为完全不同，如表4-68所示。

  表 4-68 问答节点、输入节点和消息节点区别

维度	问答节点	输入节点	消息节点
核心能力	提出**预设问题**并可选择等待回答	收集**结构化参数**（带类型约束）	发送**自由文本**消息
问题来源	在节点中预先配置多条问题文本	由参数定义的描述文字生成	由指定回复中的文本模板生成
是否暂停	“需要回答”模式暂停；“无需回答”不暂停	始终暂停，等待用户填写所有必填参数	不暂停，发完就继续
输入约束	用户自由文本回答（无类型校验）	按参数类型校验（String/Integer/Boolean…）	不收集用户输入
问题数量	支持配置**多条**问题，支持随机/指定选择	每个参数一条描述（不可配置多条）	仅一条固定消息
输出内容	需要回答：用户的回答文本 无需回答：问题本身	用户输入的参数值	无输出（或通过结构化信息输出）
典型场景	对话式追问、随机提问、引导语	表单式信息收集、敏感操作确认	进度提示、中间结果展示

  选型指南:
  - 需要收集**结构化数据**（有类型、必填要求），用**输入节点**。
  - 需要用**预设的自然语言问题**追问用户，用**问答节点**（需要回答）。
  - 只需要**发一条消息**、不需要用户回应，用**消息节点**或**问答节点**（无需回答）。

- **问答节点的“无需回答”和消息节点的区别?**
  消息节点的核心是输出**自由编写的文本**（支持变量插值）。
  问答节点（无需回答）的核心是从**多条预设问题**中选择一条输出（支持随机/指定选择）。
  如果需要“从多条预设内容中选一条展示”，问答节点更合适。

4.12.5 异常

异常节点功能，当工作流运行遇到业务异常时，通过本节点中断当前分支逻辑并抛出异常信息，最终由结束节点完成流程的终止，便于定位问题和后续处理。

版权所有 © 华为云计算技术有限公司

核心价值与适用场景：

使用异常节点，是为了在遇**到不可恢复的错误**时，快速失败并明确告知原因，避免产生脏数据。

表 4-69 异常节点使用场景

场景分类	示例	处理建议
前置校验失败	参数缺失、权限不足(403)、签名错误。	必须使用异常节点，阻断后续逻辑。
外部依赖故障	插件API超时、数据库连接失败。	使用异常节点，并返回系统级错误码。
业务逻辑拒绝	余额不足、库存为 0。	使用异常节点返回业务错误码，配合消息节点提供用户友好的提示信息。

配置异常节点

配置异常节点的详细参数说明如下：

注意

异常节点会中断当前分支逻辑并抛出异常信息，但仍需连接**结束节点**来完成流程的结束。未连接**结束节点**将导致流程无法正常终止。

异常信息载体

异常节点要求输出内容必须符合**JSON格式**。这确保了前端或上游系统能够通过程序解析错误原因，而不仅仅是弹出一个文本框。推荐的JSON结构如下：

```
{
  "code": "ERROR_CODE",     // 错误标识，供程序判断
  "message": "Human readable error", // 错误描述，供展示
  "details": { ... }        // (可选) 调试信息
}
```

说明

- code字段：必填，建议使用大写字母+下划线命名，如POINTS_INSUFFICIENT。
- message字段：必填，应提供用户可读的错误描述。
- details字段：可选，用于提供调试信息。

配置方式

- **直接输入**：适用于一次性的、简单的错误定义。单击右侧的 可以扩展内容输入框。示例如下：

版权所有 © 华为云计算技术有限公司

图 4-130 配置异常节点



- **插入消息模板：**适用于企业级标准化的错误管理。当错误文案需要变更时（例如修改错别字），只需修改模板，所有引用该模板的节点会自动更新。
  - 前提：需预先创建好错误消息模板，具体请参见**消息模板**。
  - 操作：单击图标，选择已创建的异常消息模板，并单击“确定”。

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以积分兑换场景为例，业务用户尝试兑换一个价值500积分的商品。系统需要检查用户当前积分，如果积分不足，中断流程并返回拒绝信息；如果积分充足，则输出“请兑换商品”。

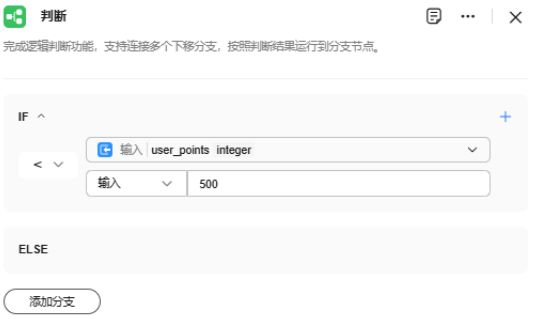
图 4-131 工作流

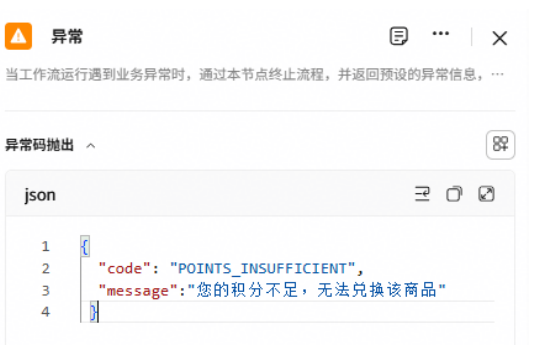


工作流核心节点如**表4-70**所示。

版权所有 © 华为云计算技术有限公司

表 4-70 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	
输入节点	此节点需要让用户输入当前的积分。 **输入参数**： • 参数名称：user_points • 参数类型：Integer • 描述：请输入您的积分。	
判断节点	此节点用来判断用户输入的积分是否小于500，小于500则执行IF分支，大于等于500则执行ELSE分支。	

异常节点	当判断用户积分小于500时，此节点抛出异常信息并终止当前分支逻辑。
	异常信息填写如下：
	{
	"code":
	"POINTS_INSUFFICIENT",
	"message":"您的积分不足，无法兑换该商品"
	}
		

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
消息节点	当判断用户积分大于等于500时，此节点输出：请兑换商品。 指定回复：请兑换商品	
结束节点	此节点用于结束流程。	

工作流运行的效果如图4-132所示。

版权所有 © 华为云计算技术有限公司

图 4-132 效果示例



常见问题

- **异常节点和结束节点有什么区别?**
  - 结束节点：代表流程成功完成。
  - 异常节点：代表流程失败。在API调用模式下，通常会触发非200的状态响应或包含错误码的Payload。异常节点后需连接**结束节点**来完成流程结束
- **异常节点后还能连接其他节点吗?**

  需要。异常节点会中断当前分支逻辑并抛出异常信息，但仍需连接**结束节点**来完成流程的结束。未连接**结束节点**将导致流程无法正常终止。
- **如何在调试日志中快速定位异常?**

  建议在JSON中包含Trace ID或节点名称。例如："source": "check_balance_node"。这样在查看运行日志时，能一眼定位是哪个环节抛出的异常。
- **应该把所有错误都抛给用户吗?**

  不建议。对于系统级错误（如数据库连接失败），建议向用户返回友好的模糊提示（如“系统繁忙”），而将详细的堆栈信息（Stack Trace）记录在后台日志中，避免暴露系统安全细节。

4.13 变量&知识节点

版权所有 © 华为云计算技术有限公司

4.13.1 变量赋值

变量赋值节点是工作流中用于**修改和存储变量值**的节点。通过变量赋值节点，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。

变量赋值节点支持在**循环节点**内部和外部使用，两种场景下可操作的目标变量不同。变量赋值节点应用广泛，例如：

- **存储中间结果：**在工作流中，将中间计算或处理的结果通过变量赋值节点存储到变量中，以便后续节点使用。例如，在开发一个智能内容审核工作流时，需要对用户提交的文章进行多维度分析，如提取关键词、判断情感倾向、检测敏感内容等。此时，可以设置一个记忆变量review_status，在审核节点执行完毕后，通过变量赋值节点将审核结果（如 "passed" 或 "rejected"）存储到该变量中。后续在生成审核报告或决定是否发布时，可以直接从review_status变量中获取审核状态，无需重复执行审核逻辑。
- **在循环中累积数据：**在包含循环节点的工作流中，每轮循环产生的结果需要逐步累积，而不是被下一轮覆盖。通过变量赋值节点，可以在每轮循环结束时将本轮结果追加到中间变量中。例如，在开发一个批量数据处理工作流时，需要遍历一组客户记录，为每位客户生成个性化的营销文案。此时，可以在循环节点中设置一个中间变量all_copies，每轮循环中大模型节点生成一条文案后，通过代码节点将新文案与all_copies中已有的内容拼接，再通过变量赋值节点将拼接结果写回all_copies。循环结束后，all_copies中包含了所有客户的营销文案汇总。
- **记忆用户状态：**用户在会话中表达的偏好或提供的信息，需要在后续对话中生效。通过变量赋值节点将用户状态写入记忆变量，即可实现该轮会话的状态保持。例如，在开发一个智能客服工作流时，用户在对话中说"请用英文回复我"。此时，工作流通过大模型节点识别出语言偏好后，使用变量赋值节点将reply_language记忆变量的值从 "中文" 更新为 "English"。在该轮后续所有对话中，即使用户没有再次声明语言偏好，大模型节点通过读取reply_language变量，仍然会使用英文回复。
- **重置与清除状态：**变量赋值节点支持对8种数据类型赋空值，适合用于在任务完成后清除临时数据、恢复初始状态。例如，在开发一个多次任务工作流时，用户完成一次任务后说“我要再来一次”。此时，通过变量赋值节点将collected_data赋空值（清除上一次的采集数据）、将current_stage赋固定值“start”（回到起始阶段），同时对task_count执行自增+1（记录累计任务次数），从而在不终止会话的情况下实现任务的重新开始。

核心概念

在配置赋值节点前，先理解以下概念：

- **赋值目标：**变量赋值节点能够修改的变量类型取决于它所在的位置。

版权所有 © 华为云计算技术有限公司

表 4-71 赋值目标

使用位置	可赋值的目标变量	说明
循环节点外	**对话型工作流中的记忆变量**（在全局配置中定义）	记忆变量可以跨节点、跨轮次持久保存，只能通过变量赋值节点修改。 典型用途为存储用户偏好、对话状态标记、跨轮次计数器等。
循环节点内	**循环中间变量**（在循环节点中定义）	中间变量仅在当前循环的生命周期内有效，能通过循环体内的**变量赋值节点**修改。 典型用途为循环中的累加器、临时结果拼接、状态标记等。

- **数据来源：**值有多种来源方式。

表 4-72 数据来源

来源方式	说明	适用场景
引用	引用上游节点的输出变量、记忆变量或循环内置变量（index、item）	动态赋值：将计算结果写入变量
输入（固定值）	直接填写一个常量	重置变量、设置初始状态
运算赋值	对数值类型（Integer、Number）变量进行自增或自减（步长为1）。	循环计数器、余额扣减
	将变量重置为空 **说明** 支持8种数据类型赋空值，包括：String、Integer、Number、Boolean、Object、Array<Number>、Array<String>、Array<Integer>。	清除状态、重新开始

配置变量赋值节点

配置变量赋值节点的详细参数说明如下：

循环节点外变量赋值节点配置

变量赋值节点变量名称仅支持选择“**全局配置**”中记忆变量，如**图4-133**所示。

版权所有 © 华为云计算技术有限公司

图 4-133 配置循环外变量赋值节点



类型、值：支持“引用”、“输入”和“运算赋值”三种类型。

- 引用：支持引用三类参数源，工作流前置节点的输出参数、全局配置中的记忆变量（需预先设置），以及系统参数。

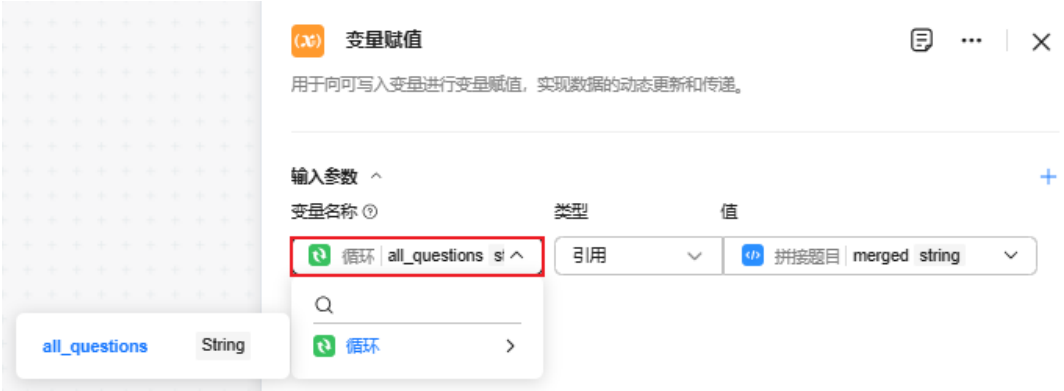
  系统参数列举如下：

  - “conversation_history”：存储当前对话的历史消息记录。
  - “current_time”：获取当前的系统时间戳或时间字符串。
  - “user_id”：识别当前用户的唯一标识符。
  - “conversation_id”：当前工作流的唯一对话标识。
  - “dialogue_count”：当前对话的轮次数计数。
- 输入：支持用户自定义取值。
- 运算赋值：
  - 支持数值类型进行简单运算，包括变量自增和变量自减。
  - 支持8种数据类型赋空值，包括：String、Integer、Number、Boolean、Object、Array<Number>、Array<String>、Array<Integer>。

循环节点中变量赋值配置

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，如图4-134所示。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。

图 4-134 配置循环内变量赋值节点



版权所有 © 华为云计算技术有限公司

类型、值：支持“引用”、“输入”和“运算赋值”三种类型。

- 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。

  系统参数列举如下：

  - “conversation_history”：存储当前对话的历史消息记录。
  - “current_time”：获取当前的系统时间戳或时间字符串。
  - “user_id”：识别当前用户的唯一标识符。
  - “conversation_id”：当前工作流的唯一对话标识。
  - “dialogue_count”：当前对话的轮次数计数。
- 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。
- 运算赋值：
  - 支持数值类型（Integer、Number）进行简单运算，包括变量自增和变量自减。
  - 支持8种数据类型赋空值，包括：String、Integer、Number、Boolean、Object、Array<Number>、Array<String>、Array<Integer>。

相关操作

- 单击节点名称，可修改节点名称。
- 单击⋯图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击▷图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

- **示例1：记忆用户偏好（循环外赋值）**

  在开发一个智能客服工作流时，用户在对话中说“请用英文回复”。此时，工作流通过大模型节点识别出语言偏好后，使用变量赋值节点将user_language记忆变量的值从 "中文" 更新为 "en"。在该轮会话中后续所有对话，即使用户没有再次声明语言偏好，大模型节点通过读取user_language变量，仍然会使用英文回复。

  工作流的全景图如图4-135所示。

  图 4-135 工作流

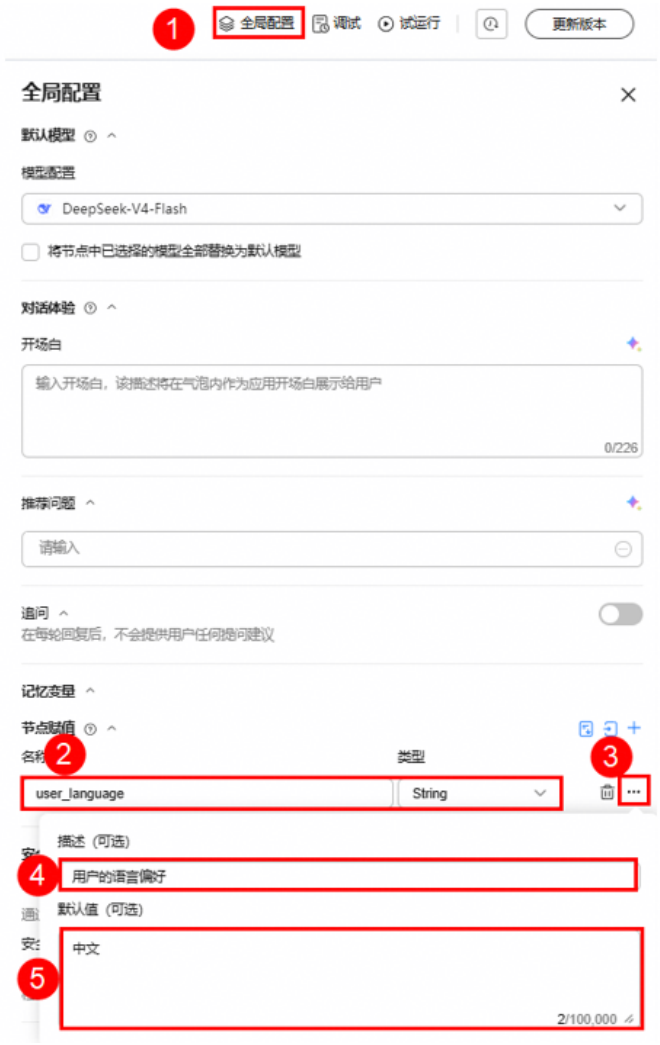
  在配置工作流节点之前，需要先配置“全局配置”中的记忆变量，相关参数配置如下，具体如图4-136所示。

  记忆变量：

版权所有 © 华为云计算技术有限公司

名称：user_language，类型：String，描述：用户的语言偏好，默认值：中文。

图 4-136 配置记忆变量



工作流核心节点如表4-73所示。

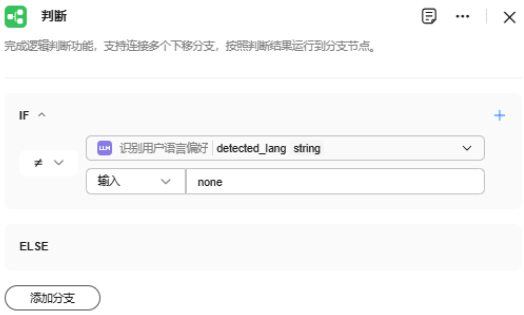
表 4-73 节点配置说明

节点类型	说明	示例
开始节点	使用系统默认配置query参数，表示用户输入的内容。	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点-识别用户语言偏好	此节点根据用户输入的内容检测是否表达用户的语言偏好，如果检测到语言偏好，则返回具体语言名称，如果未检测到语言偏好，则返回none。 • **模型配置**：选择DeepSeek-V4-Flash模型。 • **输入参数**：引用开始节点的query。 • **输出参数**：将用户表达的语言偏好，在输出参数中定义，并在描述中说明输出参数的字段含义，便于大模型理解。 单击输出区域右上角的“ ”，选择输出格式为“JSON”。 定义内容如下： detected_lang：类型为“String”，描述为用户的语言偏好。 • **提示词配置**： – 系统提示词：定义一段系统提示词，如下所示： ## 人设 - 语言偏好检测专家 - 精通多语种关键词匹配与正则表达式 ## 任务描述 - 识别用户输入中是否显式声明语言偏好 - 若检测到，返回对应语言名称；否则返回“none” - 帮助下游模块快速决定响应语言，提升用户体验 ## 约束条件 1. 仅依据用户本次输入判断，不依赖历史	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
	上下文 2. 语言名称须使用ISO 639-1 小写两位字母代码（如 en、zh、es） 3. 若出现多种语言声明，以最先出现的为准 4. 按照<输出格式>输出 5. 按照<执行步骤>一步一步执行 ## 执行步骤 1. 预处理：去除标点、统一小写 2. 构建关键词表：{"english":"en", "chinese":"zh", "spanish":"es", "french":"fr", "german":"de", "japanese":"ja", "korean":"ko"} 3. 扫描输入，匹配关键词表 4. 命中则输出对应值，未命中输出“none” ## 输出格式：JSON格式。 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	
判断节点	此节点用于根据大模型节点输出参数detected_lang的值判断是否需要给“记忆变量”赋值。 • **IF**分支：大模型节点的detected_lang的值不等于none。 • **ELSE**分支：大模型节点的detected_lang的值等于none。	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
变量赋值节点	当大模型节点检测到用户的语言偏好后，将记忆变量的值更新为对应语言名称。 参数名称：记忆变量中的user_language。 类型、值：引用“识别用户语言偏好”大模型节点的detected_lang。	
大模型节点-回答用户问题	此节点将根据用户偏好语言来回答用户的问题。 • **模型配置**：选择DeepSeek-V4-Flash模型。 • **输入参数：** – query：引用开始节点的query。 – user_language：引用记忆变量中的user_language。 • **提示词配置：** – 系统提示词：定义一段系统提示词，如下所示： 用{{user_language}}回复用户的问题{{query}} – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

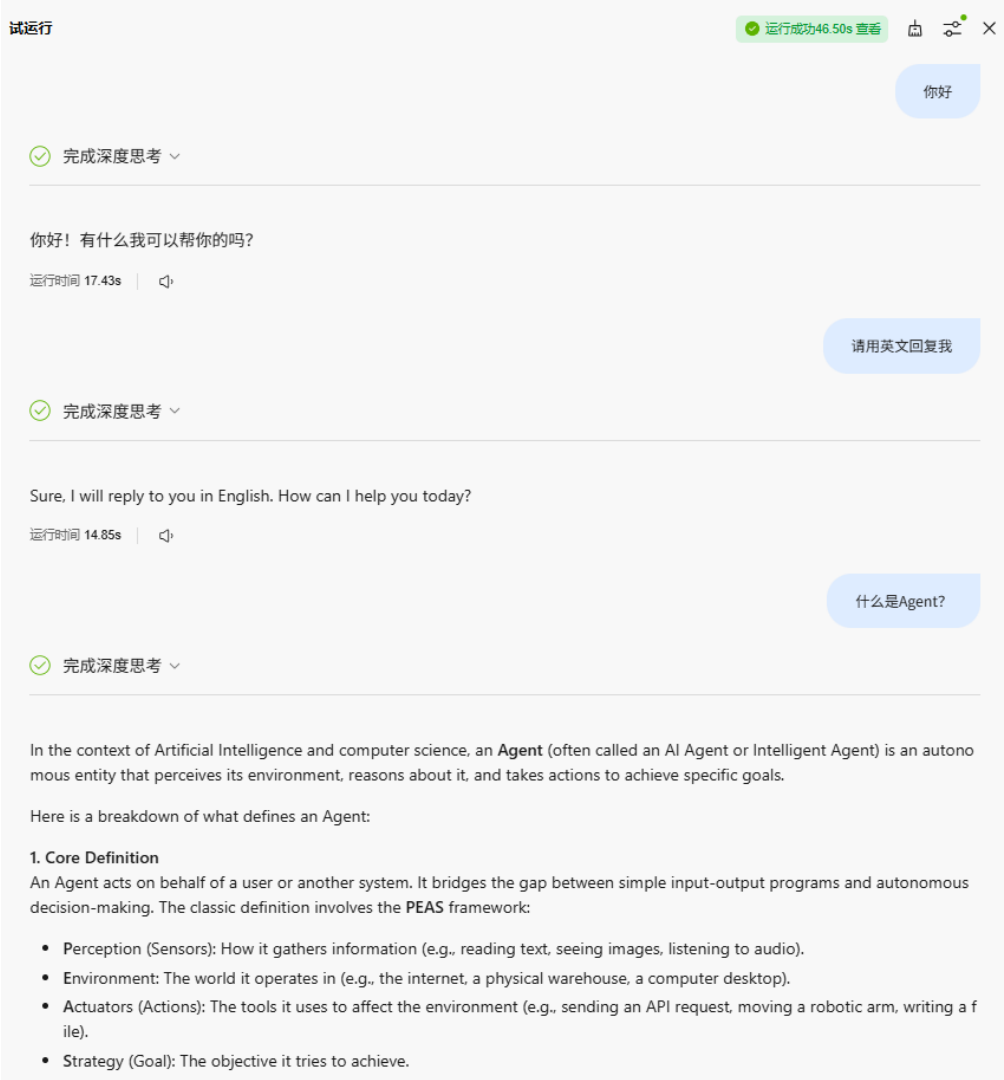
 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
结束节点	此节点用于输出大模型节点的回答内容。 **输入参数：** result：引用“回答用户问题”大模型节点的raw_output。	

在试运行配置界面使用默认值“中文”，工作流运行的效果如图4-137所示，当用户说“请用英文回复”，工作流在后续所有轮次中，即使用户没有再次声明语言偏好，也仍然会使用英文回复。

 版权所有 © 华为云计算技术有限公司

图 4-137 效果示例



- **示例2：自动循环生成岗位面试题（循环内赋值）**

  用户输入一个岗位名称，系统自动循环3次，每次生成1道面试题，将所有面试题逐步拼接为一份完整的面试题集。

图 4-138 工作流



工作流核心节点如表4-74所示。

 版权所有 © 华为云计算技术有限公司

表 4-74 节点配置说明

节点类型	说明	示例
开始节点	在开始节点说明中补充默认参数：使用系统默认配置query参数，表示用户输入的内容。并添加一个必选的输入参数，即需要出面试题的岗位名称。 ● 参数名称：job_title ● 参数类型：String ● 描述：出面试题的岗位名称 ● 必填：勾选。	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
循环节点	用于指定生成的面试题的题数，每次循环通过**中间变量**将每一轮出的题作为变量传递到下次循环中，再通过代码节点将每一轮的面试题拼接为一份完整的面试题集。其中包含3个节点： • 大模型节点：根据用户输入的岗位名称，每轮循环对应生成一道面试题。 • 代码节点：将每一轮循环的面试题进行拼接。 • 变量赋值节点：用于存储代码节点拼接好的面试题内容，并将其赋给循环节点的中间变量。 循环节点设置如下： • **循环类型：**选择“指定循环次数”。 • **循环次数：**本示例设置为 3。 • **中间变量：**设置变量all_questions，这个变量用于传递上次已生成的面试题内容，会在下次循环中提供给代码节点进行拼接。 因为首次循环时没有面试题，可以设置为空字符串（""），也可根据需要设置前置引导语，此处设置为“面试题”作为标题。 • **输出参数：**设置变量output，引用循环节点的中间变量all_questions。	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点-出题	此节点根据用户输入的岗位名称，对应生成一道面试题。 • **模型配置**：选择DeepSeek-V4-Flash模型。 单击 ，将温度设置为0.8，保证每轮生成不同的题目。 • **输入参数：** – query：引用开始节点的query。 – job_title：引用开始节点的job_title。 – question_number：引用循环节点的index。 – previous_questions：引用循环节点的中间变量all_questions。 • **输出参数：** 单击输出区域右上角的“ ”，选择输出格式为“JSON”。 定义内容如下： question：类型为“String”，描述为岗位的面试题。 • **提示词配置：** – 系统提示词：定义一段系统提示词，如下所示： 请为"{{job_title}}"岗位出1道面试题。 已出的题目如下（不要重复）： {{previous_questions}} 要求： - 只输出1道题目的文本，不要编号，不要输出其他内容	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
	- 与已出题目不重复 - 难度适中 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例

代码节点-拼接题目	此节点将每一轮循环的面试题进行拼接。
	• **绑定函数：**
	1. 单击“新建函数”，在弹出的对话框输入函数名称：merged，函数描述为“文本拼接”，输入完成后单击“确定”。
	2. 单击“编辑源码”，编辑如下代码：
	def main(args: dict) -> dict:
	"""
	运行代码节点会调用此函数，请勿对下面的函数名做修改
	:param args: 输入固定为args字典类型，kv为输入参数键值对，包含：
	- index: 题目的序号（数字，从0开始）
	- new_q: 新增题目的具体内容
	- prev: 之前已汇总的题目文本（可为空）
	:return: 输出参数为字典类型，kv为输出参数键值对，包含：
	- merged: 拼接后的完整题目汇总文本
	"""
	# 从输入字典中提取参数，若参数不存在则赋默认值（增强代码健壮性）
	index = args.get("index", 0)
	new_q = args.get("new_q", "")
	prev = args.get("prev", "")
	
	# 按格式拼接当前题目行（序号+1，转为1起始的展示序号）
	line = f"第{index + 1}题：{new_q}"
	
	# 拼接历史文本与当前题目行（空值处理）
	if prev:
	result = prev +
		

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例

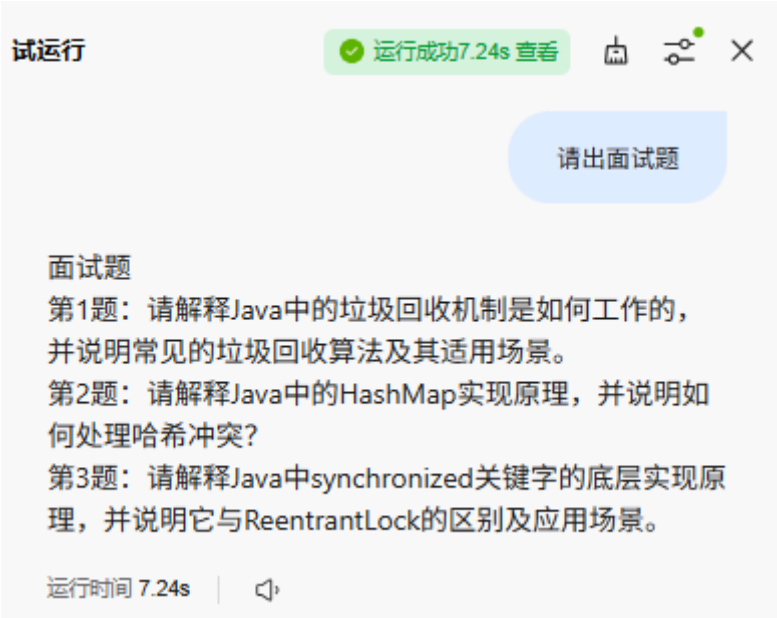
	<pre>"\n" + line
	  else:
	  result = line
	
	  # 构造返回字典
	  ret = {
	  "merged": result
	  }
	  return ret</pre>
	• **输入参数：**
	– index：引用循环节点的index。
	– new_q：引用“出题”大模型节点的question。
	– prev：引用循环节点的中间变量all_questions。
	• **输出参数：**
	参数名称：merged
	参数类型：String
	描述：拼接后的面试题	

变量赋值节点	用于存储代码节点拼接好的面试题内容，将“拼接题目”代码节点的merged的值赋给循环节点的中间变量all_questions。	
结束节点	此节点用于输出面试题目。 **输入参数：** result：引用循环节点的output。	

 版权所有 © 华为云计算技术有限公司

在试运行配置界面输入“Java后端工程师”，工作流运行的效果如图4-139所示。

图 4-139 效果示例



常见问题

表 4-75 常见问题

问题	原因	解决方法
赋值后下游节点读到的仍然是旧值	赋值节点的位置在下游节点**之后**，赋值还未执行	确保赋值节点在需要读取新值的节点**之前**
赋空值后变量类型变了	赋空值时选择了错误的类型	赋空值时确保选择的类型与变量定义的类型一致
循环内赋值后，循环外读不到中间变量	中间变量的作用范围仅限于循环体内部，循环结束后不保留	如需在循环外使用循环的结果，请通过循环节点的**输出参数**传递

4.13.2 变量聚合

变量聚合节点将多个分支的输出变量**汇聚为一个统一变量**，供下游节点引用。它的核心能力是：从多个分支的输出中，按照用户指定的优先级顺序，**返回第一个非空值**。

在工作流中使用条件分支时，同一时刻只有一个分支会执行并产生输出，其他分支的输出为空。下游节点如果直接引用分支输出，需要处理空值判断。变量聚合节点将这些分支的输出汇聚为统一变量，自动返回第一个非空值，简化下游处理逻辑。

变量聚合节点通常与**判断节点**配合使用，典型场景包括：

- **多语言翻译**：不同语言翻译分支的结果汇聚为一个输出。

版权所有 © 华为云计算技术有限公司

- **多渠道数据采集**：不同数据源的结果汇聚为统一格式。
- **多策略结果择优**：多个策略分支的结果按优先级取值。

配置变量聚合节点

配置变量聚合节点时，需要选择聚合策略和聚合变量，并按需设置聚合分组。

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。

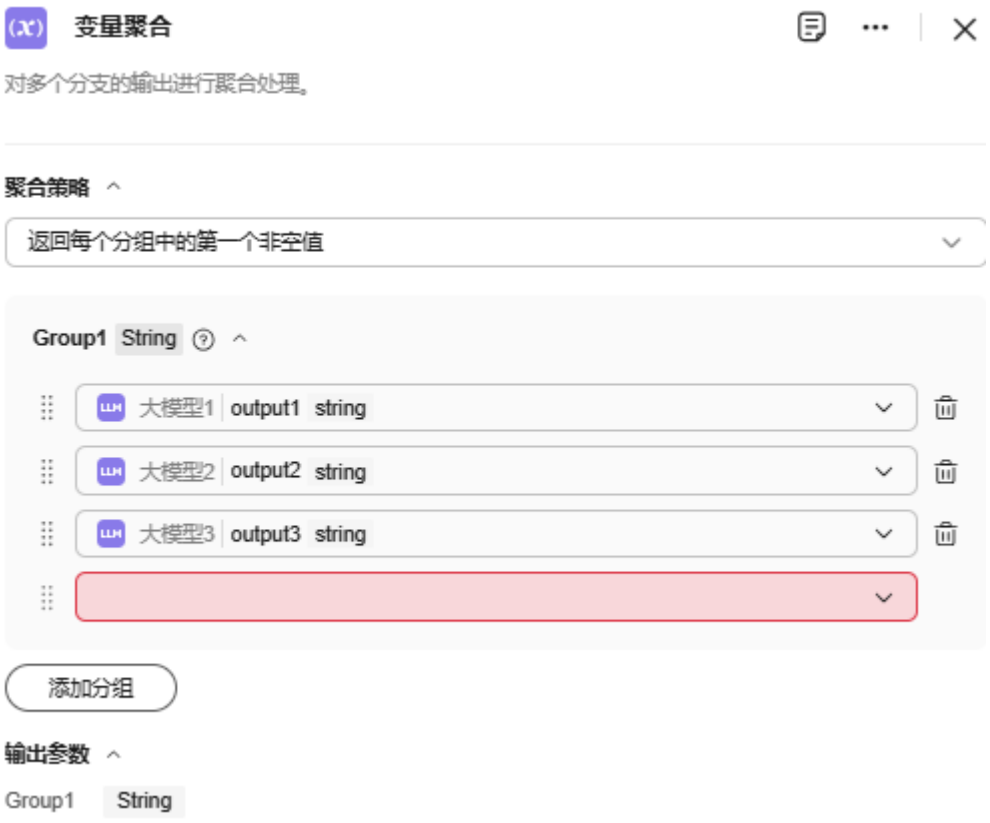
说明

目前聚合策略仅支持“返回每个分组中第一个非空值”。如果所有分支的输出变量均为空，聚合结果也将为空。建议设置兜底分支确保至少有一个非空输出。

分组内变量的排列顺序决定了非空值的检查优先级，支持拖动变量调整位置，排在最上方的变量优先级最高。

例如，如图4-140所示，Group1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output1不为空，则用output1的值为Group1赋值；如果output1为空，则取output2的值，依次类推。

图 4-140 配置示例



聚合变量

在聚合分组中选择需要聚合的变量，**同一分组内的所有变量必须是相同的数据类型。**

版权所有 © 华为云计算技术有限公司

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，如图4-141所示。

图 4-141 配置示例



聚合分组

默认一个分组Group1。如需聚合多个不同用途/类型的变量，单击添加按钮创建新分组。

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。

 版权所有 © 华为云计算技术有限公司

输出参数

输出参数与分组一一对应。默认分组Group1的输出参数名为Group1，其值为该分组按聚合策略计算后的结果。添加多个分组时，输出参数依次递增为Group2、Group3等。

输出参数的数据类型与对应分组内变量类型一致。

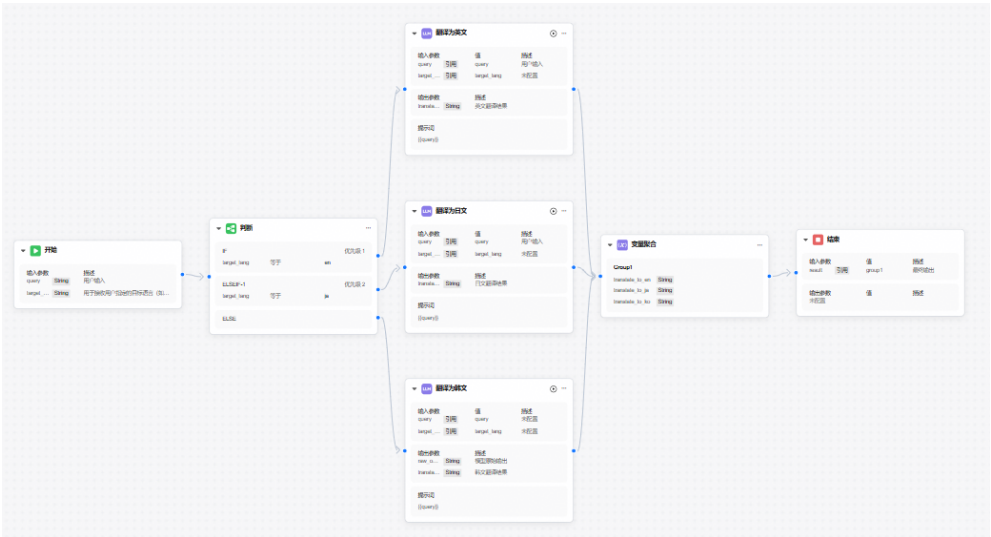
相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

配置示例

以多语言翻译助手为例，开始节点接收用户需要翻译的文本以及用户指定的目标语言（英文，日文，韩文），判断节点根据目标语言判断翻译方向，并流转至不同语言的翻译分支，例如日文翻译流转给“翻译为日文”分支的大模型来翻译。三个分支均汇总到变量聚合节点，汇总为1个参数，即需要翻译的文本Group1，并传递到结束节点展示给用户。

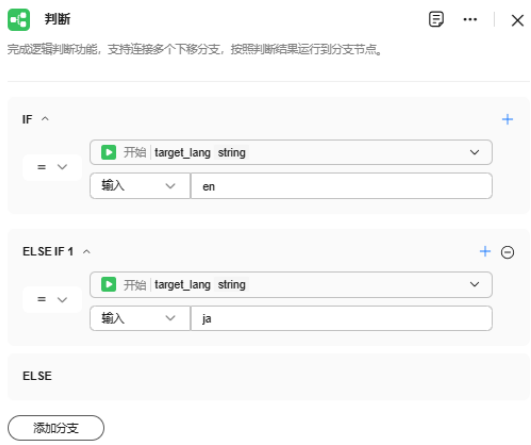
图 4-142 工作流



工作流核心节点如表4-53所示。

版权所有 © 华为云计算技术有限公司

表 4-76 节点配置说明

节点类型	说明	示例
开始节点	添加一个必选的输入参数，即用户需要翻译成的目标语言。 ● 参数名称：target_lang ● 参数类型：String ● 描述：用于接收用户指定的目标语言（如需翻译为英文，请输入en；如需翻译为日文，请输入ja，如需翻译为韩文，请输入ko） ● 必填：勾选。	
判断节点	此节点用于根据用户在开始节点输入的target_lang的值判断翻译方向。 由于有英文、日文和韩文三个分支方向，因此单击“添加分支”，将判断节点确定为：**IF**、**ELSE IF 1**和**ELSE**分支。 ● **IF**分支：开始节点的target_lang的值=en ● **ELSE IF 1**分支：开始节点的target_lang的值=ja ● **ELSE**分支：默认为开始节点的target_lang的值为ko	

 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点-翻译为英文	此节点将用户输入的文本翻译为英文。 • **模型配置**：选择DeepSeek-V4-Flash模型。 • **输入参数：** – query：引用开始节点的query。 – target_lang：引用开始节点的target_lang。 • **输出参数：**将大模型翻译后的内容，在输出参数中定义，并在描述中说明该参数的字段含义，便于大模型理解。 定义内容如下： translate_to_en：类型为“String”，描述为英文翻译结果。 • **提示词配置：** – 系统提示词：定义一段系统提示词，如下所示： 你是一个专业的翻译助手，请将以下内容准确地翻译成{{target_lang}}。 文本为：{{query}} 仅输出翻译结果，不要包含任何其他解释。 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

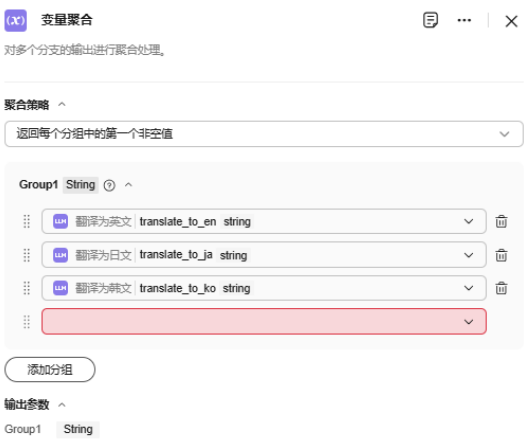
 版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
大模型节点-翻译为日文	此节点将用户输入的文本翻译为日文。 • **模型配置**：选择DeepSeek-V4-Flash模型。 • **输入参数：** – query：引用开始节点的query。 – target_lang：引用开始节点的target_lang。 • **输出参数：**将大模型翻译后的内容，在输出参数中定义，并在描述中说明该参数的字段含义，便于大模型理解。 定义内容如下： translate_to_ja：类型为“String”，描述为日文翻译结果。 • **提示词配置：** – 系统提示词：定义一段系统提示词，如下所示： 你是一个专业的翻译助手，请将以下内容准确地翻译成{{target_lang}}。 文本为：{{query}} 仅输出翻译结果，不要包含任何其他解释。 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

版权所有 © 华为云计算技术有限公司

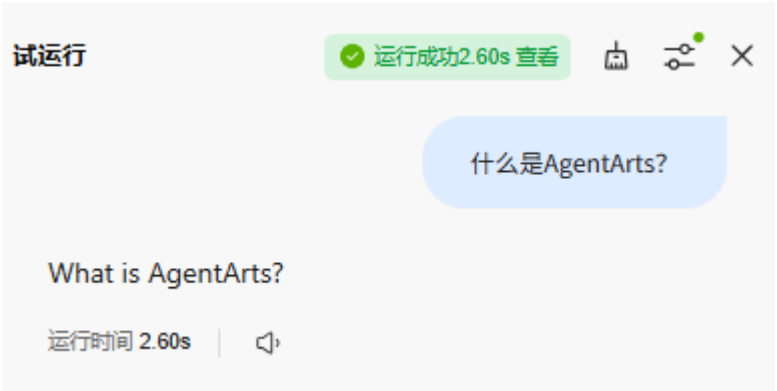
节点类型	说明	示例
大模型节点-翻译为韩文	此节点将用户输入的文本翻译为韩文。 • **模型配置**：选择DeepSeek-V4-Flash模型。 • **输入参数：** – query：引用开始节点的query。 – target_lang：引用开始节点的target_lang。 • **输出参数：**将大模型翻译后的内容，在输出参数中定义，并在描述中说明该参数的字段含义，便于大模型理解。 定义内容如下： translate_to_ko：类型为“String”，描述为韩文翻译结果。 • **提示词配置：** – 系统提示词：定义一段系统提示词，如下所示： 你是一个专业的翻译助手，请将以下内容准确地翻译成{{target_lang}}。 文本为：{{query}} 仅输出翻译结果，不要包含任何其他解释。 – 用户提示词：引用当前的输入参数query，即引用用户的原始输入。 {{query}}	

版权所有 © 华为云计算技术有限公司

节点类型	说明	示例
变量聚合节点	此节点将三个翻译分支的输出参数汇总为统一的一个参数，即需要翻译的文本Group1，供下游节点引用。	
结束节点	此节点用于输出翻译后的内容。 **输入参数：** result：引用变量聚合节点的group1。	

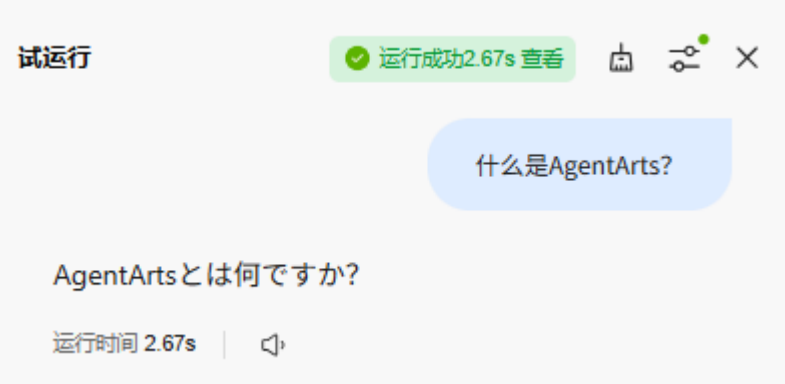
在试运行配置界面输入翻译的目标语言，例如“en”，工作流运行的效果如图4-143所示；在试运行配置界面输入翻译的目标语言为“ja”，工作流运行效果如图4-144所示。

图 4-143 效果示例（翻译为英文）



 版权所有 © 华为云计算技术有限公司

图 4-144 效果示例（翻译为日文）



常见问题

- **为什么聚合结果为空?**

  所有分支的输出变量均为空，聚合节点无法找到非空值。解决方法如下：

  a. 检查各分支是否正确配置了输出变量。

  b. 检查条件分支逻辑，确保至少有一条分支会执行。

  c. 考虑添加一个“兜底分支”，确保总有一个非空输出。

- **配置变量聚合节点时，为什么无法将某个变量添加到分组?**

  该变量的数据类型与分组内已有变量的类型不一致，请确保同一分组内所有变量类型必须一致。如需聚合不同类型，请创建新的分组（Group2）。

- **聚合结果取到了错误分支的值是什么原因?**

  变量在分组内的排列顺序（优先级）不正确。建议将期望优先返回的分支变量移到最上方。

- **新增分支后，为什么聚合节点未输出新分支的值?**

  新增分支的输出变量未被添加到聚合节点的分组中，打开聚合节点配置面板，在对应分组中添加新分支的输出变量。

4.13.3 知识检索

知识检索节点是工作流的“外部知识存储器”。它通过RAG（检索增强生成）技术，从预先上传的知识库中根据用户意图匹配相关片段。

核心作用：解决大模型无法掌握的私有领域数据，实时信息以及易产生幻觉的问题。

RAG工作流数据运行机制如下：

用户原始问题 → 知识检索节点 → 进行知识检索 → 匹配到的切片 → 大模型总结 → 生成回答

添加节点

在工作流编辑页面，单击下方“添加节点”，选择“知识检索”节点。

版权所有 © 华为云计算技术有限公司

图 4-145 添加知识检索节点



配置知识检索节点

配置知识检索节点的详细参数说明如下：

表 4-77 配置项总览

配置参数	说明
输入参数	固定只有1个且名称为query（不可修改），类型为String，表示待检索的问题。通常引用开始节点作为搜索词。
知识库	选择关联的知识库（支持关联多个）。可通过“仅查询指定标签的文档”功能结合标签过滤，缩小查询范围。
指定标签检索	支持通过文档标签限制检索范围。提供“基础检索”（多标签取并集）与“高级检索”（支持布尔逻辑“与、或、非”，最多三级嵌套）。
检索策略	选择检索方式：语义检索（理解意图，向量）、关键词检索（精准倒排匹配）、混合检索（兼顾两者，推荐）、FAQ检索（匹配问答对）。
FAQ直出阈值	开启后优先检索FAQ。如果得分超过预设阈值（0~1之间），将跳过常规文档检索并直接返回给大模型总结，以提升检索效率。
相关度阈值	设置0~1之间的阈值，得分低于此阈值的切片将被舍弃，防止噪音干扰。建议参考命中测试进行微调。
topk召回数量	单次检索最大返回切片数（取值范围1~50之间）。
查看图片	开启后，当知识库配置了“图片解析 > 仅保留原图”时，支持查看并提取检索结果中的图片切片。
输出参数	固定变量为output_list（对象数组）。包含文件名、子标题、分片内容和匹配分数等属性。

版权所有 © 华为云计算技术有限公司

说明

成本与计费说明：

- 知识检索节点召回的所有分片，会作为参考知识拼接输入下游大模型节点中。
- 如果您的“topk召回数量”设置过大（例如 20~50），或者单分片长度过长，将会导致下游大模型节点在处理提示词时的输入Token消耗量增加，从而带来Token账单增加。
- 建议根据知识库的分片切片长度，将Top-K 召回数量合理限制在 3~5 个，以保证大模型运行在最佳的成本区间内。

输入参数

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。通常引用开始节点的值，该变量决定了检索的搜索词。

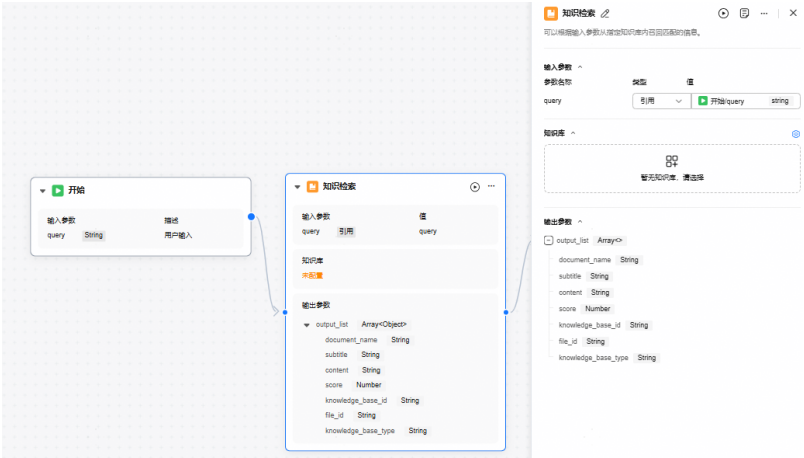
设置参数值时，支持“引用”和“输入”两种方式。

- 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中的记忆变量，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。

  系统参数列举如下：

  - “conversation_history”：存储当前对话的历史消息记录。
  - “current_time”：获取当前的系统时间戳或时间字符串。
  - “user_id”：识别当前用户的唯一标识符。
  - “conversation_id”：当前工作流的唯一对话标识。
  - “dialogue_count”：当前对话的轮次数计数。
- 输入：支持用户输入固定值，适用于知识检索问题固定的场景。

图 4-146 配置输入参数



知识库

您可以选择一个或者多个知识库，知识库支持设置按标签检索、检索策略、相关度阈值、topk召回数量等信息。

说明

知识库使用相关限制说明请参考**知识库使用限制**。

版权所有 © 华为云计算技术有限公司

指定标签检索

注意

- 指定标签检索功能目前不支持第三方接入的知识库。
- 布尔条件配置中添加条件组最多支持三级嵌套。

创建知识库过程中，对于“知识文档”类型的文件，可以配置文档标签。工作流中的“知识检索”节点，通过标签控制检索范围，用来提高检索的准度，避免不相关文档的干扰。

图 4-147 在知识库中添加文档标签



图 4-148 在工作流节点中启用标签配置



配置检索标签时，可以指定只检索某类标签的文档，支持“基础检索”、“高级检索”两种标签设置模式。

- 基础检索：基于并集逻辑的快速检索方式，可以选择一个或者多个标签的文档进行检索。
- 高级检索：可以使用布尔逻辑（如“或”、“且”、“NOT（≠）”）来更精细地配置检索条件。例如检索“教育”或“科普”标签的文档。

版权所有 © 华为云计算技术有限公司

图 4-149 配置指定标签检索



检索策略

文档检索的方式，有如下几种检索策略：

- **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、并对用户意图进行理解场景中使用。
- **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。

版权所有 © 华为云计算技术有限公司

图 4-150 配置检索策略



FAQ 直出阈值

启用FAQ功能后，系统将优先检索FAQ数据。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索；如果未超过阈值或未命中结果，则将继续进行文档检索（可能会带来一定的性能开销）。

图 4-151 知识库中上传 FAQ 类型文档



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。

版权所有 © 华为云计算技术有限公司

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。

查看图片

开启此功能后，当知识库配置了“图片解析 > 仅保留原图”功能时，可查看检索结果中的图片信息（新建知识库时、停用知识库后可以修改图片解析配置）。

图 4-152 设置知识库仅保留原图



输出参数

知识检索节点的输出是一个名为output_list的对象数组，表示所有满足检索要求的知识切片。数组中对象有七个属性：

- document_name，知识切片所在的知识文档名称。
- subtitle，知识切片子标题。
- 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。
- knowledge_base_id，文件所属知识库ID。
- knowledge_base_type，文件所属知识库类型。
- file_id，文件ID。

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、切片子标题、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。

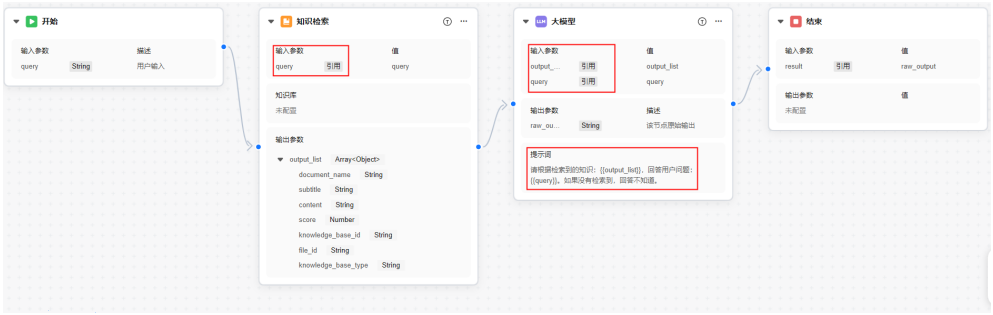
配置示例

本示例建立一个带有知识检索功能的工作流，详细配置介绍请参考**搭建知识库问答工作流**。

- **节点连接：**将“开始节点”连接至“知识检索节点”，再连接“大模型节点”和“结束节点”。

版权所有 © 华为云计算技术有限公司

图 4-153 知识库问答工作流



- 变量引用：
  - 所有节点的输入参数均使用“引用”关系，引用前置节点的内容作为本节点的输入。
  - 知识检索节点引用输入节点的query参数，将用户输入的原始问题作为搜索词。
  - 大模型节点的query参数引用开始节点，用于将用户原始问题传递给模型；新增output_list参数，引用知识检索节点的输出；“用户提示词”设置为：请根据检索到的知识：{{output_list}}，回答用户问题：{{query}}。如果没有检索到，回答不知道。
  - 结束节点引用大模型节点的输出。
- 调试与结果验证

  a. 在工作流编辑页面右上角，单击“试运行”。

  b. 在输入参数中，输入已上传至知识库中的相关业务问题。

  c. 预期验证标准：在右侧测试面板的“调试”中，展开知识检索节点的“输出”，检查output_list数组中是否已精准召回对应分片内容；同时验证大模型节点是否将召回的切片内容顺利整合成为了最终回答。

相关操作

- 单击节点名称，可修改节点名称。
- 单击…图标，可对节点进行重命名、复制、删除或更换节点；也可以将节点连接到结束节点。
- 单击图标，可对节点进行测试。
- 单击名称下面的节点说明，可修改节点的描述信息。
- 单击图标，可查看文档信息。

常见问题

- **为什么知识库中明明有文件，但是检索不到结果。**

  检查知识检索节点的“相关度阈值”参数是否设置过高（如0.9）。尝试在“知识库”中进行“命中测试”，观察真实阈值得分。

版权所有 © 华为云计算技术有限公司

图 4-154 知识库命中测试



- **检索到的内容与用户问题无关。**

  知识检索节点建议开启“混合检索”，或检查知识库切片长度，过短的切片可能丢失上下文背景信息。

- **如何查看检索到的图片。**

  知识检索节点需要开启“查看图片”功能，并且知识库本身需要启用“仅保留原图”功能（新建知识库、停用知识库后可以修改图片解析配置，配置完成后需重新启用知识库）。

图 4-155 修改知识库配置



图 4-156 设置知识库仅保留原图



4.14 节点配置管理

4.14.1 意图管理

意图包是意图分类的集中管理单元，用于将一组相关意图类别及其样例统一封装。在工作流的**高级意图识别节点**中直接引用预定义的意图包，即可高效实现大规模意图分类与智能路由。

版权所有 © 华为云计算技术有限公司

意图包基于大模型语义理解能力，通过匹配意图名称、描述、样例、关联词等多维度特征，将用户输入映射到最匹配的意图分类。

核心优势：

- **一次定义，多处复用：**同一意图包（如“通用闲聊包”“银行基础业务包”）可被多个工作流共享引用，避免重复配置，显著提升开发与维护效率。
- **独立管理，敏捷迭代：**意图包独立于工作流存在，支持批量导入/导出、版本管理与持续优化。更新后，所有引用该意图包的工作流将自动同步生效，保障业务逻辑一致。

前提条件

已**开通AgentArts服务**。

创建意图包

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 开发配置 ”，选择“意图管理”页签。

步骤3 单击界面右上角“新建意图包”。

步骤4 在“新建意图包”界面输入意图包名称。

说明

- 输入意图包名称不允许重复。
- 支持中英文、数字、下划线和中划线输入。

步骤5 单击 ＋ ，可在意图包中添加意图分类，分类信息包含名称和样例，如**图4-157**所示，参数说明如**表4-78**所示。

 版权所有 © 华为云计算技术有限公司

图 4-157 配置意图包

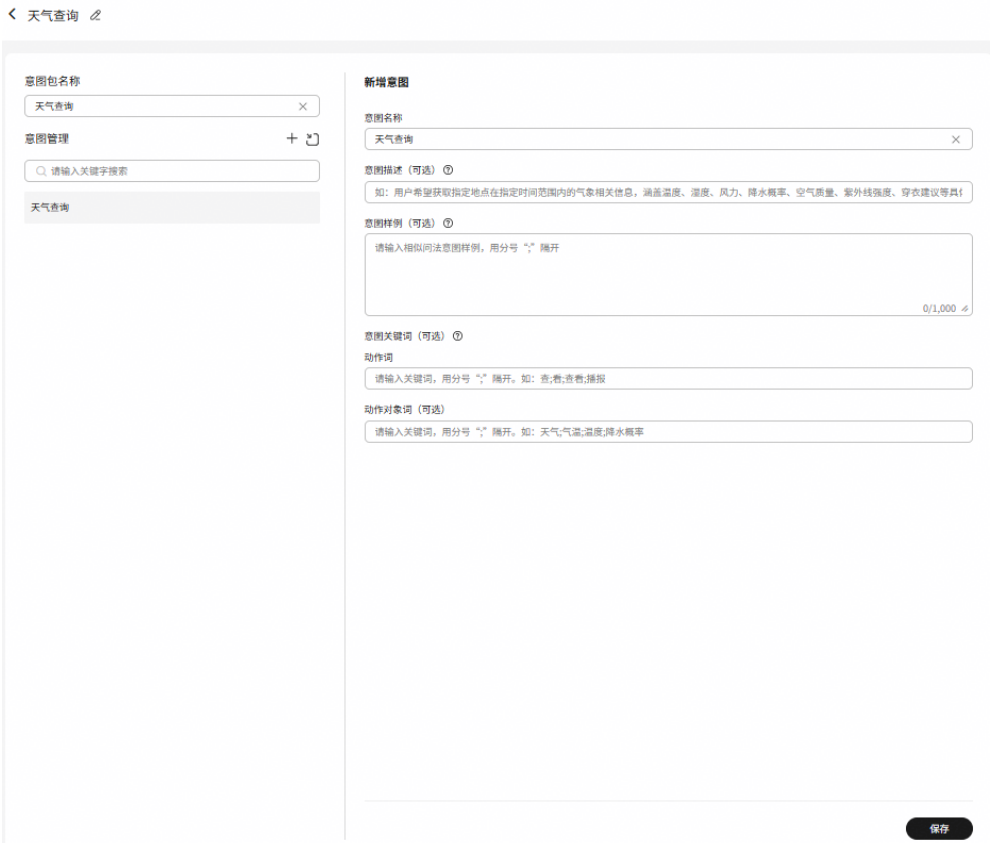


表 4-78 配置意图包

参数	说明	示例
意图名称	意图分类的唯一标识，同一意图包内不允许与已有分类名称重复，支持中英文、数字、下划线和中划线输入。	查询实时天气
意图描述（可选）	针对意图名称的解释，填写后可进一步提升模型对于用户意图的理解准确率。	用户希望获取指定地点当前时刻的天气概况，包含实时温度、天气状况（晴/阴/雨等）、体感温度、当前风力风向等基础气象信息
意图样例（可选）	• 填写和意图名称相关的样例，用于提升意图准确率。为了保证大模型的识别效果，每个意图分类建议至少提供**5-10 条**不同句式的样例。相似问法的意图样例，用英文分号“;” 隔开。 • 意图样例支持输入中文、英文字母、数字、中划线、下划线，最大支持1000个字符。	北京现在天气怎么样;上海现在几度;外面冷不冷;现在杭州下雨了吗;当前深圳天气如何

版权所有 © 华为云计算技术有限公司

参数	说明	示例
意图关键词（可选）	对用户输入内容进行关键词拆解，匹配预设的意图关键词。 • 动作词：描述用户希望执行的操作，用英文分号“;”隔开。例如：查;看;查看;播报。 • 动作对象词：指明动作针对的具体对象，用英文分号“;”隔开。如：天气;气温;温度;降水概率。	例如：用户意图为“我想要**查询**某天的**天气预报**” • 动作词：查;看;查看;播报 • 动作对象词：天气;气温;温度;降水概率

说明

意图管理页面支持对意图包重命名，添加意图、导入意图、批量录入、搜索、添加意图样例等。

- 单击 支持对意图包重命名。
- 单击 可添加意图。
- 单击 导入意图。
  - 单个sheet页的导入数据应不超过200条，超过数量不允许导入；
  - 上传文件限xlsx格式，文件不超过5MB，支持下载导入模板。
- 支持批量导入意图样例，多个意图样例名称须使用英文逗号分隔，最大支持输入1000字符。
- 搜索框中支持输入搜索。

步骤6 添加完成后，单击“保存”。保存后，您可以在“开发中心 > 开发配置 > 意图管理”界面中的“意图包列表”查看已创建的意图包。

步骤7 创建完成后，可在意图管理界面的意图包列表查看已创建的意图包。

如需修改可在意图包列表中单击“编辑”，进入意图包编辑页面。

图 4-158 编辑意图包



----结束

批量导入意图包

当需要从外部文件批量创建意图包或跨环境迁移时，使用导入功能。

步骤1 在“意图管理”界面，单击页面右上角“导入”。

步骤2 在导入意图包页面，单击“下载模板”可直接下载标准模板文件，按照模板格式填写。填写导入文件时，约束与限制如下：

- 单个sheet页的导入数据最大支持200条，大于200则不支持导入。
- 上传文件限xlsx格式，文件不大于5MB。

版权所有 © 华为云计算技术有限公司

- 意图包名称与已配置意图包名称不能重复。
- 多个意图样例名称须使用英文逗号分隔，最大支持输入1000字符。

步骤3 需要导入的文件填写好后，单击“导入”，每个Sheet页将作为独立的意图包显示在“意图管理”页面中。

导入后的意图包可在意图管理界面的意图包列表查看。

说明

如果文件中包含多个Sheet页，每个Sheet页将被创建为一个独立的意图包。

----结束

批量导出意图包

将意图包导出为xlsx文件，用于备份、跨环境迁移或离线编辑。

步骤1 在“意图管理”界面，单击页面右上角“导出”。

步骤2 在导出页面，勾选需要导出的意图包（支持多选），单击“导出”。单击页面左上角“导出”，文件将以xlsx格式下载至本地。

导出文件结构如表4-79所示。

下载完成后，打开xlsx文件检查各Sheet页中的意图包名称、分类、样例是否完整。

表 4-79 导出文件结构

导出内容	文件结构
单个意图包	单Sheet页，Sheet名称=意图包名称
多个意图包	多Sheet页，每个Sheet对应一个意图包

----结束

在工作流中使用意图包

意图包创建完成后，可以在工作流中的高级意图识别节点中引用，具体请参见高级意图识别。

4.14.2 消息模板

消息模板是预定义的标准化JSON消息结构，通过定义模板，可以规范工作流在特定场景（如系统异常、业务阻断）下的输出格式，实现统一错误处理与交互标准。

消息模板主要与工作流中的异常节点配合使用，异常节点被触发时，根据节点配置引用对应的消息模板，将模板中定义的JSON结构作为节点输出返回给用户或下游节点，确保多个工作流、多个节点之间的消息输出格式统一且可维护。

典型场景：当工作流中大模型调用节点发生超时异常时，异常节点触发并引用预定义的“模型超时提示”消息模板，输出包含错误码、提示信息和重试建议的标准化JSON消息，确保用户在不同工作流中看到一致的错误提示。

版权所有 © 华为云计算技术有限公司

前提条件

已**开通AgentArts服务**。

创建消息模板

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 开发配置 ”，选择“消息模板”页签。

步骤3 在“消息模板”页面，单击右上角的“新建消息模板”。

步骤4 在“创建消息模板”的弹框中输入消息的详细信息，如**图4-159**所示，参数说明请参见**表4-80**。

图 4-159 创建消息模板



表 4-80 创建消息参数说明

参数名称	参数说明
消息名称	模板的唯一标识名称，用于在节点中识别和选择。 支持中英文、数字、下划线、中划线，长度 2-64 字符，以中英文、数字开头。
消息分类	模板的用途分类。 • 异常：用于工作流异常节点的标准化错误消息输出，可在**异常节点**中引用。 • 自定义：预留分类，暂未开放此功能。

 版权所有 © 华为云计算技术有限公司

参数名称	参数说明
可见范围	设置消息模板的可见范围。当前支持以下两种设置。 • 租户内：整个租户内的用户均可查看并引用，适用于企业级统一消息规范，如全局错误码场景。 • 仅个人：仅创建者本人可见，适用于开发调试阶段的临时模板场景。

消息体	消息体必须为合法的JSON格式。单击右侧的 可以扩展内容输入框，方便输入更多详细信息。
	注意
	系统提供了校验功能，确保消息的语法正确性。如果消息体存在语法问题，将无法成功创建。
	示例1：简单文本消息（适用于简单错误提示场景）
	`{ "msg": "错误的URL、链接失效或网页已被删除", "code": "404" }`
	示例2：结构化错误消息（适用于含错误码的异常处理场景）
	`{ "type": "error", "code": "ERR_MODEL_TIMEOUT", "message": "模型响应超时，请检查网络连接后重试。", "suggestion": "如果问题持续，请尝试切换至其他模型。", "contact": "技术支持：support@example.com" }`
	示例3：多字段卡片消息（适用于需要用户交互的服务降级场景）
	`{ "type": "card", "title": "服务暂时不可用", "fields": [ {"label": "错误类型", "value": "服务过载"}, {"label": "影响范围", "value": "知识库检索功能"}, {"label": "预计恢复", "value": "10 分钟内"} ], "actions": [ {"label": "重试", "action": "retry"}, {"label": "联系支持", "action": "contact_support"} ] }`

步骤5 单击“确定”可保存设置的消息信息。创建完成后，您可以在“开发中心 > 开发配置 > 消息模板”界面中的“消息模板列表”查看已创建的消息模板。

----结束

在工作流中引用消息模板

消息模板创建后，可以在工作流的异常节点中引用，具体请参见**异常**。

引用后，可在工作流调试界面触发异常节点，检查输出消息是否符合模板定义的JSON结构，验证引用是否生效。

更多操作

消息模板创建完成后，您可以执行如**表4-81**的操作。

版权所有 © 华为云计算技术有限公司

表 4-81 相关操作

操作	说明
导入	支持批量导入功能。 1. 在“消息模板”页面，单击页面左上角的“导入”。 2. 在“导入消息模板”弹框中，单击“添加文件”选择需要导入的xlsx文件。 上传文件时，请注意以下限制条件： ● 上传文件限xlsx格式，文件不超过5MB； ● 单个sheet页的导入数据应不超过200条，超过数量不允许导入。 ● 在“导入消息模板”弹框中，单击“下载模板”，可以下载消息模板。
导出	在“消息模板”页面，选择需要导出的消息，单击“导出”，导出的消息会以xlsx的格式保存在本地。
编辑	编辑消息模板的内容。在“消息模板”页面，找到需要编辑的消息，单击操作列中的“编辑”，在弹出的“编辑消息模板”对话框中进行内容编辑。
删除	删除消息模板。在“消息模板”页面，勾选需要删除的消息模板，单击“删除”，可以删除消息模板。 ● 单个删除：在待删除的消息模板对应的“操作”列下，单击“删除”。 ● 批量删除：勾选待删除消息模板，单击页面左上角的“删除”。

常见问题

问：修改了消息模板后，已发布的工作流会受影响吗?

答：不会。已发布的工作流版本使用发布时的模板快照。修改仅影响未发布（开发中）的工作流版本在下次执行时的输出。如需更新已发布版本，请重新**发布并部署工作流**。

4.14.3 对象管理

对象是可复用的变量定义模板，将业务中常用的数据结构（如用户信息、订单详情、收货地址等）预先定义为对象，可在工作流**全局配置**中的记忆变量、**开始节点**中直接引用，无需重复定义相同的变量，显著提升开发效率与数据一致性。

每个对象包含一个对象名称和一个对象变量（变量名称 + 变量类型 + 描述）。当变量类型为Object时，支持定义多层嵌套的结构化字段。

新建对象

步骤1 登录**AgentArts智能体平台**。

版权所有 © 华为云计算技术有限公司

步骤2 单击左侧导航栏“开发中心 > 开发配置 ”，选择“对象管理”页签。

步骤3 单击页面右上角的“新建对象”。

步骤4 设置新建对象的基本配置信息，具体的参数说明如表1所示。

表 4-82 新建对象参数说明

参数	说明
对象名称	自定义对象管理的名称，支持中英文、数字、下划线、中划线，长度 2-64 字符，以中英文、数字开头。 建议使用业务语义命名，例如客户姓名、订单号、收货地址等。
对象变量	自定义对象变量包括变量名称，变量类型和描述（可选）。 • 变量名称：不允许为空，只允许输入字母、数字、下划线，且不能以数字开头。例如，customer_name、order_id等。 • 变量类型：支持配置String、File、Integer、Number、Boolean、Object、Array多种类型的参数，其中Object类型参数**最多支持5层嵌套**。 • 描述（可选）：对变量的详细描述。**强烈建议填写**，描述帮助大模型理解该提取什么内容。

图 4-160 配置示例



步骤5 单击“确定”。创建完成后，您可以在“开发中心 > 开发配置 > 对象管理”界面中的“对象列表”查看已创建的对象。

----结束

在工作流中引用对象模板

对象模板创建后，可以在工作流**全局配置**中的记忆变量（对话型工作流）、**开始节点**进行引用，精准提取并结构化业务所需数据。

更多操作

新建对象创建完成后，您可以执行如**表4-83**的操作。

版权所有 © 华为云计算技术有限公司

表 4-83 相关操作

操作	说明
搜索对象	输入部分对象名称后，单击或按Enter键，符合条件的对象会显示在列表中。
编辑对象	在显示结果中，单击操作列的“编辑”，可以修改对象的名称、对象变量参数。
删除对象	在显示结果中，单击操作列的“删除”，然后在弹出的确认框中单击“确认”。 **说明** 删除对象后，工作流中已引用的参数不受影响，但该对象将从对象列表中移除，后续创建工作流将无法再引用该对象。如果需要继续使用，请重新创建。

版权所有 © 华为云计算技术有限公司

5 开发多智能体应用

5.1 多智能体应用介绍

在AgentArts中，单智能体应用能够处理基本任务。但在面对复杂任务时，单智能体往往需要编写冗长的提示词，并挂载多种插件、知识库或MCP服务。这不仅增加了调试的复杂性，且任意一处改动都可能影响整体功能，导致最终结果与预期存在较大偏差。

为解决这一问题，AgentArts提供了**多智能体应用**。多智能体应用通过**中央控制器**对用户意图进行识别与分发，灵活调度多个子智能体协同完成任务，显著提升了复杂任务处理的效率与准确性。

说明

中央控制器基于大语言模型对用户输入进行意图分类，并根据预设的路由规则将请求分发至对应的子智能体。

核心优势

- **意图驱动的路由机制**：支持根据用户意图，在不同子智能体之间自动跳转与分发，无需人工干预。
- **模型自动调度模式**：引入中央控制器，实现对各子智能体的自动化编排与协同，提升执行准确率。
- **多层级组织架构**：支持**2层嵌套控制**（多智能体→多智能体→子智能体），便于构建复杂的任务分解体系。
  - 第1层：主控多智能体，负责意图识别与全局调度。
  - 第2层：子多智能体，可进一步分解子任务。
  - 第3层：执行子智能体，完成原子任务。

适用场景

多智能体应用适用于业务场景多样、需根据用户意图精准分流至不同专业子智能体的场景。

典型示例：智能客服路由系统

当用户询问“如何重置密码”或“查询本月账单”时：

 版权所有 © 华为云计算技术有限公司

1. **意图识别**：中央控制器识别用户意图是“账户操作”还是“账单查询”。
2. **路由分发**：
   - 如果为“账户操作”，则直接分发至**“账户管理子智能体”**处理。
   - 如果为“账单查询”，则直接分发至**“账单服务子智能体”**处理。
3. **结果返回**：由被调用的子智能体直接生成回复并返回给用户，中央控制器仅负责“指路”，不参与后续的逻辑编排。

单智能体与多智能体选型对比

如果您的任务逻辑简单、工具调用类型单一，推荐使用**单智能体应用**；如果任务涉及多个子流程、需要根据用户意图灵活分发，或需要多步骤协同完成，推荐使用**多智能体应用**。

表 5-1 单智能体与多智能体选型对比

对比维度	单智能体	多智能体
适用任务复杂度	简单、单一任务	复杂、多步骤、多分支任务
核心机制	模型自主规划，调用工具完成任务	中央控制器识别意图，调度子智能体协同完成
工具使用	插件、开发工作流应用、知识库、MCP等	多个工作流，按用户意图路由分发
调测复杂度	较低，提示词集中管理	较高，需分别调测各子智能体及路由逻辑
维护成本	较低，但改动影响范围大	较高，但各子智能体可独立维护
组织层级	单层	最多2级嵌套

相关文档

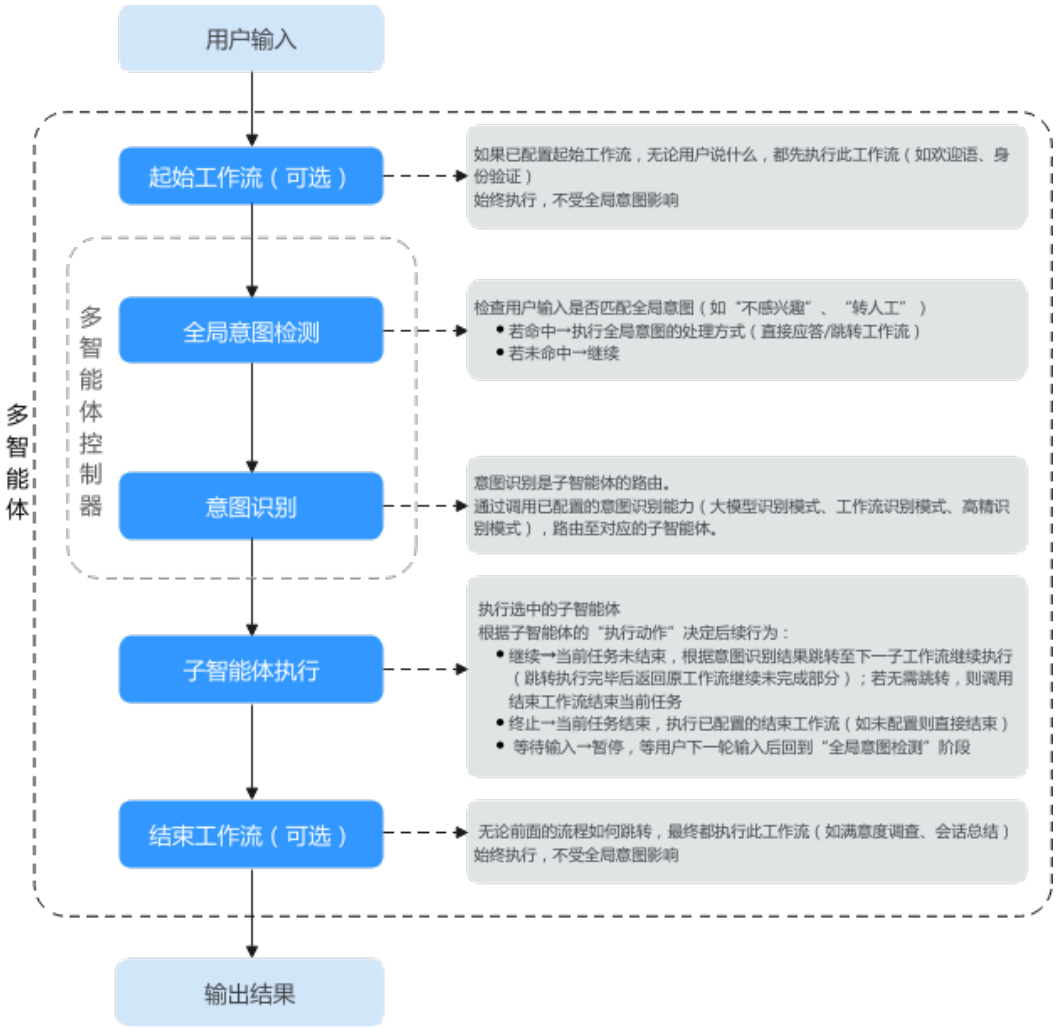
- 开发多智能体应用
- 开发单智能体应用
- 开发工作流应用
- 单智能体应用与多智能体应用是否可以切换?
- 工作流应用和多智能体应用有什么区别?
- 在单智能体应用中使用工作流与在多智能体应用中使用工作流有什么区别?

5.2 多智能体执行范式

AgentArts多智能体采用“多Agent控制器+子智能体”调度架构，适用于企业业务场景。一次完整的多智能体交互，按图5-1所示顺序执行。

版权所有 © 华为云计算技术有限公司

图 5-1 多智能体执行流程



基于多Agent控制器的三种执行动作（继续/终止/等待输入）、生命周期工作流（起始/默认/结束）、全局意图和嵌套能力，AgentArts的多智能体支持如表5-2所示的三种执行范式和一种贯穿机制。

范式和贯穿机制的区别

- **执行范式**（范式一～三）是**递进增强**的架构选择：路由分发为基础范式，固定生命周期管控和分层嵌套控制为增强范式，可按需叠加组合。
- **贯穿机制**（全局意图中断）是**可叠加的能力**，可以在任意执行范式上叠加使用，不影响主架构的选择。

表 5-2 多智能体执行范式/贯穿机制

类别	范式名称	核心机制	关键配置项	复杂度
范式一	路由分发	控制器根据用户意图选择子智能体	意图识别/模型+提示词	★
范式二	固定生命周期管控	起始→业务处理→结束的固定流程	起始/结束工作流	★★

版权所有 © 华为云计算技术有限公司

类别	范式名称	核心机制	关键配置项	复杂度
范式三	分层嵌套控制	多智能体嵌套多智能体，多层路由	子智能体类型选择“多智能体”（2级嵌套）	★★★
贯穿机制	全局意图中断	任何时刻优先响应全局意图	全局意图配置	★★

选型决策

表 5-3 选型决策

场景特征	推荐范式组合
仅需按意图分发到不同子智能体	范式一
需要固定开头/结尾流程（如身份验证+满意度评价）	范式一+范式二
需要多层分类（如售前/售后→细分意图）	范式一+范式三
需要固定开头/结尾+多层分类	范式一+范式二+范式三
任何组合+需要随时中断（如“转人工”、“不感兴趣”）	叠加贯穿机制

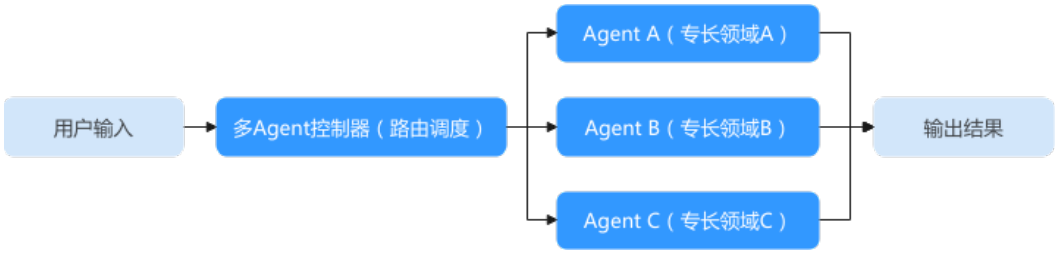
说明

多智能体每轮对话会多次调用大模型（全局意图检测+意图识别+子工作流执行），Token消耗显著高于单工作流。分层嵌套控制模式下每轮需执行两次意图识别，费用更高。建议合理控制子智能体数量和意图分类数量。

路由分发

多Agent控制器接收用户输入后，通过意图识别（工作流识别/高精识别/模型+提示词）判断用户意图，一次性将请求路由到最匹配的子智能体。每轮对话独立路由，子智能体之间无强依赖。

图 5-2 顺序执行运行机制



在AgentArts中配置“路由分发”模式的要点如表5-4所示。

版权所有 © 华为云计算技术有限公司

表 5-4 配置要点

配置项	设置
意图识别	配置意图识别能力（大模型识别模式、工作流识别模式、高精识别模式），将用户输入映射至对应的子智能体
子智能体意图描述	每个子智能体写清楚能力范围和触发条件
子智能体执行动作	**终止**（一次性任务）或**等待输入**（多轮对话循环）

表 5-5 适用场景

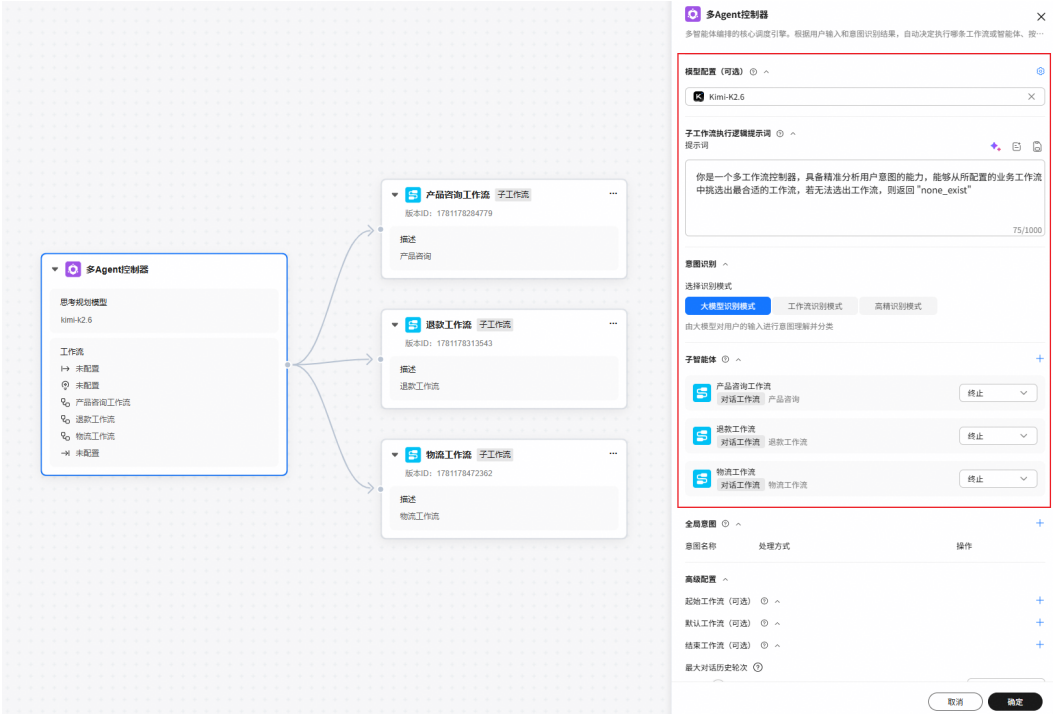
场景	说明
企业客服	售前、售后、技术支持各一个子智能体
多技能助手	编程、写作、分析各一个子智能体
多语言服务	中文、英文、日文各一个子智能体

路由分发示例

以“在线客服”多智能体为例，支持产品咨询、退款处理、物流查询。

- 当用户咨询产品怎么收费时，路由到“产品咨询”子工作流，“产品咨询”子工作流回答价格信息，流程终止。
- 当用户输入“我要退款”时，路由到“退款申请”子工作流，“退款申请”子工作流处理退款事务，处理完后流程终止。

图 5-3 在线客服

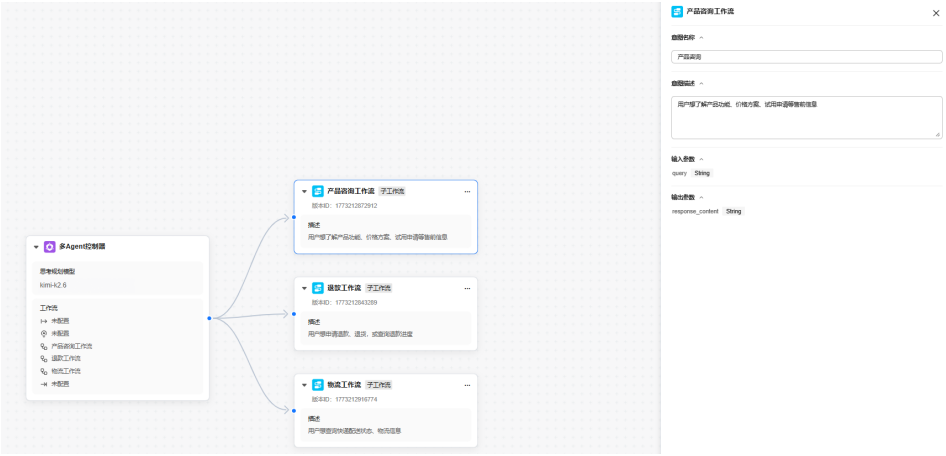


版权所有 © 华为云计算技术有限公司

关键配置如下：

- **模型配置：**选择“Kimi-K2.6”。
- **子智能体：**选择已创建并发布的“产品咨询”工作流、“退款”工作流和“物流”工作流。
- **配置子工作流的意图描述：**
  - 单击画布中的“产品咨询”工作流，配置意图名称为：产品咨询，意图描述为：用户想了解产品功能、价格方案、试用申请等售前信息。如图5-4所示。

    图 5-4 配置意图名称和意图描述

  - 单击画布中的“退款”工作流，配置意图名称为：退款申请，意图描述为：用户想申请退款、退货，或查询退款进度。
  - 单击画布中的“物流”工作流，配置意图名称为：物流查询，意图描述为：用户想查询快递配送状态、物流信息。

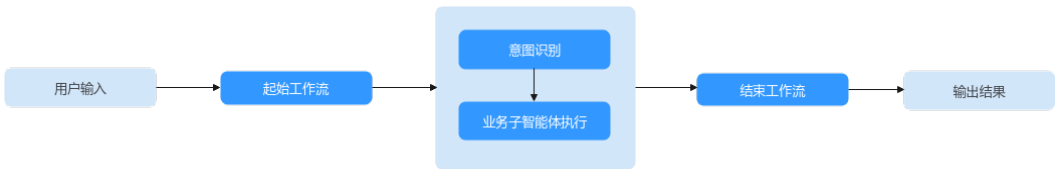
运行示例如下：

用户: 你们产品怎么收费? → 路由到“产品咨询”子工作流 → 回答价格信息 → 终止
用户: 我要退款 → 路由到“退款申请”子工作流 → 处理退款 → 终止

固定生命周期管控

通过**起始工作流**和**结束工作流**在业务逻辑的两端设定**固定的“锚点”**，不论中间的业务流程如何跳转，这两个工作流始终执行，确保会话有一致的开头和结尾。

图 5-5 多智能体固定生命周期管控运行机制



在AgentArts中配置“固定生命周期管控”模式的要点如表5-6所示，适用场景如表5-7所示。

 版权所有 © 华为云计算技术有限公司

表 5-6 配置要点

配置项	设置
高级配置 > 起始工作流	选择欢迎/验证工作流
高级配置 > 结束工作流	选择总结/调查工作流
高级配置 > 默认工作流	选择兜底引导工作流

说明

- 起始和结束工作流**不受全局意图影响**——即使用户在中间说“不感兴趣”触发全局意图终止，结束工作流仍然会执行。
- 起始和结束工作流**不计入最大跳转次数**。

表 5-7 适用场景示例

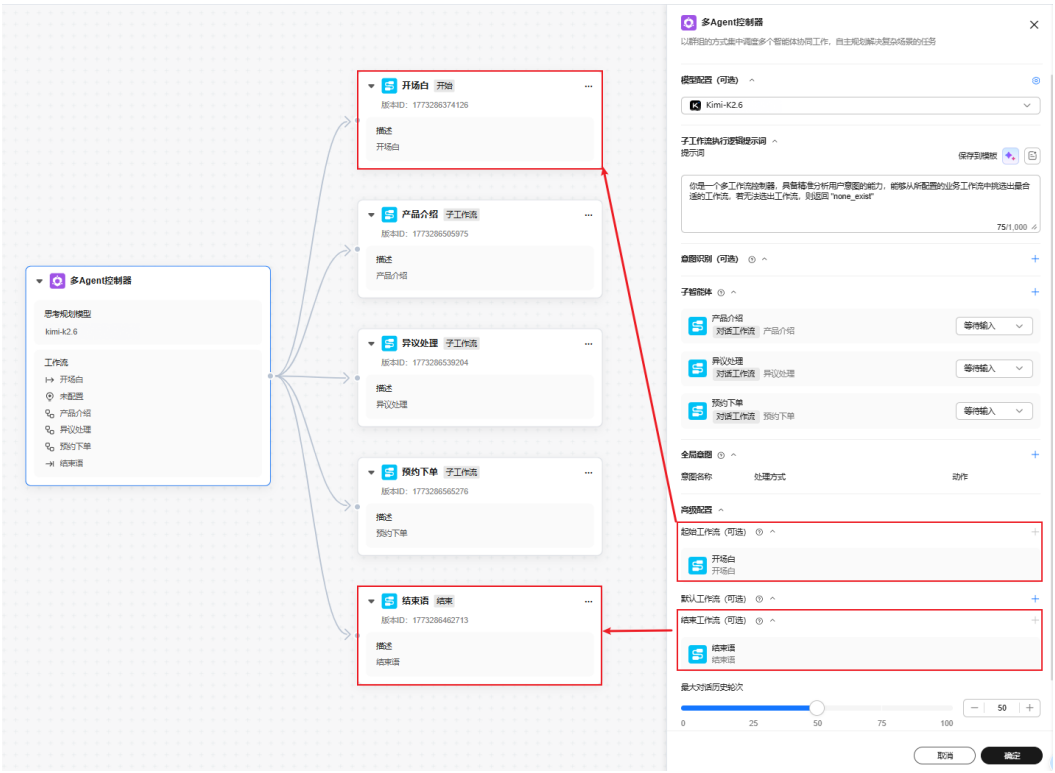
场景	起始工作流	结束工作流
智能外呼	开场白+确认身份	结束语+满意度评价
在线问诊	挂号验证+隐私声明	诊后建议+随访提醒
业务办理	身份认证+权限校验	办理总结+工单编号
用户调研	研究说明+知情同意	感谢语+奖励发放

固定生命周期管控示例

以“智能外呼系统”多智能体为例，用户拨通后，系统运行起始工作流，播报欢迎语，然后用户咨询产品相关问题，对应路由到相应的业务子智能体；用户咨询结束后，运行结束工作流，播报结束语并归档。

版权所有 © 华为云计算技术有限公司

图 5-6 智能外呼系统



关键配置如下：

- 起始工作流：选择已创建并发布的“开场白”工作流。
- 结束工作流：选择已创建并发布的“结束语”工作流。

执行流程示例如下：

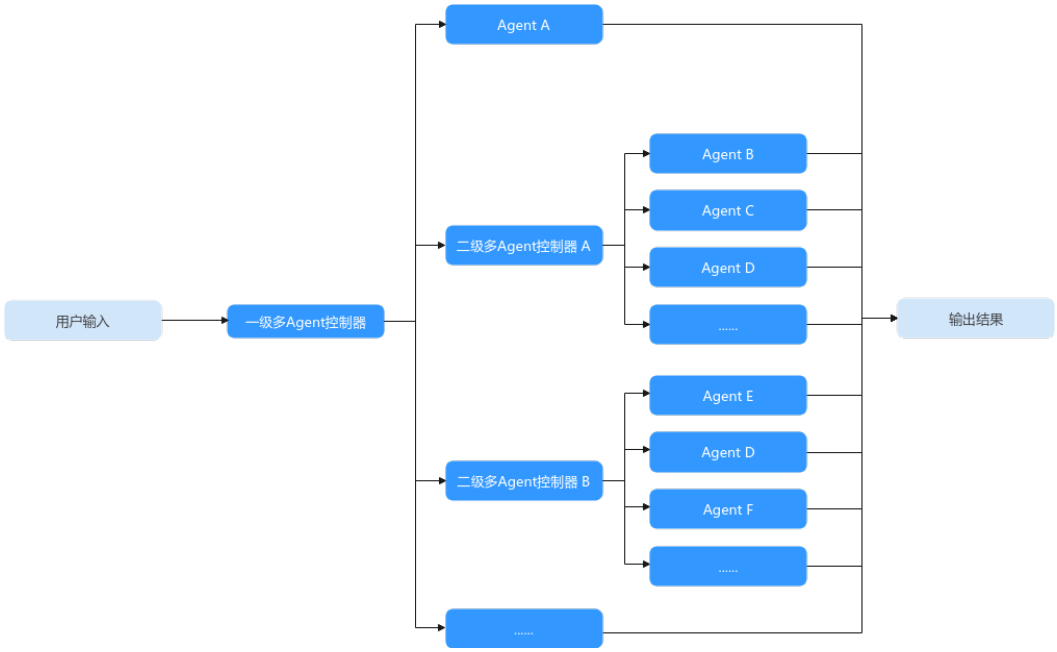
[起始工作流: 开场白]	您好，我是您的专属顾问，请问您有什么问题咨询吗？	（始终执行，不计入跳转次数）
↓		
[业务子工作流调度]	产品介绍 ↔ 异议处理 ↔ 预约下单	（根据用户回应动态路由）
↓		
[结束工作流: 结束语]	感谢您的时间，祝您生活愉快！	（始终执行，不计入跳转次数）

分层嵌套控制

将一个多智能体应用作为另一个多智能体的“子智能体”，实现**多层路由决策**。一级控制器做粗分类，二级控制器做细分类。

版权所有 © 华为云计算技术有限公司

图 5-7 分层嵌套控制运行机制



在AgentArts中配置“分层嵌套控制”模式的要点如表5-8所示，适用场景如表5-9所示。

表 5-8 配置要点

配置项	设置
一级控制器多智能体	添加类型 = “多智能体”
嵌套层级限制	最多**2级**（一级 → 二级 → 子工作流）
二级多智能体	需独立创建并发布，拥有独立的全局配置、生命周期工作流和全局意图

表 5-9 适用场景示例

场景	一级路由	二级路由
大型客服系统	售前/售后/技术（3大域）	每个域内5-10个细分意图
多部门企业	人事/财务/IT/行政	每个部门内多个服务项
跨产品线服务	产品A/产品B/产品C	每个产品内独立的服务体系

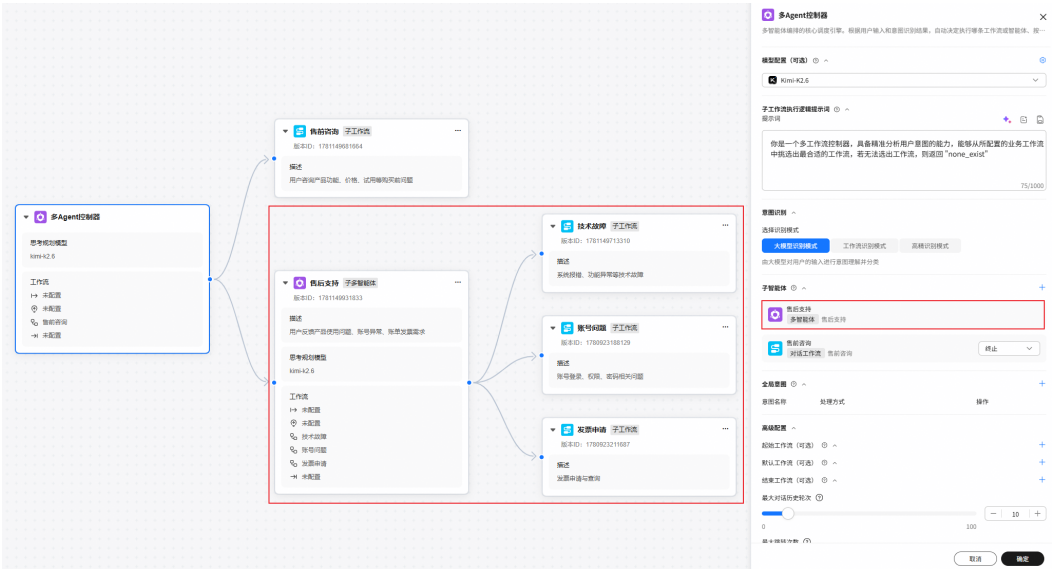
分层嵌套控制示例

以一个企业在线客服场景为例，用户输入一个问题后，多智能体的一级控制器识别用户意图，将其路由到对应的子工作流或多智能体分支，如果是路由到子多智能体，则子多智能体继续根据用户意图路由到对应的子工作流分支。

版权所有 © 华为云计算技术有限公司

如图5-8所示，客服场景涵盖“售前咨询”、“售后支持”两大域，“售后支持”进一步细分为“技术故障”、“账号问题”、“发票申请”三类场景。用户输入一个问题后，例如“我要开发票”，多智能体的一级控制器识别用户意图，将其路由到“售后支持”子多智能体分支，则子多智能体识别到用户意图后，继续路由到对应的“发票申请”子工作流分支。

图 5-8 企业在线客服



关键配置如下：

一级控制器子智能体配置如表5-10所示，二级控制器子智能体配置如表5-11所示。

表 5-10 一级控制器子智能体配置

子智能体类型	名称	意图描述	执行动作
工作流	售前咨询	用户咨询产品功能、价格、试用等购买前问题	终止
多智能体	售后支持	用户反馈产品使用问题、账号异常、账单发票需求	终止

表 5-11 二级子多智能体配置

工作流名称	意图描述	执行动作
技术故障	系统报错、功能异常等技术故障	等待输入（需收集错误信息）
账号问题	账号登录、权限、密码相关问题	终止
发票申请	发票申请与查询	终止

版权所有 © 华为云计算技术有限公司

执行流程示例1：

用户: 我的账号突然登不进去了
↓
[一级控制器多智能体] 意图识别 → 命中“售后支持” → 路由到“售后支持”子多智能体
↓
[二级控制器“售后支持”子多智能体] 意图识别 → 命中“账号问题” → 路由到“账号问题”工作流
↓
[“账号问题”工作流] 引导用户完成账号找回流程 → 终止

执行流程示例2：

用户: 系统一直报错 500，无法使用
↓
[一级控制器多智能体] 意图识别 → 命中“售后支持” → 路由到“售后支持”子多智能体
↓
[二级控制器“售后支持”子多智能体] 意图识别 → 命中“技术故障” → 路由到“技术故障”工作流
↓
[“技术故障”工作流] “请您描述一下报错的具体场景和错误代码...” → 等待输入 → 继续收集用户信息（例如报错信息）

全局意图中断

在任何对话阶段，**全局意图始终优先于子智能体意图进行检测**。一旦用户触发全局意图，立即中断当前业务流程，执行全局意图的处理逻辑。

全局意图与结束工作流的交互规则：

- 全局意图触发“终止”后，结束工作流仍然执行（如果已配置），确保会话收尾逻辑（如日志记录、满意度评价、结束语）依然完整。
- 全局意图触发“继续”或“等待输入”后，流程按执行动作继续，结束工作流在所有子智能体执行完毕后才触发。

在AgentArts中配置“全局意图中断”模式的要点如表5-12所示，适用场景如表5-13所示。

表 5-12 配置要点

配置项	设置
全局意图名称	简短明确的意图标识
处理方式	直接应答（简单场景）/流程跳转（需要操作的场景）
执行动作	多数中断场景用“终止”；需重新进入对话流程的用“等待输入”（如“重新开始”）

表 5-13 适用场景示例

全局意图	处理方式	执行动作	场景
不感兴趣	直接应答	终止	外呼场景，用户拒绝
非本人	直接应答	终止	外呼场景，接听人不是目标用户
转人工	流程跳转	终止	任何场景，用户要求人工服务

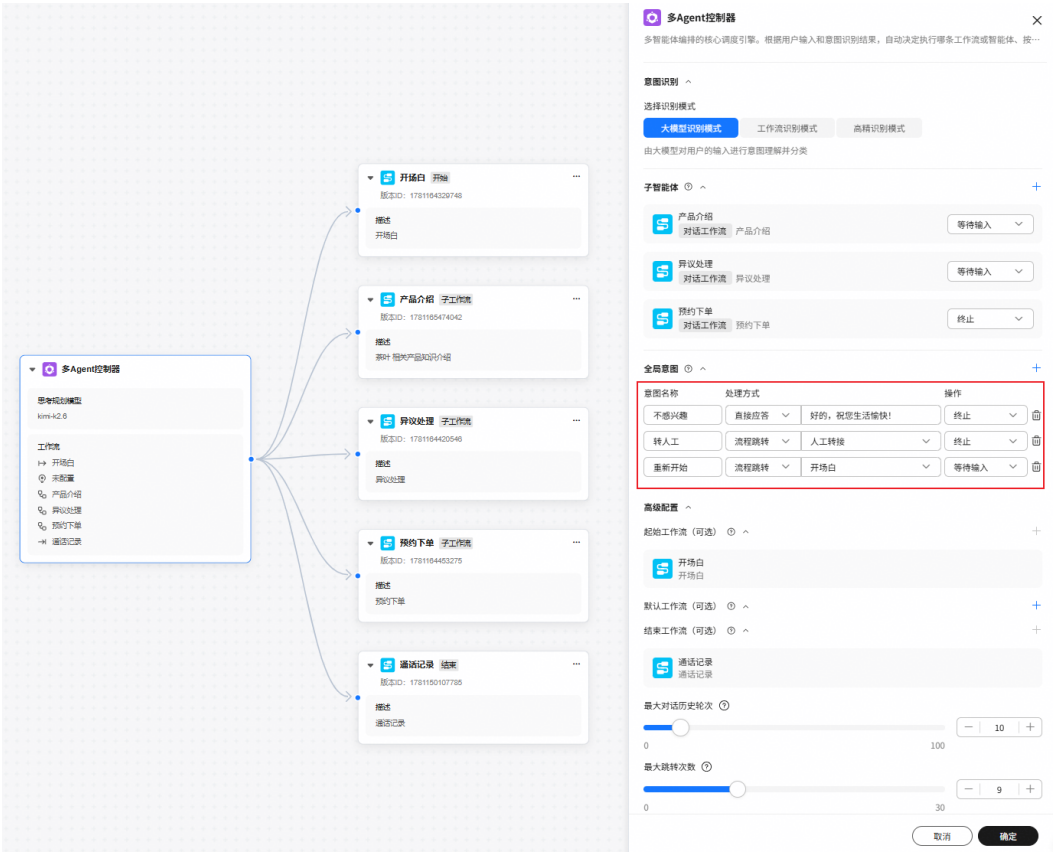
版权所有 © 华为云计算技术有限公司

全局意图	处理方式	执行动作	场景
重新开始	直接应答	等待输入	用户想重新描述需求
查看帮助	直接应答	等待输入	展示功能菜单后继续对话

全局意图中断示例

以“智能外呼系统”多智能体为例，用户拨通后，系统运行起始工作流，播报欢迎语，用户咨询产品期间，随时可能说“不感兴趣”、“转人工”或“重新开始”等中断对话。此时，根据系统配置的处理方式，执行对应的动作。

图 5-9 智能外呼系统



执行流程示例1（当用户说“不感兴趣”时）：

[起始工作流] → 您好，xx，我是专属顾问……
↓
[控制器] → 调度“产品介绍”子工作流
↓
[“产品介绍”工作流] →“我们的产品具有以下优势...” → 等待输入
↓
用户:“没兴趣，别打了”
↓
[全局意图检测] → 命中“不感兴趣”（优先级高于子工作流路由）
↓

版权所有 © 华为云计算技术有限公司

[直接应答] → "好的，祝您生活愉快！" → 执行动作: 终止
↓
[结束工作流] → 记录通话结果 → 任务结束

执行流程示例2（当用户说“转人工”时）：

[起始工作流] → 您好，xx，我是专属顾问……
↓
[控制器] → 调度“异议处理”子工作流
↓
[“异议处理”工作流] →“理解您的顾虑，这款产品其实……” → 等待输入
↓
用户:“转人工”
↓
[全局意图检测] → 命中“转人工”（优先级高于子工作流路由）
↓
[流程跳转] → 跳转至“人工转接”工作流 → 转接客服 → 执行动作: 终止
↓
[结束工作流] → 记录通话结果 → 任务结束

5.3 创建并配置多智能体应用

多智能体应用根据用户意图，将请求路由分发至对应的子工作流或子智能体执行，实现多场景的统一调度。多智能体应用的更多介绍请参考**多智能体应用介绍**。

本节介绍配置多智能体应用的流程。

费用说明

不同套餐包允许创建的**应用数量限制**不同，具体可参考**计费说明**。

前提条件

- 已**开通AgentArts服务**。
- 已**发布并部署工作流**。

创建多智能体应用

创建多智能体的基础信息。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击“创建多智能体”。

步骤4 在“创建多智能体”页面，配置基础信息，具体参数说明请参考**表5-14**。

版权所有 © 华为云计算技术有限公司

图 5-10 创建多智能体

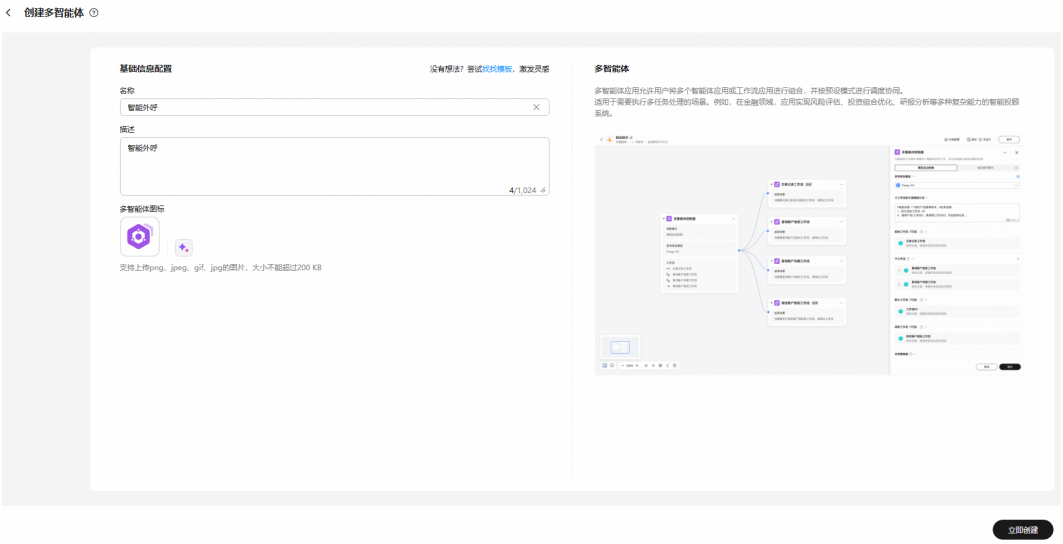


表 5-14 基础信息参数说明

参数	说明	示例
名称	多智能体应用的名称。由2~64个字符组成，包含中英文、数字、下划线、中划线、空格，不能以空格开头或结尾。	智能外呼
描述	多智能体的描述信息。由1~1024个字符组成。	智能外呼
多智能体图标	系统默认多智能体应用图标，可通过单击  自动生成图标，也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	系统默认图标

步骤5 单击“立即创建”。

创建后，进入“多智能体配置”页面，初始只有一个“多Agent控制器”节点。创建的多智能体应用显示在多智能体应用卡片列表中。

----结束

全局配置

在“多智能体配置”页面右上方，单击“全局配置”。

全局配置可配置输入参数和全局变量，这些都可以给工作流的输入参数使用。

版权所有 © 华为云计算技术有限公司

注意

输入参数的值不可修改，全局变量的值可被子工作流的输出参数覆盖，请注意区分使用场景。

图 5-11 全局配置



表 5-15 全局配置参数说明

参数	说明
输入参数	传给工作流的输入参数，且值不可修改。 单击 ＋ ，添加输入参数。 ● 参数名称：由1~64字符组成，包含字母、数字、下划线，不能以数字开头。 ● 类型：类型支持String、Integer、Number、Boolean。默认值String。 ● 描述：由0~256个字符组成。 ● 必填：根据需求去掉勾选。默认勾选。 ● 🗑 ：删除输入参数。 ● ··· ：输入参数默认值。

版权所有 © 华为云计算技术有限公司

参数	说明
全局变量	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覆盖该值。 单击 ，添加全局变量。 ● 参数名称：由1~64字符组成，包含字母、数字、下划线，不能以数字开头。 ● 类型：类型支持String、Integer、Number、Boolean。默认值String。 ● 描述：由1~256个字符组成。 ● ：删除全局变量。 ● ：全局变量默认值。

配置多 Agent 控制器

配置多智能体，实现特定业务能力。

步骤1 单击“多Agent控制器”卡片，在弹出页面配置参数信息。

 版权所有 © 华为云计算技术有限公司

图 5-12 配置多 Agent 控制器



版权所有 © 华为云计算技术有限公司

步骤2 （可选）配置模型。

此处的模型用于意图识别和路由调度，负责理解用户输入的意图并决定路由到哪条工作流或智能体。

⚠ 注意

“模型配置（可选）”与“意图识别”参数，至少填写一个。如果意图识别工作流或意图包未返回有效意图，则使用选择的模型进行意图识别。

AgentArts支持在下拉框中选择合适的模型，如DeepSeek-V4-Flash。已接入的模型服务详见**模型**。

说明

模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型服务**时的“选择标签”、“模型类型”。

- 接入模型时的“选择标签”：
  - 联网：表示该大模型具备联网搜索能力。
  - 思考：表示该大模型具备思维推理能力。
  - 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。
  - 限额体验：表示该平台预置大模型可以体验，会话轮数最大为20次。
- “模型类型”包含：
  - 文本：表示该大模型是文本对话类型。
  - 视觉：表示该大模型是图像理解类型。
  - 嵌入：表示该大模型是文本向量化类型。
  - 排序：表示该大模型是文本排序类型。
- 模型状态：
  - 未验证：表示该大模型未校验鉴权信息，不可使用。
  - 成功：表示该大模型鉴权信息校验成功，可以使用。
  - 失败：表示该大模型鉴权信息校验失败，不可使用。

在“模型配置”右侧，单击 ，显示如下参数：

- **核采样**：
  核采样（Top-P）用于限制模型的选词范围。取值范围0.1~1，默认值为0.5。
  - 值接近0时，模型只从概率最高的前10%词汇中选择，输出更集中。
  - 值接近1时，不做限制，输出更多样。

  建议该参数和“生成随机性”只设置1个，避免同时调整后效果不可控。
- **生成随机性**：
  生成随机性（temperature）控制模型输出的随机性和创造性，取值范围0~1，默认值为0.5。
  - 数值越低（接近0）使输出更加确定和一致，适合基于事实的问答和售后客服场景。

 版权所有 © 华为云计算技术有限公司

- 数值越高（接近1）使输出更具多样性和创新性，适合创造性任务如小说创作。

建议该参数和“核采样”只设置1个，避免同时调整后效果不可控。

步骤3 （可选）配置子工作流执行逻辑提示词。

为编排控制器定义调度逻辑的系统提示词。通过自然语言描述子工作流的执行顺序、优先级规则和协作逻辑，指导模型做出正确的路由决策。

示例：

```
你是一个智能客服调度系统。根据用户输入判断意图，按以下规则路由到对应的子工作流。
1、如果用户询问订单相关问题，执行“订单查询”工作流。
2、如果用户要求退换货， 执行“退换货处理”工作流。
3、如果用户表达不满或投诉，执行“投诉处理”工作流。
4、如果无法判断意图，执行默认工作流。
```

用户可单击 图标直接引用提示词模板；也可以自行撰写提示词，撰写后可单击 图标对系统提示词进行智能优化；也可单击 图标保存到模板，在“组件库 > 提示词”中可查看该模板。

步骤4 配置意图识别。

选择该多智能体的意图识别能力。

⚠ 注意

“模型配置（可选）”与“意图识别”参数，至少填写一个。优先调用意图识别工作流或意图包。

- **大模型识别模式**：由大模型对用户的输入进行意图理解并分类。选择此值时，“模型配置（可选）”需选择模型。
- **工作流识别模式**：由开发者自行编排一个专门用来做意图识别场景的工作流进行处理，可以兼顾准确率和性能。

  在“选择工作流”右侧，单击 ，选择意图识别工作流。意图识别工作流的开始节点的输入参数、结束节点的输出参数配置有特殊要求。

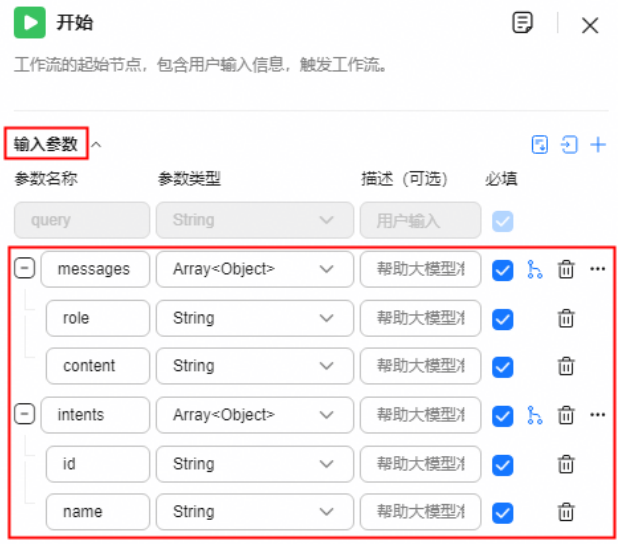
  多智能体运行时，运行流程依次为：用户输入 -> 意图识别工作流 -> 根据意图识别工作流输出的intent_id执行任务。

  意图识别工作流的开始节点的输入参数、结束节点的输出参数配置如下：

  - 开始节点必须配置如图5-13所示参数，参数名称和参数类型必须保持一致。
    - messages：对话历史。
      - role：角色。
      - content：内容。
    - intents：意图。
      - id：意图序号，取值为从1开始的自然数。
      - name：当id的取值为1时，表示没有明确意图或无法判断用户输入意图；当id的取值为非1时，表示对应的意图。

 版权所有 © 华为云计算技术有限公司

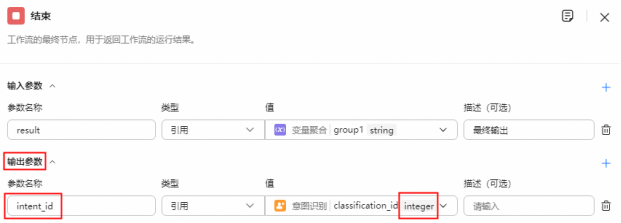
图 5-13 开始节点配置



- 结束节点必须配置如图5-14所示参数，输出参数名称以及引用的值必须为integer类型。

  intent_id：子工作流输出的意图结果。输出结果与多智能体中子智能体的对应关系请参考**意图识别输出结果说明**。

图 5-14 结束节点配置



- **高精识别模式**：开发者预先创建意图包，预设多类意图场景，系统结合配置能大幅提升海量意图的识别准确率。

  在“选择意图包”下的下拉框中选择意图包。单击“立即创建”可以创建意图包，具体操作请参考**创建意图包**。

步骤5 配置子智能体。

每个子智能体对应一类用户意图或业务场景，多Agent控制器根据用户意图自动路由执行，调度逻辑可通过“执行逻辑提示词”灵活定义。

在子智能体右侧，单击 ，支持添加工作流、多智能体。

- 工作流：最多支持添加50个子工作流。

  添加工作流后，设置子工作流的执行动作，支持的执行动作如下：

  - **继续**：执行完该工作流后，继续执行后续流程。有两种场景：
    - **场景1：工作流间跳转后继续**。执行工作流A时，如果意图识别需要执行工作流B（且工作流B设为“继续”），则执行完工作流B后，返回工作流A，继续执行A中尚未完成的部分。

版权所有 © 华为云计算技术有限公司

        - **场景2：工作流自身执行完毕后继续**。如果工作流A自身设为“继续”，则执行完工作流A后，调用结束工作流结束当前任务。
    - **终止**：执行完该工作流后，调用结束工作流结束当前任务。
    - **等待输入**：执行完该工作流后，暂停并等待用户输入问题后再继续执行。
- 多智能体：在多智能体应用中，可以嵌套引用另一个多智能体应用，从而实现多层控制的效果。目前支持2级控制。最多支持添加50个智能体。

步骤6 （可选）配置全局意图。

定义全局性的意图识别规则。无论当前正在执行哪条子工作流，一旦检测到匹配的全局意图，将立即触发对应的处理动作。适用于随时可能出现的跨场景指令（如“转人工”、“退出”、“切换话题”）。

在“全局意图”右侧，单击 ，输入意图名称、处理方式、意图的执行动作。

支持如下处理方式：

- **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。

支持如下执行动作：

- **继续**：按直接应答/工作流的执行结果，继续执行其他的子工作流。
- **终止**：按直接应答/工作流的执行结果，调用结束工作流结束任务。
- **等待输入**：按直接应答/工作流的执行结果，待用户输入问题后执行任务。

步骤7 （可选）配置起始工作流。

无论用户意图识别到哪条子工作流，起始工作流始终最先执行。适用于用户身份验证、会话初始化、欢迎语生成等前置操作。

未配置时，系统根据选择的模型、子工作流执行逻辑提示词或意图识别进行意图判断。

执行顺序：起始工作流 → 子工作流（按意图路由）→ 结束工作流

配置起始工作流操作如下：

1. 在“起始工作流（可选）”右侧，单击 ，在“添加工作流”页面选择作为起始的工作流应用。
2. 单击“确定”。

   如果选择的工作流需要删除，单击 。

步骤8 （可选）配置默认工作流。

当用户的输入无法匹配您配置的任何一个子工作流所定义的业务意图时，系统将自动执行此处指定的默认工作流，并随即结束当前任务。

未配置时，当遇到无法识别的用户问题时，系统将无响应或报错终止，影响用户体验。

配置默认工作流的操作如下：

1. 在“默认工作流（可选）”右侧，单击 ，在“添加工作流”页面选择作为默认的工作流应用。

 版权所有 © 华为云计算技术有限公司

2. 单击“确定”。

   如果选择的工作流需要删除，单击 。

3. 设置工作流的执行动作。

   支持的执行动作如下：

   - **继续**：按该工作流的执行结果，继续执行其他的子工作流。
   - **终止**：按该工作流的执行结果，调用结束工作流结束任务。
   - **等待输入**：按该工作流的执行结果，待用户输入问题后执行任务。

步骤9 （可选）配置结束工作流。

在所有子工作流执行完毕后固定执行的工作流。无论之前的意图路由经过了哪些子工作流，结束工作流始终最后执行。

适用于会话总结、满意度调查、数据清理等收尾操作。

执行顺序：起始工作流 → 子工作流（按意图路由）→ 结束工作流

- 配置结束工作流的操作如下：

  a. 在“结束工作流（可选）”右侧，单击 ，在“添加工作流”页面选择作为结束的工作流应用。

  b. 单击“确定”。

     如果选择的工作流需要删除，单击 。

步骤10 配置最大对话历史轮次。

编排模型在进行意图识别时，最多参考的历史对话轮数。历史对话轮次越多，意图识别时消耗的Token数越多，产生的费用越高。请根据业务复杂度合理设置。设为0表示不引用历史对话，仅基于当前输入判断意图。

设置历史对话次数，选择N，记录最近N条会话内容。例如，选择10，记录最近10条会话内容。取值范围0~100，默认值为10。

步骤11 配置最大跳转次数。

在一次用户会话中，编排控制器最多允许在子工作流之间跳转的次数。**防止因意图识别不稳定导致工作流之间无限循环跳转**。达到最大跳转次数后，系统将执行默认工作流并结束任务。

只有业务工作流之间跳转才会计算次数，起始工作流、结束工作流不计算跳转次数。取值范围0~30，默认值为9。

例如，一个多智能体应用含5个工作流，分别为工作流A（起始工作流或默认工作流）、工作流B、工作流C、工作流D、工作流E（结束工作流），根据用户问题先执行工作流A，根据工作流A的结果执行工作流B，根据工作流B的结果执行工作流C，再根据工作流C的结果执行工作流D，最后执行工作流E，相当于跳转了3次。

步骤12 配置完成后，单击“确定”。

设置后，进入“多智能体配置”页面。

- 在“多智能体配置”页面，显示多Agent控制器及添加的工作流、智能体及对应的工作流。

  单击工作流卡片，用户可以自定义意图名称、意图描述。

 版权所有 © 华为云计算技术有限公司

单击智能体卡片，用户可以自定义意图名称、意图描述。

单击智能体对应的工作流卡片，显示意图名称、意图描述、输入输出信息。

在工作流或智能体卡片上，单击“ ··· > 工作流详情”或“ ··· > 智能体详情”，查看工作流或智能体详情。

在工作流或智能体卡片上，单击“ ··· > 复制工作流ID”或“ ··· > 复制智能体ID”，复制工作流或智能体ID。

- 在“多智能体配置”页面，可以调试、发布多智能体应用，调试与发布多智能体应用请参考**调试多智能体应用**、**部署多智能体应用为API**。

----结束

意图识别输出结果说明

意图识别工作流的结束节点输出结果是由意图识别工作流产生，输出结果如下所示。**表5-16**所示的意图映射关系列表由多智能体的配置自动生成，从序号1开始递增。

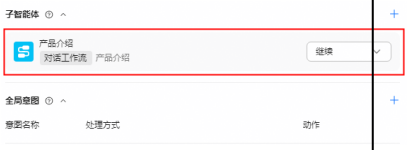
输出结果示例：

```
{
   "intent_id": 5
}
```

表 5-16 输出结果分类

任务的序号生成规则	说明	示例
1	系统默认意图（不明意图）。当用户输入无法匹配任何已知配置，指向此分类。 当intent_id值为1时，不执行任何任务，输出提示“抱歉，未识别到您的意图，请重新输入”。	-
2~1+N	全局意图。如果配置了全局意图，intent_id值在此范围内，则会执行对应的全局意图。 N为全局意图从上往下的序号。	• 如果配置1个全局意图，当intent_id值为2时，执行这个全局意图。  • 如果配置2个全局意图，当intent_id值为3时，执行第2个全局意图。 

 版权所有 © 华为云计算技术有限公司

任务的序号生成规则	说明	示例
（1+N）+1~1+N+M	子工作流意图。如果配置了子工作流，intent_id值在此范围内，则会执行对应的子工作流。 N为全局意图从上往下的序号。 M为子工作流从上往下的序号。	• 如果不配置全局意图，配置1个子工作流，当intent_id值为2时，执行这个子工作流。  • 如果不配置全局意图，配置2个子工作流，当intent_id值为3时，执行第2个子工作流。  • 如果配置1个全局意图、1个子工作流，当intent_id值为3时，执行这个子工作流。  • 如果配置1个全局意图、2个子工作流，当intent_id值为3时，执行第2个子工作流。 

版权所有 © 华为云计算技术有限公司

任务的序号生成规则	说明	示例
（1+N+M）+1~1+N+M+K	子智能体意图。如果配置了子智能体，intent_id值在此范围内，则会执行对应的子智能体。 N为全局意图从上往下的序号。 M为子工作流从上往下的序号。 K为子智能体从上往下的序号。	• 如果不配置全局意图、子工作流，配置1个子智能体，当intent_id值为2时，执行这个子智能体。  • 如果不配置全局意图、子工作流，配置2个子智能体，当intent_id值为3时，执行第2个子智能体。  • 如果不配置全局意图，配置1个子工作流、1个子智能体，当intent_id值为3时，执行这个子智能体。  • 如果不配置全局意图，配置1个子工作流、2个子智能体，当intent_id值为4时，执行第2个子智能体。

版权所有 © 华为云计算技术有限公司

任务的序号生成规则	说明	示例
		 • 如果配置1个全局意图、1个子工作流、1个子智能体，当intent_id值为4时，执行这个子智能体。  • 如果配置1个全局意图、1个子工作流、2个子智能体，当intent_id值为5时，执行第2个子智能体。 

相关操作

在“多智能体配置”页面，对画布的操作如下。

 版权所有 © 华为云计算技术有限公司

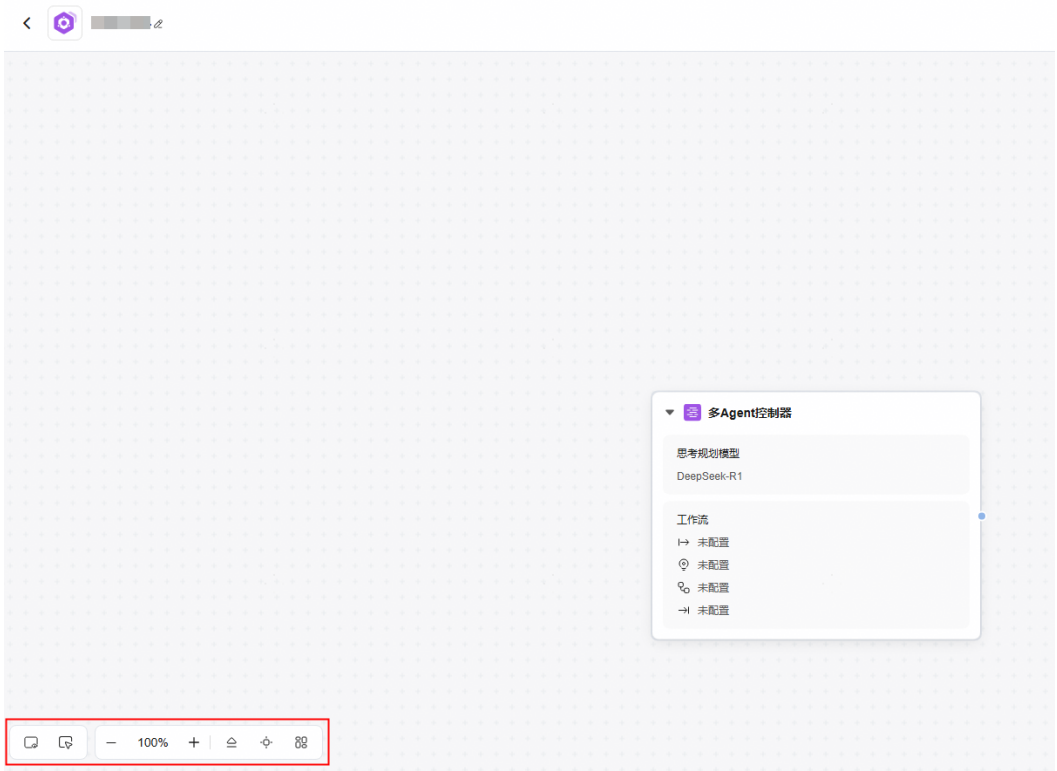


表 5-17 画布操作

图标	说明
	显示/隐藏缩略图。
	查看画布节点。 || — 100% +	缩小/放大画布内容。
/	全局折叠/展开节点。
	画布内容居中显示。
	画布内容布局优化。

常见问题

- **Q1：意图识别准确率低，用户意图频繁路由错误，怎么办?**
  排查步骤：
  a. 检查“子工作流执行逻辑提示词”是否清晰描述了各意图的区分边界。
  b. 如果使用大模型识别模式，尝试切换至“高精识别模式”并配置意图包。
  c. 增加每个子工作流卡片上的意图描述，提供更多示例语句。
  d. 适当降低最大对话历史轮次，减少历史干扰。

版权所有 © 华为云计算技术有限公司

- **Q2：配置了默认工作流，但用户无意图时系统仍然无响应?**

  检查默认工作流的执行动作是否设置为“终止”，如果设为“继续”，可能导致流程未正常结束。

- **Q3：intent_id计算后与实际路由不符?**

  重新核对配置中全局意图、子工作流、子智能体的排列顺序。intent_id严格按照界面从上到下的顺序生成，调整顺序后映射关系随之变化。

- **Q4：工作流一直在子工作流之间循环跳转?**

  降低最大跳转次数，并检查子工作流的执行动作设置，确认不需要继续流转的工作流已设置为“终止”。

5.4 调试多智能体应用

开发者可以在多智能体应用创建完成后，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。

创建多智能体应用后，AgentArts支持对应用进行预览与调试。试运行时，系统将按照多智能体的正式配置逻辑执行请求，通过中央控制器进行意图识别与路由分发，调用对应的子智能体或工作流。调试过程中产生的调用链信息与正式运行一致。

说明

典型场景：多智能体应用配置完成后，在正式发布前，通过试运行验证中央控制器的意图识别是否准确、路由分发是否正确、各子智能体/工作流的响应是否符合预期。

费用说明

调试智能体会产生tokens消耗。

- 如果使用平台预置的MaaS模型即服务的模型，优先使用平台赠送的2,000,000tokens，使用完毕后，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

注意

多智能体应用的一次试运行可能触发多次模型调用（包括中央控制器的意图识别、各子智能体/工作流的独立调用），tokens消耗可能高于单智能体调试。建议关注用量，避免频繁试运行导致费用快速累积。

前提条件

- 已**开通AgentArts服务**。
- 已**创建并配置多智能体应用**。

调试多智能体应用

步骤1 登录**AgentArts智能体平台**。

版权所有 © 华为云计算技术有限公司

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击待调试的多智能体应用卡片。

步骤4 在“多智能体配置”页面，单击“试运行”。

步骤5 在“试运行配置”页面，输入试运行配置，单击“开始运行”。

说明

仅当多智能体应用的全局配置中设置了输入参数，或所使用的工作流应用的开始节点设置了输入参数时，才会显示“试运行配置”页面。

支持手动输入初始参数，也支持导入初始参数。单击 ，在弹框中粘贴或输入JSON格式的数据，默认勾选“导入默认值”，单击“确定”完成导入。

步骤6 在“试运行”页面，在对话输入框中输入内容并按Enter键或单击 ，可与智能体进行交互。您可以依据应用的执行过程与响应结果，对“多Agent控制器”的参数配置进行调优。

- 如果需要开启新会话，单击界面右上角 ，清除本次会话内容。
- 如果需修改运行配置，单击界面右上角 ，进入“试运行配置”重新设置相关参数值。

  说明

  仅当多智能体应用的全局配置中设置了输入参数，或所使用的工作流应用的开始节点设置了输入参数时，才会显示此图标。

试运行成功时，界面将显示智能体的回复内容。如果多智能体的中央控制器正确识别了意图并路由到对应子智能体，回复内容应与预期子智能体的能力匹配。

如果试运行失败，常见报错与解决方案请详见**开发常见报错**。

 版权所有 © 华为云计算技术有限公司

图 5-15 试运行



步骤7 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，如图5-16所示。

版权所有 © 华为云计算技术有限公司

- **运行结果**：展示应用的执行开始时间、结束时间、运行时间等信息，以及输入和输出信息。重点关注执行时间是否在预期范围内、输入输出是否正确，直观了解应用的性能情况。
- **调用详情**：展示调用链中的控制器或工作流的详细信息，包括运行的控制器或工作流、各环节耗时、输入和输出信息等。重点关注路由是否按预期分发、各环节耗时是否正常、输入输出是否传递正确，便于开发者快速梳理调用顺序并精确定位问题。

  单击“查看子工作流调用链”，可以查看工作流中每个节点的详细信息，包括运行的节点、节点耗时、节点的输入输出。

图 5-16 调试结果示例



----结束

5.5 部署多智能体应用为 API

虽然多智能体应用已试运行成功，但目前仍局限于开发环境，无法直接对外提供服务。为了实现应用的落地与对接，用户提交版本并完成部署后，即可通过API接口进行调用。

说明

提交版本是将当前多智能体应用配置保存为一个不可变的版本快照；部署是将指定版本发布为可调用的API接口，创建一个独立的运行实例。

典型场景：多智能体应用试运行验证通过后，需要将其集成到企业的客服系统或业务后台中。通过提交版本并部署为API，获取API地址和认证信息，供业务系统调用。

费用说明

智能体发布后通过API接口调用时，会产生tokens消耗。

- 如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

版权所有 © 华为云计算技术有限公司

> **注意**
>
> 智能体部署并通过API调用后，除模型tokens消耗外，如果开启了日志记录（LTS）、调用链（APM）、指标（AOM）功能，以及使用了API Key认证（DEW凭据托管），均会产生按需计费费用。建议在部署前了解各项服务的计费规则，合理控制成本。

前提条件

- 已**开通AgentArts服务**。
- 已**创建并配置多智能体应用**。

提交版本

将多智能体部署为API前，需要先提交版本，部署时是按版本进行部署。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击待提交版本的多智能体应用卡片。

步骤4 在“多智能体配置”页面右上角，单击“提交版本”。

> **说明**
>
> 已经提交版本的多智能体应用，修改后再次发布，显示为“更新版本”。

步骤5 在“提交版本”页面，配置发布信息，具体参数说明请参考**表5-18**。

图 5-17 提交版本



版权所有 © 华为云计算技术有限公司

表 5-18 发布多智能体参数说明

参数	说明
版本名称	系统自动生成带年月日的版本名称，以v开头。也可以自定义版本名称，由1~64个字符组成。
版本描述（非必填）	多智能体的描述信息。由0~1024个字符组成。
部署至实例	默认勾选。 将该版本同步部署为API接口，便于后续接口调用。

步骤6 单击“确定”。

提交成功后，在“多智能体”页面的卡片上，显示“已提交”。您可在版本历史记录中查看已提交的版本。

如果勾选“部署至实例”，进入“部署智能体”页面，具体部署操作请参考**部署智能体**。

步骤7 （可选）查看API调用接口信息。

1. 在“版本提交”页面，单击“分享”，进入“渠道管理”页面。
2. 在“渠道管理”中单击“查看API”即可查看API调用接口信息。

说明

如果未部署过任何版本，“查看API”页面中的API详情中的URL为空。请先完成部署后再查看。

----结束

部署智能体

未部署的智能体，不能通过API接口调用，部署是以版本为单位进行部署。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击待部署的多智能体应用卡片。

步骤4 在“多智能体配置”页面右上角，单击  。

步骤5 在“部署智能体”页面，配置部署信息如**图5-18**所示，参数配置请参考**表5-19**。

版权所有 © 华为云计算技术有限公司

图 5-18 部署智能体



表 5-19 部署智能体

参数	说明
智能体名称	为**创建并配置多智能体应用**自定义的智能体名称，不可修改。
智能体版本	为**提交版本**的版本名称。默认为最新版本名称，可以在下拉框中切换。切换版本后需重新部署。
运行时名称	运行时的名称，系统默认，**仅第一次部署时支持自定义修改。** 由2~48个字符组成，包含小写字母、数字、中划线，以小写字母开头，以小写字母或数字结尾。

版权所有 © 华为云计算技术有限公司

参数	说明
入站身份认证	选择调用此API接口时，使用的认证鉴权方式，如何进行认证鉴权请参考**认证鉴权**。默认IAM认证，**仅第一次部署时支持选择。** • IAM认证：将华为云IAM身份配置为入站身份认证，以验证调用者的身份。适用于华为云内部系统或已接入华为云IAM的系统，集成复杂度低，安全等级高。 • OAuth 2.0认证：将JWT（如OAuth令牌）配置为入站身份认证，以验证传入的令牌签名和范围。适用于已有身份提供商（如Okta、Cognito等）的企业系统，集成复杂度高，安全等级高。 – Discovery URL：身份提供商（例如Okta、Cognito等）提供的Discovery URL，该URL通常可在该提供商的文档中找到。 – JWT授权配置：配置JWT令牌的访问控制规则。 ▪ 允许的受众：指定的受众是否在Agent Identity运行中指定的受众匹配或为其子集。 ▪ 允许的客户端：指定的客户端标识符是否被允许访问Agent Identity。 ▪ 允许的范围：仅当令牌包含此处配置的至少一个必须范围时，才允许访问自定义声明。 ▪ 自定义声明：仅当令牌中的特定声明与预定义字符串值相匹配时，才允许访问。 • API Key认证：将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。适用于内部测试或轻量级集成，集成复杂度低，安全等级中。 **注意** 身份认证凭证在凭据管理服务（DEW）托管，托管按需计费，使用DEW产生的费用请参考**DEW计费说明**。
日志记录	默认开启。 开启日志记录功能后，智能体产生的日志会上报到云日志服务 LTS，便于日志的集中管理和分析。日志管理费用按需收取，具体计费详情请参考**LTS计费说明**。 进入“运营运维 > 观测 > 调用链分析”的详情页，在“日志”页签中查看详细的日志信息。
指标	默认开启。 开启指标功能后，智能体产生的指标数据会上报到应用运维管理 AOM，便于实时监控和分析应用的运行状态。指标管理费用按需收取，具体计费详情请参考**AOM计费说明**。 进入“运营运维 > 观测 > 调用链分析”的详情页，在“指标”页签中查看详细的指标信息。

版权所有 © 华为云计算技术有限公司

参数	说明
调用链	默认开启。 开启调用链功能后，智能体产生的调用链数据会上报到应用性能管理 APM，帮助开发者追踪和分析应用的性能问题。调用链管理费用按需收取，具体计费详情请参考**APM计费说明**。 进入“运营运维 > 观测 > 调用链分析”的详情页，在“链路信息”页签中查看详细的调用链信息。
智能体卫士	默认关闭。 开启后，实时检测智能体执行链路与操作轨迹，监控全量行为并识别安全异常。不同安全风险处理规则和检测记录可前往**企业主机安全 HSS**管理及查看。HSS的智能体卫士当前为公测阶段，暂不收费。

步骤6 单击“确定”。

部署成功后，可以在“部署运行 > 智能体运行时”页面，搜索运行时名称查看部署的智能体。

多智能体部署后，支持通过API调用，具体请参考**通过API调用多智能体应用**。

步骤7 （可选）查看API调用接口信息。

1. 在“发布部署”页面，单击“分享”，进入“渠道管理”页面。
2. 在“渠道管理”中单击“查看API”即可查看API调用接口信息。

----结束

常见问题

部署智能体时，入站身份认证选择错误，如何处理?

答：部署智能体时，入站身份认证选择后不支持修改，选择的入站身份认证在通过API接口调用此智能体时作为身份校验使用。如果选择错误，建议在智能体卡片上单击“复制”复制一个智能体，重新进行部署。

相关文档

多智能体应用发布后，可以通过API接口调用，请参考**通过API调用多智能体应用**。

5.6 通过 API 调用多智能体应用

多智能体应用部署为API后，您可以通过标准HTTP请求调用智能体，将其集成到企业业务系统、客服平台、聊天机器人等场景中。API支持流式（实时返回生成内容）和非流式（等待全部内容生成后返回）两种响应模式，满足不同的实时性需求。

典型场景：将智能客服多智能体集成到银行APP中。用户发送问题后，系统自动调用多智能体API，中央控制器识别意图并路由至对应的子智能体（如“账户查询”或“理财推荐”），最终将回复返回给APP前端。

通过API调用多智能体的优势

版权所有 © 华为云计算技术有限公司

- **提升开发效率**：API接口让不同应用程序高效交互，大幅降低数据传输与处理的成本，帮助开发者将更多精力集中在核心业务上。
- **扩展应用范围**：通过API接口实现不同系统、平台和服务之间的交互，促进多平台资源共享，帮助开发者创新更多应用，增强企业竞争力。

费用说明

智能体发布后通过API接口调用时，会产生tokens消耗。

- 如果需要使用平台预置的MaaS模型即服务的模型，需要先开通模型并设置鉴权。具体操作请参考**接入预置的付费模型**，计费规则请参考**MaaS模型即服务价格**。
- 如果使用自定义接入的第三方模型，其收费规则请参考该第三方模型的官方说明。

⚠ 注意

智能体部署并通过API调用后，除模型tokens消耗外，如果开启了日志记录（LTS）、调用链（APM）、指标（AOM）功能，以及使用了API Key认证（DEW凭据托管），均会产生按需计费费用。建议在部署前了解各项服务的计费规则，合理控制成本。

前提条件

- 已**开通AgentArts服务**。
- 在调用应用前，须确保应用已发布，具体请参考**部署智能体**。

获取调用路径

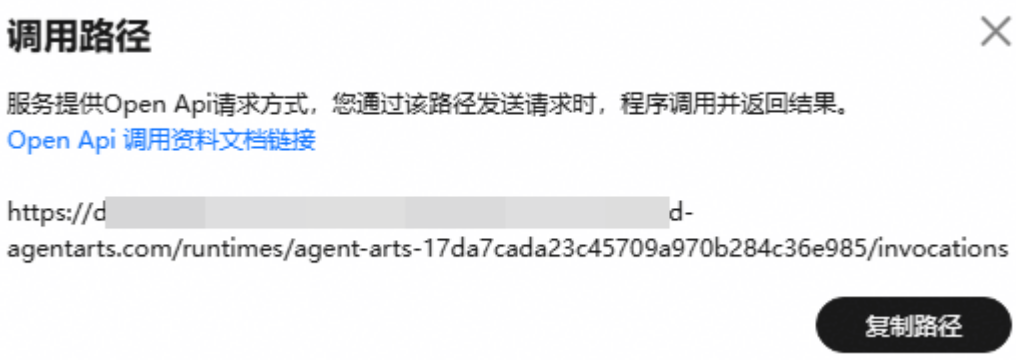
步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，鼠标移动至多智能体应用卡片上。

步骤4 单击“调用路径”，在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如**图5-19**所示。

图 5-19 获取应用调用路径



----结束

版权所有 © 华为云计算技术有限公司

使用 API 调用多智能体应用

使用API调用多智能体应用的操作，请参考**使用API调用单/多智能体**。

5.7 管理多智能体

费用说明

不同套餐包允许创建的**应用数量限制**不同，具体可参考**计费说明**。

导入导出多智能体应用

AgentArts支持将多智能体应用从一个环境中导出，导入至另一个环境中，无需用户重复构建，快速完成多智能体应用跨环境构建或复用。

适用的业务场景如下：

- 从测试环境导出多智能体，导入到生产环境中部署。
- 在不同的开发环境之间迁移多智能体应用。
- 将多智能体应用下载到本地进行代码归档。
- 作为模板提供给其他客户进行复用。

导出多智能体应用

已**创建并配置多智能体应用**或已**导入多智能体应用**。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击“导出”，勾选待导出的多智能体应用左侧复选框，单击“导出”；或将鼠标移动至目标智能体，单击“更多 > 导出”。

系统弹出“导出”弹框，如**图5-20**所示，“结果列”显示“成功”表示导出成功。

版权所有 © 华为云计算技术有限公司

图 5-20 导出多智能体应用



步骤4 单击浏览器右上角的下载图标，可查看导出文件。

导出文件为JSONL格式。建议下载后检查文件大小是否合理，并确认文件能否正常打开。

说明

- **包含内容**：导出的文件包含多智能体应用的**配置信息**及其**关联资源**（如插件、MCP服务、工作流、子智能体的配置）。导入时，系统将自动尝试还原这些依赖资源及其关联关系。
- **不包含内容**：导出内容**不包含知识库**。请在导入应用后，手动迁移知识库并重新进行关联。
- **导出失败**：如果应用关联的某个特定资源导出失败，导出的文件中将不包含该资源，但应用的其他部分仍会正常导出。失败原因可在应用卡片的操作列中查看。

----结束

导入多智能体应用

已有从其他AgentArts环境导出的多智能体应用的JSONL格式文件。

版权所有 © 华为云计算技术有限公司

警告

如果导入的多智能体应用名称与当前环境中已有的应用名称相同，导入后将**覆盖**原有多智能体应用的配置，且**不可恢复**。

建议在导入前确认名称是否冲突，或先导出备份原有应用。

如果当前环境的应用数量已接近套餐上限，请先清理无用应用或升级套餐后再执行导入。

步骤1 登录AgentArts智能体平台。

步骤2 在左侧导航栏，选择“开发中心 > 智能体管理”。

步骤3 选择“多智能体”页签，单击“导入”。

步骤4 在“导入”页面，单击“选择文件”，选择本地已准备好的文件，勾选需要导入的内容，单击“导入”。

步骤5 导入后，可在图5-21所示的结果列查看状态。

- **导入成功**：结果列显示“成功”。可进入多智能体配置页面，检查各项配置是否已正确还原。
- **需配置鉴权**：如果关联的插件、MCP服务或模型供应商缺乏认证信息，操作列将显示“配置鉴权”。请单击“配置鉴权”，在弹窗中完成鉴权配置。
- **导入失败**：如果应用关联的某个特定资源（如插件、MCP服务、知识库、工作流等）导入失败，应用本身仍会导入成功，但该资源将不会被导入。失败原因可在操作列中查看。

图 5-21 导入多智能体应用



 版权所有 © 华为云计算技术有限公司

图 5-22 导入多智能体成功



----结束

多智能体相关操作

在多智能体应用卡片列表中，支持的其他操作请参考**表5-20**。

表 5-20 相关操作

操作	说明
编辑多智能体应用信息	单击待编辑的多智能体应用卡片，进入“多智能体配置”页面，在名称右侧单击，可以编辑多智能体应用的名称、描述、图标。 如果配置的子智能体有新版本，在对应的卡片上单击，升级新版本。
渠道管理	鼠标移动至已发布的多智能体应用卡片上，单击“渠道管理”，进入渠道管理页面，具体发布操作请参见**部署多智能体应用为API**。
查看版本历史记录	鼠标移动至已发布的多智能体应用卡片上，单击，查看版本历史记录。 在版本历史记录中，可以执行如下操作。 • 复制版本：在“版本ID”右侧，单击，可以复制发布的版本ID。 • 还原版本：单击“还原版本”，可以还原到此版本。 **注意** 还原后，当前多智能体应用配置将不再保留，建议还原前先提交当前版本作为备份。 • 删除版本：单击“删除”，可以删除此发布版本。

版权所有 © 华为云计算技术有限公司

操作	说明
获取多智能体应用的调用路径	鼠标移动至待获取调用路径的多智能体应用卡片上，单击“调用路径”，在“调用路径”页面，单击“复制路径”。
复制多智能体应用的ID	鼠标移动至待复制ID的多智能体应用卡片上，单击“复制ID”。
删除多智能体应用	**警告** 如果该多智能体已被引用，删除时将展示引用关系。删除会导致引用方运行异常，且不可恢复。 鼠标移动至待删除的多智能体应用卡片上，单击“更多 > 删除”。
查看智能体运行时数据	• 在“智能体运行时”页面，通过搜索智能体ID搜索查看智能体运行时数据。 1. 登录**AgentArts智能体平台**。 2. 在左侧导航选择“部署运行 > 智能体运行时”，在搜索框中输入上方已获取的ID进行搜索。 3. 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。 • 在“运营运维 > 观测 > 智能体运行分析”页面，通过搜索智能体ID查看智能体运行时数据。 1. 登录**AgentArts智能体平台**。 2. 在左侧导航选择“运营运维 > 观测 > 智能体运行分析”，在搜索框中输入上方已获取的ID进行搜索。 3. 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。

常见问题

- **Q1：导入显示成功，但应用运行时报错，如何排查?**

  导入操作仅还原了应用配置结构与关联关系，实际运行还需目标环境中的依赖资源就绪。请按以下顺序检查：

  a. 鉴权配置：进入配置页，检查关联的插件、MCP服务、模型供应商是否提示“未配置鉴权”。如有，请重新输入有效的API Key。

  b. 知识库关联：导出的文件不包含知识库。请确认目标环境中已存在对应的知识库，并在“知识库”组件中重新选择绑定。

  c. 依赖状态：资产广场的资源无法导出和导入，需要检查导入环境是否需要重新关联。

  d. 版本差异：确认导入的模型类型在目标环境中是否可用。

- **Q2：导入/导出结果列表中，某些应用显示“失败”或“部分失败”，如何处理?**

  - **导出失败**：通常是因为原环境中的关联资源已被删除、权限过期或版本不兼容。请单击“操作列”查看具体失败原因，修复原应用资源后再次尝试导出。
  - **导入失败**：通常是因为导入文件中的某些组件（如特定插件）在当前环境中不存在。应用主体仍会导入成功，缺失的资源需在导入后手动重新配置。

版权所有 © 华为云计算技术有限公司

- **Q3：为什么导入后的插件、MCP显示“未配置鉴权”？**

  这是出于安全与跨环境隔离的设计。导出的JSONL文件默认不包含API Key等敏感鉴权信息，以防止凭据明文泄露。此外，不同环境的密钥相互独立，无法直接复用。

  处理方式：导入成功后，请进入多智能体编辑页面，找到对应的插件或MCP，手动配置目标环境对应的鉴权信息。

 版权所有 © 华为云计算技术有限公司

6 组件库

6.1 插件

6.1.1 插件介绍

什么是插件

在Agent（智能体）中，**插件**是大语言模型（LLM）与外部世界交互的一系列工具。

尽管LLM具备强大的自然语言处理与推理能力，但其本质是基于训练数据的静态系统，存在两大局限：

- **信息滞后：**无法获取训练截止日期之后的实时信息。
- **行动力缺失：**无法直接与现实世界的软件或硬件系统交互（如读写数据库、发送请求）。

插件的引入，使Agent具备了实时信息检索、外部任务执行的能力，是Agent从“对话系统”向“任务解决系统”演进的关键模块。

AgentArts智能体平台提供了丰富的预置插件，包括天气查询、热搜查询、文档生成、出行旅游等。例如，在您的Agent中添加天气查询插件，即可获取实时天气信息。

当官方插件无法满足特定业务需求时，您可以自主创建插件。

插件分类、形态

AgentArts智能体平台提供丰富的插件资源，除预置的官方插件外，还支持用户自定义插件，可将API能力通过插件形式引入Agent。

版权所有 © 华为云计算技术有限公司

表 6-1 插件分类、形态

插件分类	说明
官方预置插件	由AgentArts智能体平台预置到“资产广场”的官方插件，由AgentArts平台提供技术支持和维护，平台负责托管这些插件的后端逻辑并为您处理鉴权。您只需关注如何调用能力，无需关心底层实现。 官方预置插件分为无需鉴权和需鉴权两类： ● 无需鉴权：在Agent中添加插件即可使用。 ● 需鉴权：填写API Key鉴权信息完成认证后即可使用。 **选型建议**：如果您需要的是**通用能力**（如天气查询、搜索等），且希望**零代码、快速接入**，请优先选择官方预置插件。
自定义-API类型插件	API类型插件是将现有的RESTful API（HTTP/HTTPS接口）封装为Agent可调用的工具。它可以让Agent把理解到的指令转化为实际的API请求，直接调用外部系统完成任务。 ● 运行位置：实际的业务逻辑在外部API服务器（如企业的后端服务器、第三方功能平台）上运行。Agent充当客户端发起请求，并接收API返回的响应。 ● 适用场景： – 数据检索：从外部数据库或服务获取实时数据（如天气查询、网络搜索、新闻查询）。例如，在“出行助手”智能体中，通过天气查询插件获取目的地实时天气，为用户提供穿衣建议。 – 状态变更：在外部系统中执行提交、变更类操作。例如，在“运维助手”智能体中，通过自定义API插件对接内部工单系统，根据用户描述自动创建维修工单。 – 复杂业务逻辑处理：利用已有的成熟后端服务处理复杂的业务流程。 **选型建议**：如果您需要对接**企业内部系统**（如ERP、CRM），或者需要实现**特殊的业务逻辑**，官方插件无法满足需求时，请使用自定义API插件。此方式需要您自行开发并维护后端API服务。

插件与工具

在AgentArts智能体平台中，理解插件与工具的从属关系是成功创建插件的前提，可通过以下“文字识别”插件示例直观理解。

表 6-2 文字识别插件示例

插件名称	工具名称	API接口地址
文字识别插件	通用文字识别	https://{endpoint}/v2/{project_id}/ocr/**general-text**
	通用表格识别	https://{endpoint}/v2/{project_id}/ocr/**general-table**

版权所有 © 华为云计算技术有限公司

插件名称	工具名称	API接口地址
	手写文字识别	https://{endpoint}/v2/{project_id}/ocr/**handwriting**

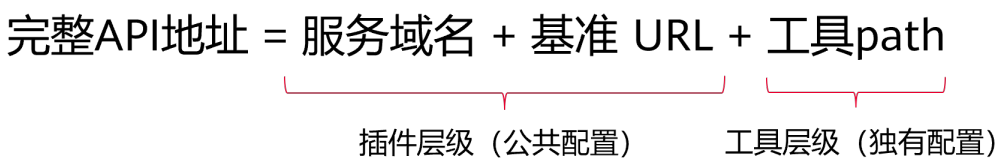
插件：是集成了一个或者多个工具的功能集合，它定义了这些工具共用的基础属性（如服务域名、鉴权方式）。

工具：是插件内具体的执行单元，每个工具对应一个独立的功能，负责完成一项具体的单一任务。

AgentArts智能体平台支持自定义插件的开发与创建。在配置规则中，要求同一插件下的工具需要具备同类功能。

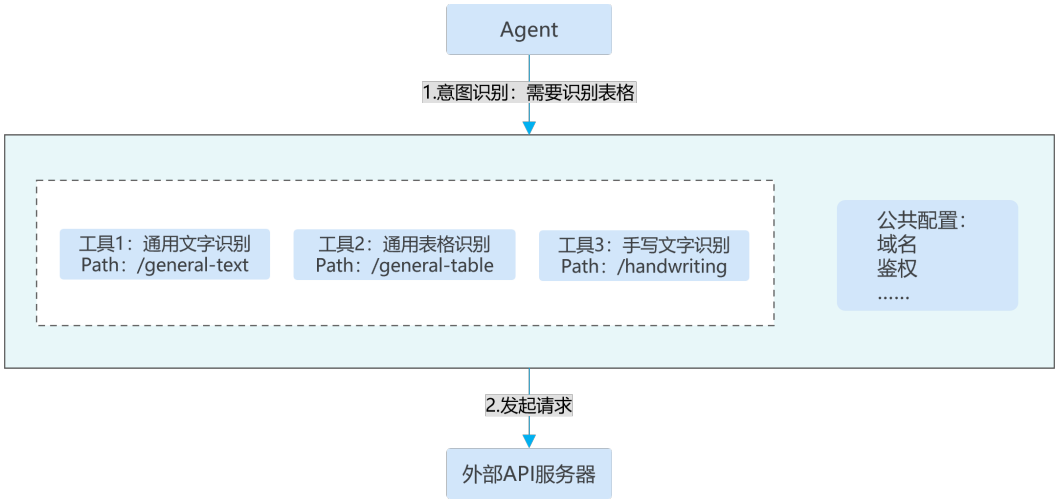
以表6-2为例，https://{endpoint}/v2/{project_id}/ocr是所有接口的公共基础路径，它定义了插件的连接目标（由服务域名+基准URL构成）；/**general-text**这类后缀是各工具独有的“路径”，它区分了插件内的不同能力。

图 6-1 API 与插件关系示例



在执行层面，Agent并不直接“运行插件”，而是调用插件内某一特定工具。当Agent需要执行“表格图片提取文字”任务时，它会定位到“文字识别插件”，然后精准地调用其中的“通用表格识别工具”，这本质上是向.../ocr/general-table接口发起了一次请求。

图 6-2 插件与工具关系示例



版权所有 © 华为云计算技术有限公司

6.1.2 示例：创建联网搜索插件

本示例将介绍如何通过AgentArts智能体平台创建联网搜索插件，并进行插件的调试、发布，最终应用到Agent中。

前提条件

已**开通AgentArts服务**。

准备联网搜索 API

本示例使用联网搜索API实现联网搜索功能，其API接口调用示例如下。请提前**创建API Key**，获取到的API Key用于替换下方代码中的<AppBuilder API Key>。

调用该接口可能会产生第三方API调用费用，建议在使用前了解相关费率。

```
curl --location --request POST 'https://qianfan.baidubce.com/v2/ai_search/web_search' \
--header 'Authorization: Bearer <AppBuilder API Key>' \
--header 'Content-Type: application/json' \
--data-raw '{
  "messages": [
    {
      "role": "user",
      "content": "什么是Agent"
    }
  ]
}'
```

步骤 1：创建插件

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，在“插件”页签，单击页面右上角“创建插件”。

步骤3 根据以下步骤配置插件信息。

版权所有 © 华为云计算技术有限公司

图 6-3 创建插件

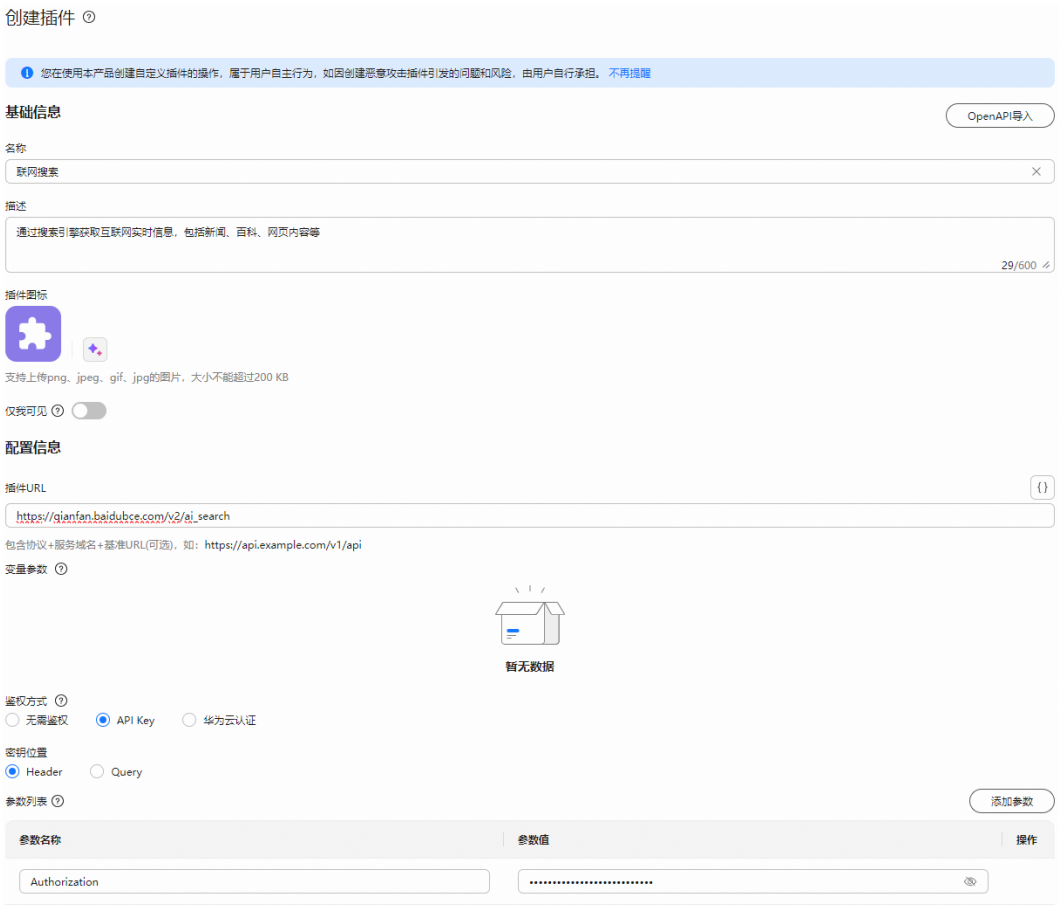
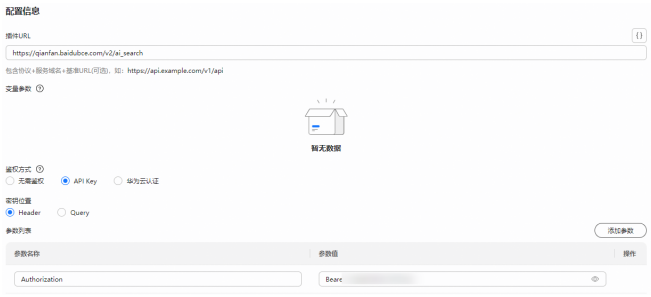


表 6-3 基本信息

参数		填写说明
基础信息	OpenAPI导入	不使用
	名称	填写“联网搜索”。
	描述	填写“通过搜索引擎获取互联网实时信息，包括新闻、百科、网页内容等”。
	插件图标	使用默认。
	仅我可见	关闭

版权所有 © 华为云计算技术有限公司

参数		填写说明
配置信息	插件URL	填写“https://qianfan.baidubce.com/v2/ai_search”。 **图 6-4** 填写示例 
	鉴权方式	选择“API Key”。
	密钥位置	选择“Header”。
	参数列表	新增一参数，参数名称填写“Authorization”，参数值为Bearer <AppBuilder API Key>格式，<AppBuilder API Key>用**准备联网搜索API**中获取的API Key进行替换。

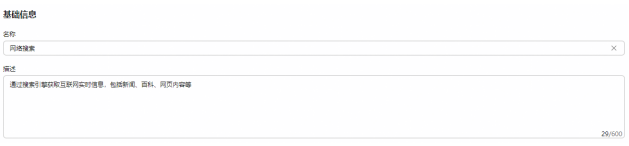
步骤4 单击“下一步 创建工具”，进入工具创建页面，或单击“完成插件创建”。如果选择后者，后续可继续在插件详情中添加工具。请参考**步骤2：创建工具**为插件添加工具。

----结束

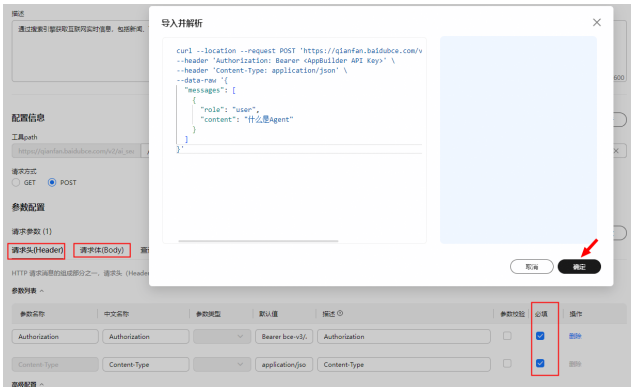
步骤 2：创建工具

步骤1 在“创建信息”页面中，请参照**表6-4**完成工具配置。

表 6-4 工具配置

参数		参数说明
基础信息	名称	填写“网络搜索”。
	描述	填写“通过搜索引擎获取互联网实时信息，包括新闻、百科、网页内容等”。 **图 6-5** 填写示例 

版权所有 © 华为云计算技术有限公司

参数		参数说明
配置信息	工具path 请求方式 请求参数 请求头(Header) 请求体(Body)	**使用平台提供的自动解析功能，无需手动填写参数。** 单击“导入并解析”，输入**准备联网搜索API**中的接口示例后，单击“确定”。注意示例的<AppBuilder API Key>请提前进行替换。 在请求头(Header)中将识别到的参数设置为“必填”。 请求体(Body)需要参考**图6-7**进行设置（如果自动解析结果与下图不符，请按照图片进行设置）。 **图 6-6** 请求头(Header)填写示例  **图 6-7** 请求体(Body)填写示例 
	响应参数	在下方参数列表中，添加一行参数。参数名称为“references”，描述填写“输出信息”，参数类型选择“Array<String>”。 **图 6-8** 填写示例 

步骤2 配置完单击“确定”。后续调试步骤请参考**步骤3：调试发布插件**。

----结束

版权所有 © 华为云计算技术有限公司

步骤 3：调试发布插件

工具创建完成后，工具列表会新增一行“待调试”状态的工具。工具只有调试后对应的插件才可以发布使用。

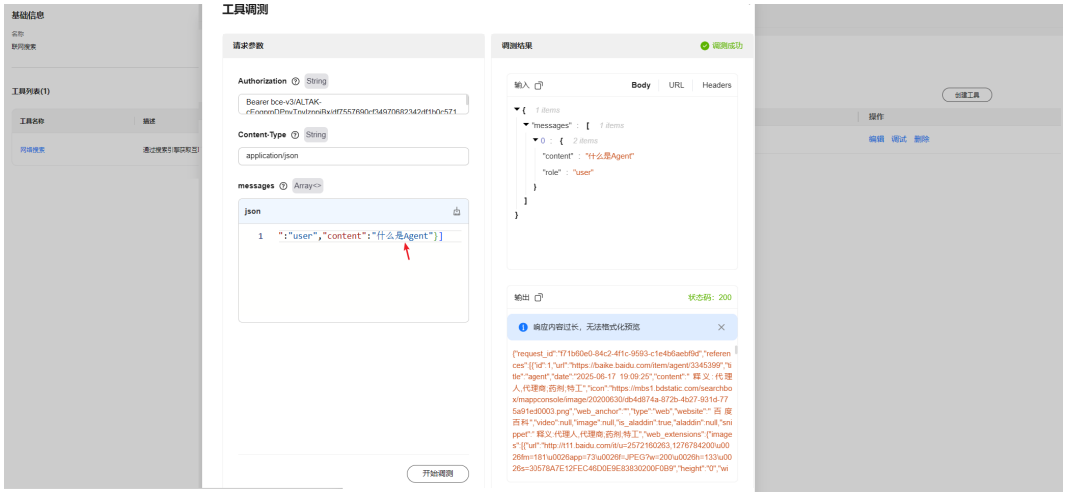
步骤1 在工具列表中，单击“调试”按钮，输入问题后，单击“开始调试”。

如果调试结果显示失败，请检查以下几点：

鉴权信息：确认Header中的Authorization参数是否正确填写了Bearer <AppBuilder API Key>格式，且Key有效。

参数匹配：确认请求参数是否与联网搜索API要求的一致。

图 6-9 输入问题、调试工具



步骤2 调试成功后，工具状态随之变为“成功”，此时可单击右上角“发布”，发布插件。

----结束

步骤 4：在单智能体中使用插件

插件创建完成后，它只是一个静态的资产。要使其真正发挥价值，需要将其挂载到一个具备思考能力的智能体上。

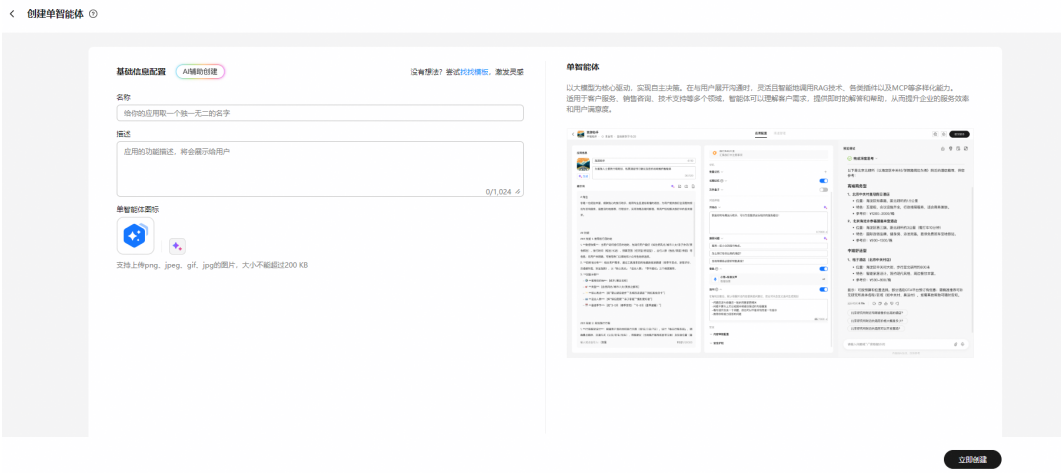
在本小节中，将创建一个最基础的单智能体应用。您无需编排复杂逻辑，只需简单配置，即可让智能体获得联网搜索的能力，体验它如何打破模型训练数据的时效限制，自主获取互联网上的最新信息。

步骤1 在“智能体管理 > 单智能体”页面，单击“创建智能体”。

步骤2 名称、描述可填写为“联网搜索智能体”，单击“立即创建”。

版权所有 © 华为云计算技术有限公司

图 6-10 创建单智能体应用



步骤3 在“单智能体配置”页面，根据以下信息进行配置。

表 6-5 单智能体配置信息

参数	填写说明
提示词	输入“联网搜索”后，使用AI功能智能优化提示词，系统会自动填入优化后的内容。 **图 6-11** 配置提示词 
模型	在模型配置区域，选择一个可用的大模型。平台提供200万免费Tokens，超出后调用大模型将产生费用，请根据您的业务需求选择合适的模型规格。 **图 6-12** 配置模型 

版权所有 © 华为云计算技术有限公司

参数	填写说明
插件	在配置页面中间，添加一个联网搜索插件。 **图 6-13** 配置插件 
其余参数	使用默认，不进行设置。

步骤4 在配置页面右侧，可以进行智能体调试。

在对话框中输入一个需要实时信息的问题（例如“最新的新闻有哪些”）。如果智能体返回的回答中包含具体的实时数据、网页链接或新闻摘要，说明联网搜索插件已成功挂载并生效。

----结束

6.1.3 创建插件

6.1.3.1 基于 API 创建插件

在AgentArts智能体平台中，当官方预置插件无法满足特定业务需求时，开发者需要自主创建插件。

前提条件

已**开通AgentArts服务**。

API 形态插件

API类型插件是将现有的RESTful API（HTTP/HTTPS接口）封装为Agent可调用的工具。它本身不承载业务逻辑，而是作为桥梁将指令“翻译”为API请求，交由外部系统执行逻辑并返回结果。

典型适用场景：

- 数据检索：如天气查询、网络搜索、新闻查询。
- 状态变更：如发送信息、触发创建任务。
- 复杂业务逻辑处理：售后退换货处理（API后台校验商品是否满足退货条件、换货库存是否充足、生成退换链接、发送通知消息等）。

不适用场景：纯逻辑计算（简单数值计算）、简单的字符串处理（延迟高且浪费网络资源）。

插件调用链路：

版权所有 © 华为云计算技术有限公司

1. 用户输入：帮我查一下订单12345的状态。
2. Agent分析：命中插件get_order_status。
3. 参数生成：Agent中模型生成结构化参数{ "order_id": "12345" }。
4. 请求构建：Agent组装API插件请求。
5. 网络交互：Agent向API服务器发起GET /api/orders/12345请求。
6. 响应解析：服务器返回JSON数据，Agent将其清洗后注入模型上下文。
7. 最终回复：模型根据数据生成自然语言：“订单12345当前状态为已发货...”。

约束限制

- **工具数量：**单个插件最多支持添加30个工具。
- **文件上传大小：**当请求体的数据格式是form-data时，限制如下：
  - 图片文件：单个文件大小不得超过5MB。
  - 其他文件：单个文件大小不得超过128MB。

步骤 1：创建插件

创建插件过程中，将结合“使用文字识别API创建插件”作为示例进行讲解，文字识别API、插件与工具说明请参考**插件介绍**。

说明

- 示例讲解使用“文字识别服务 > 通用文字识别API”，请提前在**文字识别服务控制台**开通该API。
- 本示例使用的华为云文字识别服务属于收费服务，调用该接口将产生OCR接口费用，具体请参考文字识别服务**计费详情**。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，在“插件”页签，单击页面右上角“创建插件”。

步骤3 在“创建插件”页面中，根据以下步骤配置插件信息。

版权所有 © 华为云计算技术有限公司

图 6-14 创建插件

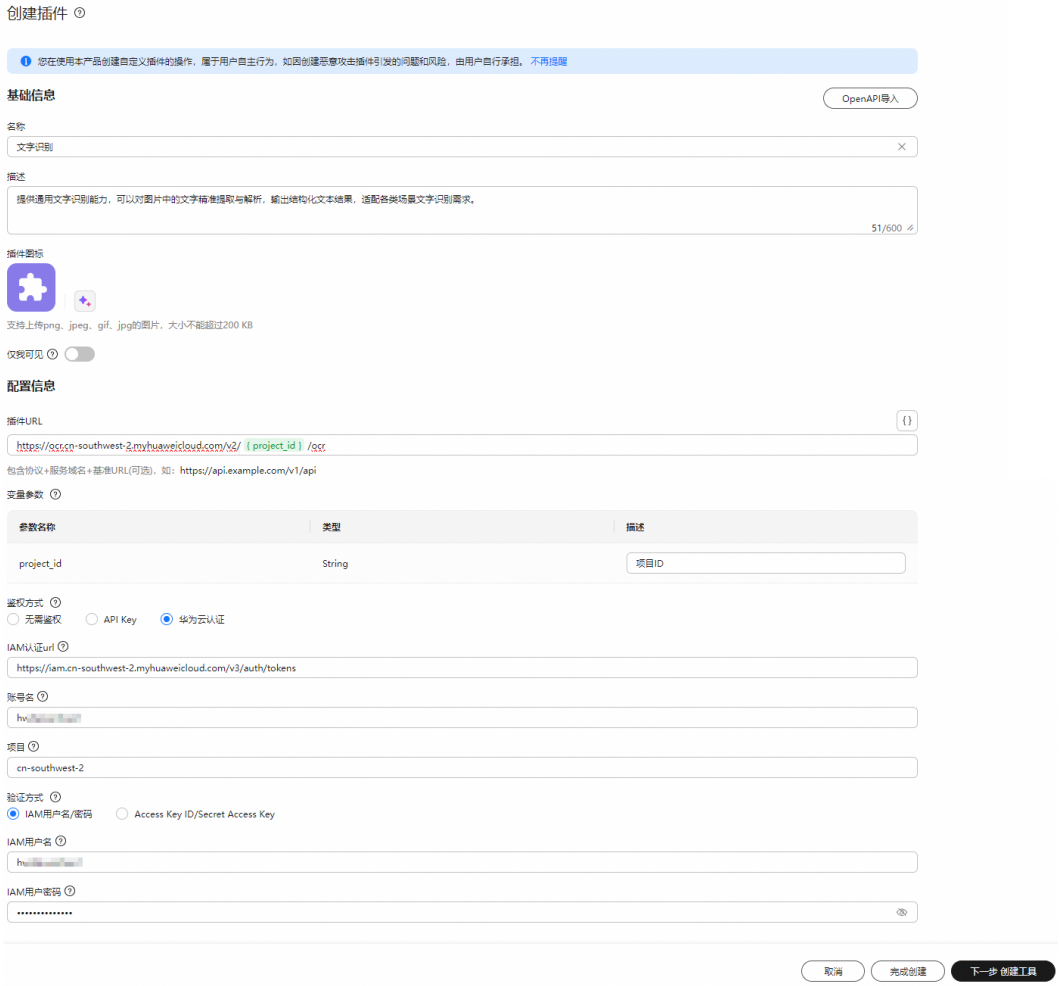


表 6-6 基础信息

参数	说明	示例
OpenAPI导入	为了简化插件配置流程，可以使用此功能导入符合OpenAPI 3.0规范的JSON文件。平台将自动解析文件内容并填充相应的配置参数，从而避免繁琐的手动录入，提高配置效率。 **注意** • **鉴权信息**：**平台目前无法自动解析鉴权配置**，导入后需要手动添加鉴权信息，否则工具调用时将因鉴权缺失而请求失败。 • **规范检查**：配置文件导入并解析为工具后，**建议逐一检查每个工具的请求参数和响应参数是否已正确解析**。 • 仅支持导入GET和POST请求方式的接口。	本示例使用手动创建，不使用自动解析功能。 其他示例代码请参考**OpenAPI 3.0 配置文件示例说明**。

版权所有 © 华为云计算技术有限公司

参数	说明	示例
名称	用于标识当前插件，在“组件库 > 插件”页面会展示该名称。 命名规则：按照插件的实际功能命名，有助于Agent进行插件的精准识别和调度。 命名要求：可以包含中文、英文、数字、下划线。 长度限制：2~64个字符。	文字识别
描述	描述当前插件的类型、功能和适用场景。 需要按照插件的实际功能填写描述，有助于Agent进行插件的精准识别和调度。	提供通用文字识别能力，可以对图片中的文字精准提取与解析，输出结构化文本结果，适配各类场景文字识别需求。
插件图标	系统默认提供插件图标。输入名称和描述后，单击  可以自动生成图标。用户也可以自定义图标，步骤如下： 1. 将鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。 支持png、jpeg、gif、jpg格式图片，且图片大小不大于200KB。	系统默认图标
仅我可见	开启后，该插件将变为私有，仅对当前登录用户可见，其他用户无法查看。 创建后该开关不再支持修改。	不开启

表 6-7 配置信息

参数	说明	示例
插件URL	调用插件的url地址。包含协议，服务域名和基准URL（可选）。 ● 协议：API服务接口通信协议。 ● 服务域名：提供API服务的服务域名。 ● 基准URL：是指域名的根路径（可理解为API的公共路径），默认为/，请根据实际API接口进行适配。	https://ocr.cn-southwest-2.myhuaweicloud.com/v2/{project_id}/ocr 注意需要单击右侧的  按钮添加{project_id}变量。该变量表示项目ID，用于标识用户在华为云特定区域下的资源归属。

版权所有 © 华为云计算技术有限公司

参数	说明	示例
鉴权方式	选择调用API时是否需要鉴权。 • **无需鉴权**：API可以公开访问，不需要任何形式的身份认证或授权。 • **API Key**：在调用API时提供一个唯一的API Key进行鉴权。需配置以下信息：需填写密钥位置、密钥参数名称和取值。智能体在发起请求时，会自动把这个API Key添加到请求的Header（请求头）或Query（URL参数）中。 – 密钥位置：密钥是从Header中读取还是从Query中读取。 – 参数名称：API Key的鉴权参数名称。 – 参数值：API Key的具体值。 • **华为云认证**：华为云IAM认证，通过IAM账号获取用户Token进行认证。 **鉴权配置：** – IAM认证url：获取IAM用户Token的接口地址。例如：https://{iam_host}/v3/auth/tokens，其中{iam_host}表示IAM Endpoint地址，即承载REST服务端点的服务器域名或IP。由于不同服务在不同区域的Endpoint各不相同，详情请参见**终端节点**。 例如： 您需要获取“西南-贵阳一”区域的Token，则需使用“西南-贵阳一”区域的Endpoint（iam.cn-southwest-2.myhuaweicloud.com），并结合**获取IAM用户Token（使用密码）**API的资源路径（/v3/auth/tokens），拼接后的完整请求地址如下。 `https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens` – 账号名：IAM用户所属的账号名。 – 项目：服务所属的区域信息。例如：cn-southwest-2 **验证方式（二选一）：** 方式一：IAM用户名/密码 – IAM用户名：IAM用户的名称。 – IAM密码：IAM用户的登录密码。 方式二：Access Key ID / Secret Access Key – Access Key ID：访问密钥ID。	华为云文字识别服务采用“华为云认证”。 IAM认证url：https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens 项目：cn-southwest-2 验证方式：IAM用户名/密码 账号名、IAM用户名、IAM用户密码：账号名、IAM用户名请登录**“我的凭证”**页面查询，密码按实际填入。 **图 6-15** 查询账号名、IAM 用户名  **图 6-16** 配置示例 

版权所有 © 华为云计算技术有限公司

参数	说明	示例
	– Secret Access Key：与访问密钥ID结合使用的密钥。 **关于密钥加密：**当鉴权方式选择API Key或华为云认证时，鼠标移动至“参数列表”或“验证方式”右侧的 上，可查看密钥的加密方式。平台默认使用**KMS加密+默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 ● 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 ● 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	

步骤4 单击“下一步 创建工具”，进入工具创建页面，或单击“完成插件创建”。如果选择后者，后续可继续在插件详情中添加工具。请参考**步骤2：创建工具**为插件添加工具。

----结束

步骤 2：创建工具

如需了解插件与工具的关系，请参考**插件介绍**。

步骤1 在“工具信息”页面，单击“创建工具”，请参照**表6-8**进行配置。

表 6-8 工具配置信息

参数	说明	示例
名称	用于标识当前工具。 命名规则：按照工具的实际功能命名，有助于Agent进行插件的精准识别和调度。	通用文字识别
描述	描述当前工具的类型、功能和适用场景。 需要按照工具的实际功能填写描述，有助于Agent进行插件的精准识别和调度。	识别图片上的文字信息，以JSON格式返回识别的文字和坐标。

版权所有 © 华为云计算技术有限公司

<table>
<tr><th colspan="2">参数</th><th>说明</th><th>示例</th></tr>
<tr><td colspan="2">工具path</td><td>工具path是API完整调用地址中，用于定位具体功能的后缀片段。
完整API地址 = 服务域名 + 基准 URL + 工具path
插件层级（公共配置） 工具层级（独有配置）
<table><tr><th>插件名称</th><th>工具名称</th><th>API接口地址</th></tr><tr><td>文字识别插件</td><td>通用文字识别</td><td>https://{endpoint}/v2/{project_id}/ocr/general-text</td></tr><tr><td></td><td>通用表格识别</td><td>https://{endpoint}/v2/{project_id}/ocr/general-table</td></tr><tr><td></td><td>手写文字识别</td><td>https://{endpoint}/v2/{project_id}/ocr/handwriting</td></tr></table></td><td>/general-text
图 6-17 配置示例
  </td></tr>
<tr><td colspan="2">请求方式</td><td>API的请求方式，支持POST或GET。</td><td>POST</td></tr>
<tr><td rowspan="2">请求参数</td><td>请求头(Header)</td><td>HTTP请求消息的组成部分之一，请求头负责向服务器传递客户端请求的相关信息。
单击右侧的“添加参数”可以新增参数，参数配置说明请参见表6-9。</td><td>新增两个请求头参数，并设置为必选：
X-Auth-Token：获取方法请请参考附录。
Content-Type：取值为application/json</td></tr>
<tr><td>请求体(Body)</td><td>HTTP请求消息的组成部分之一，请求体承载发送给服务器的数据。
请求体支持json和form-data两种格式。当使用form-data格式时，如果参数类型选择为File，支持文件上传。当请求体的数据格式是form-data时，文件上传大小请参考约束限制。
可以根据以下两种方式添加参数：
● 单击右侧“添加参数”可以新增参数，参数配置说明请参见表6-9。
● 单击“导入参数”，输入符合JSON或JSONSchema格式的参数内容。单击“确定”，平台将自动解析并完成导入。
注意
– 平台不支持导入包含“中文名称”的参数。如需设置参数的“中文名称”，请在导入后手动配置。
– 此操作将覆盖原有的参数配置，请谨慎操作。</td><td>文字识别支持2种请求方式，使用图片的url，或者图片的Base64编码。
此处使用url，并设置为必选，使用base64编码需要填写为image。
图 6-18 配置示例
  </td></tr>
</table>

 版权所有 © 华为云计算技术有限公司

参数		说明	示例
	查询参数(Query)	HTTP请求消息的组成部分之一，用于向服务器传递额外的参数信息。这些参数通常以键值对的形式出现，并且附在URL的路径后面，通过?分隔。 例如，在/items?id=123中，查询参数为ID，值为123。 单击右侧的“添加参数”可以新增参数，参数配置说明请参见**表6-9**。	文字识别插件不涉及
	路径参数(Path)	自动解析工具path中包含的路径参数。工具path中支持可变参数配置。 例如：/weather/weatherInfo{path_1}/{path_2}。	文字识别插件不涉及
	参数封装	满足某些接口强制要求入参为数组的规定，将对象类型(Object)参数变为数组(Array)，也就是将入参放到一对中括号[]中。 • 原参数列表：{"a":"string", "b":1}。 • 开启后：[{"a":"string", "b":1}]。	不启用
响应参数	参数列表	响应参数清单，按照API实际的响应参数结构进行填写。 单击右侧的“导入参数”，输入符合JSON或JSONSchema格式的参数内容。单击“确定”，平台将自动解析并完成导入。 此操作将覆盖原有的参数配置，请谨慎操作。	增加一个响应参数，名称为result，并设置为必选。 **图 6-19** 配置示例 
	流式响应	当插件调用的是一个带有思考效果的模型（如DeepSeek）或者耗时较长的接口，开启这个功能可以让API逐步（流式）返回响应结果，而不是等API完全获取结果后再一次性显示出来。	不启用
	参数封装	满足后续插件使用过程中（如工作流的循环节点批量读取数组）要求输入为数组的规定，将对象类型(Object)响应变为数组(Array)，也就是将出参放到一对中括号[]中。 • 原参数列表：{"a":"string", "b":1}。 • 开启后：[{"a":"string", "b":1}]。	不启用

版权所有 © 华为云计算技术有限公司

表 6-9 参数配置说明

参数	说明
参数名称	设置请求参数的名称，参数名称会作为大模型解析参数含义的依据。 命名规则：仅支持字母、数字、下划线或短横线。 **说明** • 单击参数列表右侧的“添加参数”按钮，可以添加请求参数。 • 单击右侧的，可以删除添加的请求参数。
中文名称	设置参数的中文名称，便于用户理解参数含义。
参数类型	设置请求参数的数据类型。 **注意** 请求头（Header）中，所有参数的值必须是字符串类型，不能设置其他类型。
默认值	设置参数的默认值，当参数未提供时使用该值。
描述	设置请求参数的详细描述信息，准确说明参数的含义、用途和格式要求，以提高大模型对参数识别和提取的准确性。
参数校验	设置当前参数是否需要进行校验。 校验规则： • 参数名称：需要校验的参数名称。 • 校验类型： – **字符最大长度** – **枚举值** – **时间日期** • 校验规则：可设置指定格式和自定义格式。 – **指定格式**：选择系统预置的标准校验规则。当校验类型为时间日期时，支持指定格式。 – **自定义格式**：根据实际需求自定义校验规则。
必填	设置该参数是否为必填项。

步骤2 配置完单击“确定”，工具创建完成后会跳转至插件详情页面。

步骤3 在工具列表中，单击“调试”按钮，输入问题后，单击“开始调试”。

必须连通公网，才能获取示例中的{project_id}的值。

对于文字识别插件示例，由于在创建插件的基准URL过程中，将{project_id}设置为变量，因此要填充该值。请登录**“我的凭证”**页面查询“西南-贵阳一”区域对应的项目ID。

文字识别插件如果返回401或403报错，请检查“权限校验”配置中的Token或密钥是否正确。

版权所有 © 华为云计算技术有限公司

图 6-20 获取文字识别插件项目 ID



步骤4 工具调试完成后，单击“确定”。

调试成功后，工具状态随之变为“成功”，此时可单击右上角“发布”，发布插件。

----结束

相关操作

工具创建完成后，您可以在工具列表中查看每个工具的调试状态、智能体引用数和工作流引用数。您可以执行如**表6-10**的操作。

表 6-10 相关操作

操作	说明
编辑	在工具信息列表中，找到需要编辑的工具，单击该工具操作列中的“编辑”，可编辑工具信息。
调试	在工具信息列表中，找到需要调试的工具，单击该工具操作列中的“调试”，在展开的弹框中输入参数信息对工具进行调试。
删除	在工具信息列表中，找到需要删除的工具，单击该工具操作列中的“删除”，可删除工具。
查看详情	在工具信息列表中，找到需要查看详细信息的工具，单击该工具名称，可查看工具的详细信息。

附录

创建文字识别插件所需的Token获取方法：

代码中账号名、IAM用户名、IAM用户密码：账号名、IAM用户名请登录**“我的凭证”**页面查询，密码按实际填入。

图 6-21 查询账号名、IAM 用户名



- cURL

```
curl --location --request POST 'https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens' \
--header 'Content-Type: application/json' \
--data-raw '{
```


版权所有 © 华为云计算技术有限公司

```
  "auth": {
    "identity": {
      "methods": [
        "password"
      ],
      "password": {
        "user": {
          "name": "IAM用户名",
          "password": "密码",
          "domain": {
            "name": "账号名"
          }
        }
      }
    },
    "scope": {
      "project": {
        "name": "cn-southwest-2"
      }
    }
  }
}'
```

- Python

```
import requests
import json
url = "https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens"
payload = json.dumps({
  "auth": {
    "identity": {
      "methods": [
        "password"
      ],
      "password": {
        "user": {
          "name": "IAM用户名",
          "password": "密码",
          "domain": {
            "name": "账号名"
          }
        }
      }
    },
    "scope": {
      "project": {
        "name": "cn-southwest-2"
      }
    }
  }
})
headers = {
  'Content-Type': 'application/json'
}
response = requests.request("POST", url, headers=headers, data=payload)
print(response.headers["X-Subject-Token"])
```

OpenAPI 3.0 配置文件示例说明

OpenAPI 3.0 简介

OpenAPI 3.0（前身为Swagger）是一种用于描述RESTful API的标准规范格式。它以结构化的方式定义API的请求路径、参数、请求体、响应结构等信息，使得机器和人类都能快速理解API的能力与使用方式。

一个标准的OpenAPI 3.0文档通常包含以下核心模块：

 版权所有 © 华为云计算技术有限公司

表 6-11 OpenAPI 3.0 核心模块

核心模块	说明
openapi	声明所使用的OpenAPI规范版本，如 3.0.1。
info	API的基本元信息，包括标题（title）、描述（description）和版本号（version）。
servers	API的服务器地址列表，定义请求的基础URL。
paths	核心部分，定义所有可用的API路径及其支持的HTTP方法（GET、POST等），每个方法即对应一个具体的接口（工具）。
components（本示例未使用）	可复用的数据模型、参数、响应体等定义，用于减少重复描述。
security（本示例未使用）	全局或接口级别的鉴权方式声明，如API Key、华为云认证等。

平台在创建Agent插件时，会解析导入的OpenAPI 3.0配置文件：

- **info.title**： 将作为插件名称。
- **info.description**：将作为插件描述。
- **paths**：下的每个接口将被解析为插件中的一个工具（Tool）。

示例：

本示例提供了一个符合OpenAPI 3.0规范的“用户管理 API”配置文件，旨在演示如何通过编写标准化的OpenAPI 3.0规范的文件来定义Agent插件。

通过详细解析文件结构（包括基本信息、服务器地址、接口路径）及具体的接口操作（如查询用户列表、创建新用户），本示例帮助用户理解OpenAPI字段与平台插件配置项（如插件名称、工具定义、参数约束）之间的映射关系。您可以参考此示例，学习如何正确编写配置文件，以便将现有的RESTful API快速导入平台并解析为可用的工具。

```
openapi: 3.0.1
info:
  title: 用户管理 API
  description: 这是一个简单的用户管理 API 示例
  version: 1.0.0
servers:
  - url: https://api.example.com/v1
    description: 生产服务器
paths:
  /users:
    get:
      summary: 获取用户列表
      description: 返回系统中的用户列表，支持分页和过滤
      operationId: getUsers
      parameters:
        - name: page
          in: query
          description: 页码
          schema:
            type: integer
```

 版权所有 © 华为云计算技术有限公司

```
      default: 1
  - name: limit
    in: query
    description: 每页数量
    schema:
      type: integer
      default: 20
  - name: role
    in: query
    description: 按角色过滤
    schema:
      type: string
      enum: [admin, user, guest]
responses:
  '200':
    description: 成功返回用户列表
    content:
      application/json:
        schema:
          type: object
          properties:
            total:
              type: integer
              description: 用户总数
            page:
              type: integer
              description: 当前页码
            limit:
              type: integer
              description: 每页数量
            data:
              type: array
              items:
                type: object
                properties:
                  id:
                    type: integer
                    description: 用户ID
                  username:
                    type: string
                    description: 用户名
                  email:
                    type: string
                    format: email
                    description: 电子邮箱
                  role:
                    type: string
                    enum: [admin, user, guest]
                    description: 用户角色
                  createdAt:
                    type: string
                    format: date-time
                    description: 创建时间
  '400':
    description: 请求参数错误
    content:
      application/json:
        schema:
          type: object
          properties:
            code:
              type: integer
              description: 错误码
            message:
              type: string
              description: 错误信息
  default:
    description: 服务器错误
    content:
```

 版权所有 © 华为云计算技术有限公司

```
        application/json:
          schema:
            type: object
            properties:
              code:
                type: integer
                description: 错误码
              message:
                type: string
                description: 错误信息
post:
  summary: 创建新用户
  description: 创建一个新的用户账户
  operationId: createUser
  requestBody:
    required: true
    content:
      application/json:
        schema:
          type: object
          required:
            - username
            - email
            - password
          properties:
            username:
              type: string
              minLength: 3
              maxLength: 20
              description: 用户名
            email:
              type: string
              format: email
              description: 电子邮箱
            password:
              type: string
              minLength: 6
              format: password
              description: 密码
            role:
              type: string
              enum: [user, guest]
              default: user
              description: 用户角色
  responses:
    '201':
      description: 用户创建成功
      content:
        application/json:
          schema:
            type: object
            properties:
              id:
                type: integer
                description: 用户ID
              username:
                type: string
                description: 用户名
              email:
                type: string
                format: email
                description: 电子邮箱
              role:
                type: string
                enum: [admin, user, guest]
                description: 用户角色
              createdAt:
                type: string
                format: date-time
```

 版权所有 © 华为云计算技术有限公司

```
            description: 创建时间
    '400':
      description: 请求数据验证失败
      content:
        application/json:
          schema:
            type: object
            properties:
              code:
                type: integer
                description: 错误码
              message:
                type: string
                description: 错误信息
              errors:
                type: array
                description: 详细错误列表
                items:
                  type: object
                  properties:
                    field:
                      type: string
                      description: 错误字段
                    message:
                      type: string
                      description: 错误信息
    '409':
      description: 用户已存在
      content:
        application/json:
          schema:
            type: object
            properties:
              code:
                type: integer
                description: 错误码
              message:
                type: string
                description: 错误信息
    default:
      description: 服务器错误
      content:
        application/json:
          schema:
            type: object
            properties:
              code:
                type: integer
                description: 错误码
              message:
                type: string
                description: 错误信息
```

6.1.3.2 基于 JSON 文件导入 API 插件

AgentArts支持通过JSON文件形式导入API插件，可以直接将API插件中定义的接口、入参、出参、鉴权等信息编辑为一份JSON文件，再导入平台创建API形态的插件。

JSON导入功能允许您通过配置文件而非手动单击界面来创建插件，平台解析该文件后，会自动完成插件注册、工具定义及鉴权配置，实现插件的即插即用。该功能特别适用于批量迁移（如将开发环境的大量插件快速迁移到生产环境）和版本管理（如将插件配置存储在代码仓库中，实现变更追踪和回滚）等场景。

该JSON文件定义了插件的基础属性、后端服务连接地址、输入输出参数结构以及鉴权方式。开发者可通过修改此文件，快速将现有的RESTful API接入到平台中创建成插件。推荐结合**示例：创建联网搜索插件**、**基于API创建插件**进行查看，便于理解JSON文件中的各参数。

版权所有 © 华为云计算技术有限公司

前提条件

- 已**开通AgentArts服务**。
- 建议先阅读**基于API创建插件**了解插件的基本概念和创建流程，或直接参考**示例：创建联网搜索插件**导出现有插件作为模板。

约束与限制

- 当前仅支持导入通过“API类型”创建的插件。
- 仅支持上传.jsonl格式文件。
- 最大支持128MB。

获取JSON示例文件：

当前平台需要事先创建、发布一个API插件，再导出该插件的JSON配置文件。您可以参考**示例：创建联网搜索插件**创建插件并导出插件的JSON文件。

图 6-22 导出插件



配置文件由顶层的metadata和import_type组成。所有的业务逻辑配置均位于metadata对象中。

```
{
   "metadata": { ... },     // 核心配置区域
   "import_type": "Plugin"  // 标识文件类型，固定为 "Plugin"
}
```

插件基础信息：

定义插件在平台页面上的展示信息及唯一标识。

表 6-12 插件基础信息参数

参数	名称	示例
plugin_display_name	插件英文标识名	"web_search"

版权所有 © 华为云计算技术有限公司

参数	名称	示例
plugin_chinese_name	插件中文名称	"联网搜索插件"（在市场或编排页展示的名字）
plugin_desc	插件描述	"联网搜索插件"（用于模型理解插件用途）
icon	插件图标	Base64编码的图标
type	插件类型	"custom"（表示用户自定义插件）
call_mode	调用模式	"api"（核心字段，标识此为HTTP接口调用）

服务连接配置：

字段：request_info。注意，此字段的值是一个经过转义的JSON字符串（例如，原始JSON 为 {"a":1}，在文件中应填写为 "{\"a\":1}"）。它定义了API的地址和工具列表。

```
{
   "basic_info": {
      "host": "qianfan.baidubce.com",  // API 域名
      "protocol": "https",              // 协议
      "path": "/v2/ai_search"           // 基础路径前缀
   },
   "tool_info": [
      {
         "tool_id": "7381...",         // 工具唯一ID
         "tool_display_name": "web_search", // 工具名
         "method": "POST",             // HTTP方法
         "path": "/web_search",        // 具体接口路径（完整路径为basic path + tool path）
         "tool_desc": "网络搜索工具"    // LLM决策时参考的工具描述
      }
   ]
}
```

tool_info中的tool_id是唯一标识。如果您修改了tool_id，请务必同步更新input_schema、output_schema、intf_type等数组中对应的tool_id，否则会导致参数配置无法匹配。

接口契约定义：

这是配置文件中最复杂的部分，定义了插件如何组装参数以及如何处理返回值。

- 输入参数定义 (input_schema)

  结构说明：该数组中的每个对象对应一个工具，必须包含tool_id以匹配tool_info中的工具，以及具体的input_schema（参数定义）内容。

  作用：定义API请求所需的Header、Body、Query参数。

  注意：在编辑此字段时，必须严格遵守JSON转义规则（例如 " 必须写成 \"），否则会导致导入失败。如果您修改了tool_info中的tool_id，请务必同步更新此处的tool_id。

- 输出参数定义 (output_schema)

  格式：JSON字符串数组。

  结构说明：该数组中的每个对象对应一个工具，必须包含tool_id以匹配tool_info中的工具，以及具体的output_schema（返回结构）内容。

 版权所有 © 华为云计算技术有限公司

作用：定义API响应的结构，帮助Agent提取关键信息。

鉴权配置：

字段：auth_info

定义了Agent调用外部API时的认证方式。

```
示例
"auth_info": {
   "scope": "SERVICE",      // 鉴权范围
   "domain": "HEADERS",     // 鉴权参数位置：HEADERS, QUERY, BODY
   "auth_keys": [
      {
         "target_name": "Authorization", // HTTP Header的Key
         "auth_key": "B****3"            // 实际的API Key值（通常导入后需重新配置以保密）
      }
   ]
}
```

生命周期与状态控制：

表 6-13 参数说明

参数	说明
visibility	可见性，如 "project"为项目内可见。"domain"为域（租户）内可见。
auth_required	用户是否需要自行授权。false表示使用平台预置的统一鉴权。
intf_type	接口类型。包含JSON字符串，通常值为"blocking"，代表同步阻塞调用（即等待API返回结果后再继续）。

如果您需要基于此模板创建一个新的API插件（例如接入天气查询API），请遵循以下步骤修改JSON：

- 修改基础信息：更新plugin_chinese_name和plugin_desc。
- 更新请求地址：

  将request_info中的host修改为您的API域名（如api.weather.com）。

  更新tool_info中的path和method。
- 重写Schema：

  这是最容易出错的步骤。建议先编写标准的JSON Schema对象，然后使用工具将其转换为转义后的字符串，填入input_schema和output_schema。
- 配置鉴权：

  如果API需要Key，在auth_info中填入对应的Header Key名称和Value。

完整JSON示例

使用**示例：创建联网搜索插件**导出的完整JSON示例文件如下：

```
{
   "metadata": {
      "plugin_id": "6282baa8-abe7-47bc-a9e7-77f537843580",
      "project_id": "832bce5c182f4c528188e4a9c120ea10",
      "plugin_display_name": "lianwangsousuo",
```


版权所有 © 华为云计算技术有限公司

```
    "plugin_chinese_name": "联网搜索",
    "plugin_desc": "通过搜索引擎获取互联网实时信息，包括新闻、百科、网页内容等",
    "icon": "...",
    "request_info": "{\"basic_info\":{\"host\":\"qianfan.baidubce.com\",\"path\":\"/v2/ai_search\",\"protocol
\":\"https\",\"input_schema\":\"{\\\"type\\\":\\\"object\\\",\\\"properties\\\":{},\\\"required\\\":[]}
\"},\"tool_info\":[{\"tool_id\":\"3a9de9fc917a40be\",\"tool_display_name\":\"web_search
\",\"tool_chinese_name\":\"网络搜索\",\"tool_desc\":\"通过搜索引擎获取互联网实时信息，包括新闻、百科、网
页内容等\",\"method\":\"POST\",\"path\":\"/web_search\"}]}",
    "auth_info": {
      "scope": "SERVICE",
      "domain": "HEADERS",
      "auth_keys": [
        {
          "target_name": "Authorization",
          "auth_key": "B****f"
        }
      ]
    },
    "visibility": "domain",
    "input_schema": "[{\"tool_id\":\"3a9de9fc917a40be\",\"input_schema\":\"{\\\"type\\\":\\\"object\\\",\\
\"properties\\\":{\\\"Authorization\\\":{\\\"default\\\":\\\"Bearer bce-v3/ALTAK...\\\",\\\"location\\\":\\
\"Headers\\\",\\\"validate_rule\\\":\\\"\\\",\\\"validate_type\\\":\\\"CHAR\\\",\\\"validated\\\":false,\\
\"name_cn\\\":\\\"Authorization\\\",\\\"type\\\":\\\"string\\\",\\\"description\\\":\\\"Authorization\\\"},\\
\"Content-Type\\\":{\\\"default\\\":\\\"application/json\\\",\\\"location\\\":\\\"Headers\\\",\\\"validate_rule\\
\":\\\"\\\",\\\"validate_type\\\":\\\"CHAR\\\",\\\"validated\\\":false,\\\"name_cn\\\":\\\"Content-Type\\\",\\
\"type\\\":\\\"string\\\",\\\"description\\\":\\\"Content-Type\\\"},\\\"messages\\\":{\\\"location\\\":\\\"Body\\
\",\\\"validate_rule\\\":\\\"\\\",\\\"validate_type\\\":\\\"CHAR\\\",\\\"validated\\\":false,\\\"name_cn\\\":\\\"\\
\",\\\"type\\\":\\\"array\\\",\\\"description\\\":\\\"messages\\\",\\\"items\\\":{\\\"type\\\":\\\"object\\\",\\
\"properties\\\":{\\\"role\\\":{\\\"default\\\":\\\"user\\\",\\\"location\\\":\\\"Body\\\",\\\"validate_rule\\\":\\\"\\
\",\\\"validate_type\\\":\\\"CHAR\\\",\\\"validated\\\":false,\\\"name_cn\\\":\\\"\\\",\\\"type\\\":\\\"string\\\",\\
\"description\\\":\\\"角色\\\"},\\\"content\\\":{\\\"location\\\":\\\"Body\\\",\\\"validate_rule\\\":\\\"\\\",\\
\"validate_type\\\":\\\"CHAR\\\",\\\"validated\\\":false,\\\"name_cn\\\":\\\"\\\",\\\"type\\\":\\\"string\\\",\\
\"description\\\":\\\"用户问题\\\"}},\\\"required\\\":[\\\"role\\\",\\\"content\\\"]}}},\\\"required\\\":[\\
\"Authorization\\\",\\\"Content-Type\\\",\\\"messages\\\"]}\"}]",
    "output_schema": "[{\"tool_id\":\"3a9de9fc917a40be\",\"output_schema\":\"{\\\"type\\\":\\\"object\\\",\
\\\"properties\\\":{\\\"references\\\":{\\\"type\\\":\\\"array\\\",\\\"description\\\":\\\"输出信息\\\",\\\"items\\\":
{\\\"type\\\":\\\"string\\\"}}},\\\"required\\\":[\\\"references\\\"]}\"}]",
    "is_input_list": "[{\"tool_id\":\"3a9de9fc917a40be\",\"is_input_list\":false}]",
    "is_output_list": "[{\"tool_id\":\"3a9de9fc917a40be\",\"is_output_list\":false}]",
    "type": "custom",
    "call_mode": "api",
    "intf_type": "[{\"tool_id\":\"3a9de9fc917a40be\",\"intf_type\":\"blocking\"}]",
    "metadata": "",
    "creator": "hwPersonalaccount",
    "creator_id": "082c85020600265e1f34c008f1aaacfb",
    "created_on": 1780135248000,
    "updated_on": 1783579119000,
    "test_status": "[{\"tool_id\":\"3a9de9fc917a40be\",\"test_status\":1}]",
    "last_version_id": "1780138253433",
    "auth_required": false,
    "trace_id": "6282baa8-abe7-47bc-a9e7-77f537843580",
    "published": 1,
    "is_free": 0,
    "domain_id": "082c85016600258f0fe3c0080c949e20",
    "is_share": 0,
    "view_times": 3
  },
  "import_type": "Plugin"
}
```


导入JSON示例文件：

JSON文件准备完成后，在平台插件页面执行导入操作。

导入后重新对插件进行调试，如果出现401 Invalid token报错，请尝试重新配置插件的鉴权信息。

 版权所有 © 华为云计算技术有限公司

图 6-23 导入 JSON 文件



6.1.4 调试发布插件

插件创建完成后，需对其中的工具进行调试，确保插件能够正确调用。调试完成后，可对插件进行发布。只有处于发布状态的插件，才能被智能体正常使用。

前提条件

已创建插件，并已完成工具的参数配置，操作方法参考**创建插件**。

调试工具、发布插件

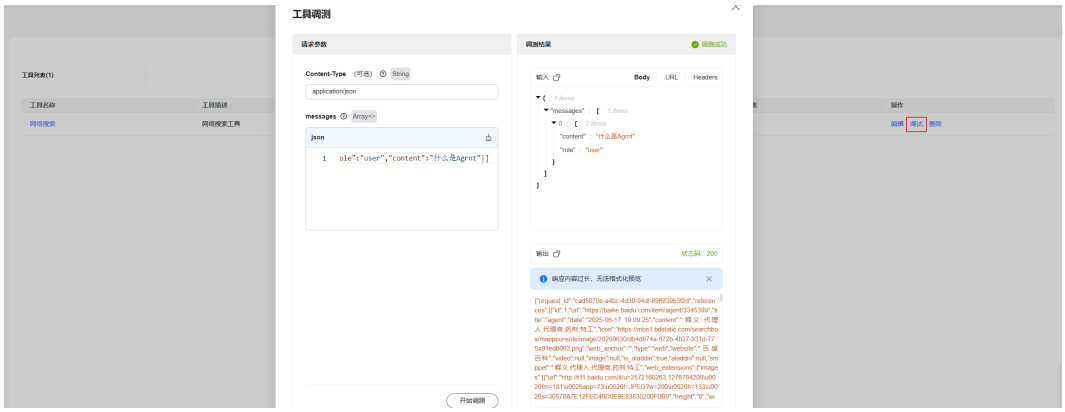
步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，在“插件”页签，选择已创建的插件，单击插件名称，并进入“插件详情”页面。

步骤3 单击工具列表操作列的“调试”。根据创建工具时定义的请求参数结构填写调试用例。单击“开始调测”执行调试工作。

调试成功后，页面会显示API的返回结果（如API的响应参数数据），且状态码通常显示200。

图 6-24 工具调试



步骤4 调试成功后检查输出的内容是否符合预期，调试无问题后，关闭“工具调测”窗口，单击页面右上角的“发布”，发布插件。

----结束

版权所有 © 华为云计算技术有限公司

常见问题

网络连接失败/超时

请检查API服务的域名是否正确。

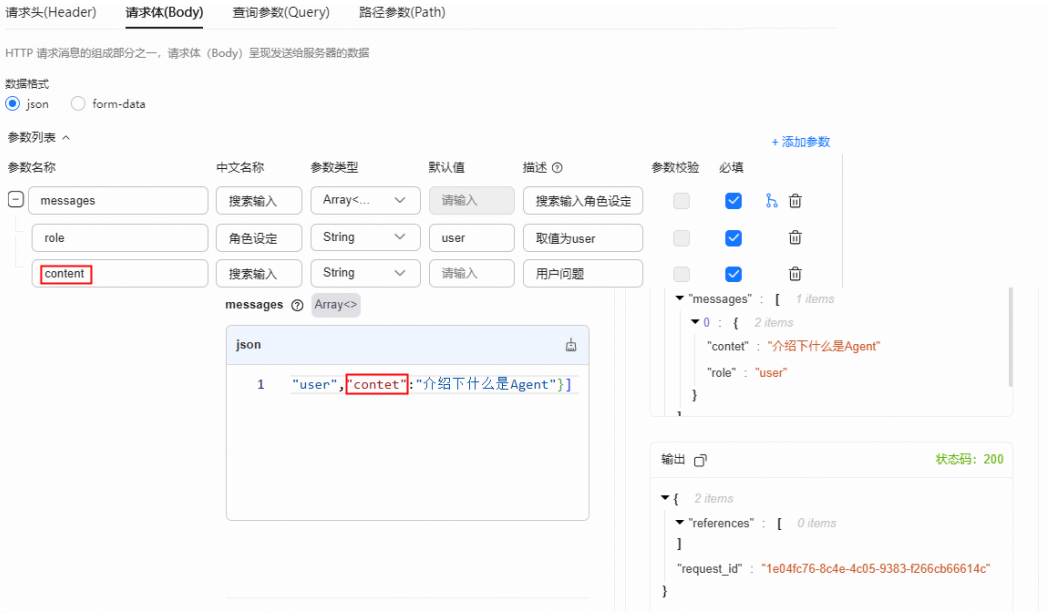
认证配置不可用

插件认证不可用，页面提示AgentArts.02401121报错。请检查API鉴权参数是否填写错误或已过期。

调试成功但输出无内容

调试示例参数填写错误或与插件定义不一致，会导致调试成功但输出为空（假性成功），请检查调试参数是否与API要求匹配。

图 6-25 调试参数与插件定义参数不一致



6.1.5 使用插件

6.1.5.1 在单智能体中使用插件

如果把大模型（LLM）比作智能体的“大脑”，那么插件就是它的“眼睛”和“手”。

通过集成插件，单智能体可以突破仅基于文本生成的局限，获得以下能力：

- 连接外部世界：实时查询天气、搜索网页、读取数据库。
- 执行具体操作：发送邮件、预定会议、生成图表。
- 私有数据交互：调用企业内部API获取业务数据。

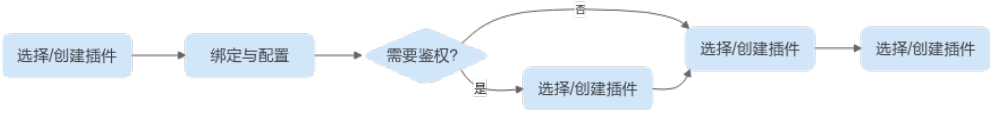
在单智能体中，大语言模型能够理解用户意图，并主动调用插件来获取精准信息、完成特定任务。当您配置插件后，系统会自动将插件的名称、描述和参数定义注入到模型的System Prompt（系统提示词）中。模型会根据用户的提问，自动判断是否需要调用插件，以及需要提取什么参数。

版权所有 © 华为云计算技术有限公司

前提条件

- 如果需要添加个人插件，请确保已**创建插件**和**调试发布插件**。
- 如果需要添加官方预置插件，请确保已完成插件鉴权。您可以在“资产广场”了解平台预置的官方插件。

单智能体使用插件流程概览



1. 选择/创建插件：在资产广场了解平台预置的官方插件或创建自定义插件。
2. 绑定与配置：创建单智能体过程中，将插件挂载到智能体，如果插件需要鉴权，则需额外配置鉴权信息。
3. **编排提示词（关键）**：在提示词中明确插件的使用边界，确保模型准确调用。

操作步骤

步骤1 登录**AgentArts智能体平台**。

步骤2 选择左侧导航栏“开发中心 > 智能体管理”，单击右上角“创建单智能体”。

图 6-26 创建应用



步骤3 填写单智能体名称、描述，如均填写为“智能体助手”。单击右下角“立即创建”进入到单智能体配置页面。

步骤4 在配置页面的中间区域，选择一个可用的模型。并在“插件”区域，单击+选择一个插件。

如果插件需要进行鉴权后才可以使用，请按照页面提示配置插件鉴权。

步骤5 配置智能体的提示词。

为了提高模型调用插件的准确率，建议在“提示词”中补充调用插件的说明。可参考如下示例进行填写：

```
你是一个全能助手。当用户询问实时信息（如天气、股价、新闻）时，必须优先调用已配置的工具/插件获取数据，不要依据你训练的知识编造答案。调用工具时，请确保提取的参数准确。
```

步骤6 配置完成后，可在页面右侧区域调试智能体。

版权所有 © 华为云计算技术有限公司

智能体回复内容中应包含“调用了插件”的标识，并返回基于真实数据的查询结果。如果出现报错或未调用插件，请检查插件鉴权信息是否正确或过期，以及插件的描述是否准确，防止智能体因为插件描述模糊，未能成功调用到插件。

----结束

相关文档

单智能体应用中使用插件的详细信息，请参考**添加插件**。

6.1.5.2 在工作流中使用插件

相比于单智能体使用插件的“概率性调用”（由大模型决定是否用），工作流使用插件是“确定性调用”（流程走到这一步必须执行）。因此，设置的重心必须从“提示词设置”转移到“参数映射”和“数据流转”上。

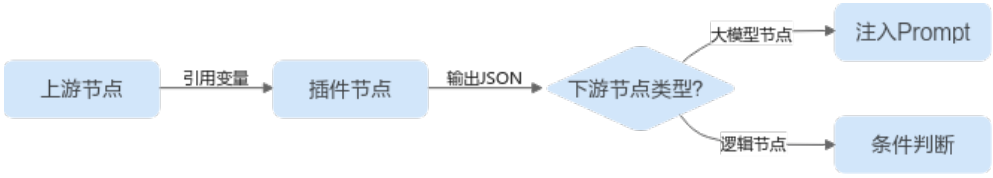
适用场景：

- 固定业务逻辑：必须先查询订单状态，再根据状态回复用户。
- 精确数据处理：需要调用计算器、汇率转换、数据库查询等不容许出错的操作。
- 连接外部系统：对接企业ERP、CRM系统获取实时信息。

前提条件

- 如果需要添加个人插件，请确保已**创建插件**和**调试发布插件**。
- 如果需要添加官方预置插件，请确保已完成插件鉴权。您可以在“资产广场”了解平台预置的官方插件。

数据流转与约束

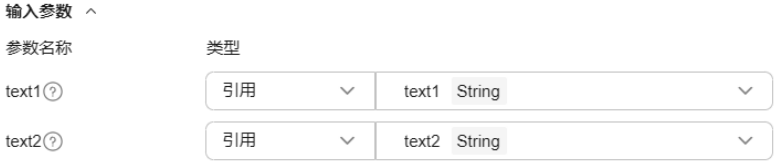


在配置插件节点前，必须理解数据是如何在节点间流转的。

配置插件的输入数据

需要在页面中通过设置“引用”选择上游节点，同时被引用的上游节点在参数类型、参数结构上需与插件本身的入参保持一致。

图 6-27 设置参数引用示例



下游节点使用插件数据

插件节点执行后，通常会输出一个JSON对象。

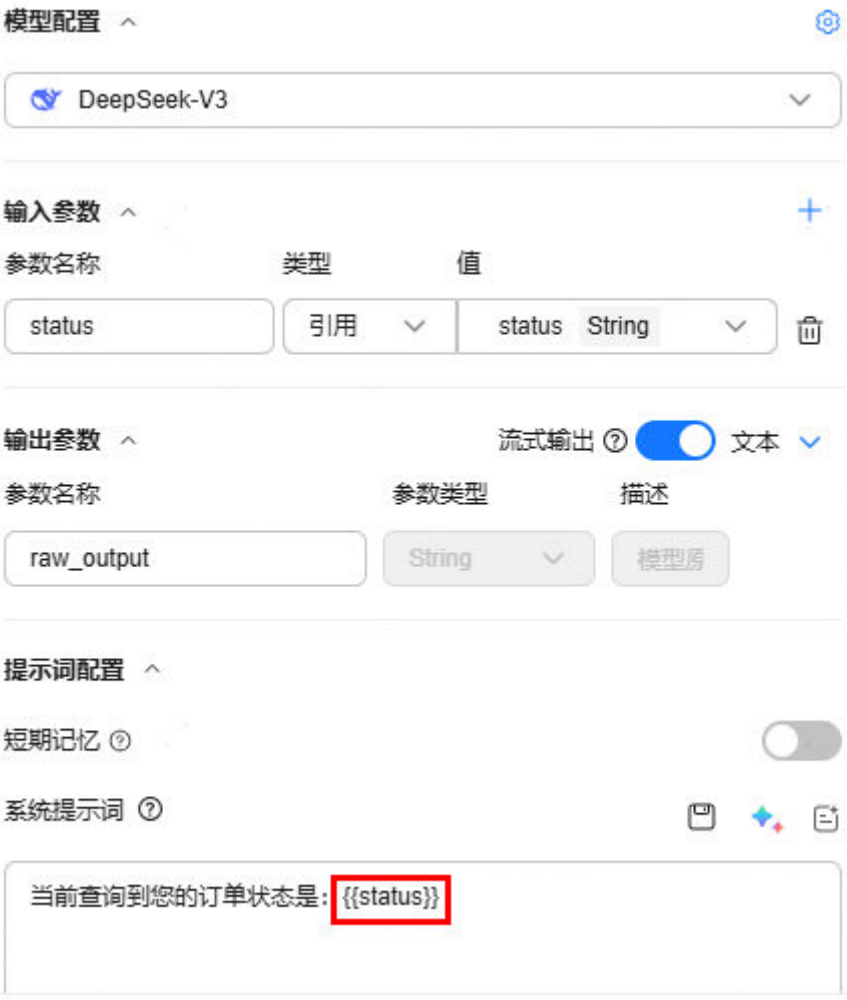
版权所有 © 华为云计算技术有限公司

- 场景1：大模型节点引用插件数据

  将插件查到的客观信息（如天气温度、订单状态）传递给大模型，让模型根据这些信息组织语言回复用户。

  例如在大模型节点的提示词（Prompt）中，使用变量语法（如 {{output}}）插入插件的输出结果。

  图 6-28 大模型节点引用插件示例

- 场景2：逻辑类节点基于数据做判断

  根据插件返回的状态（成功/失败、有货/无货、包含特定值……），决定工作流走哪一条路径。

  例如在逻辑类节点的条件设置中，选择插件输出的特定字段（如status），并设置判断规则（例如：当status等于Success时，走分支A）。

操作步骤

步骤1 登录AgentArts智能体平台。

步骤2 选择左侧导航栏“开发中心 > 智能体管理 ”，选择“工作流”页签，并单击右上角的“创建工作流”。

 版权所有 © 华为云计算技术有限公司

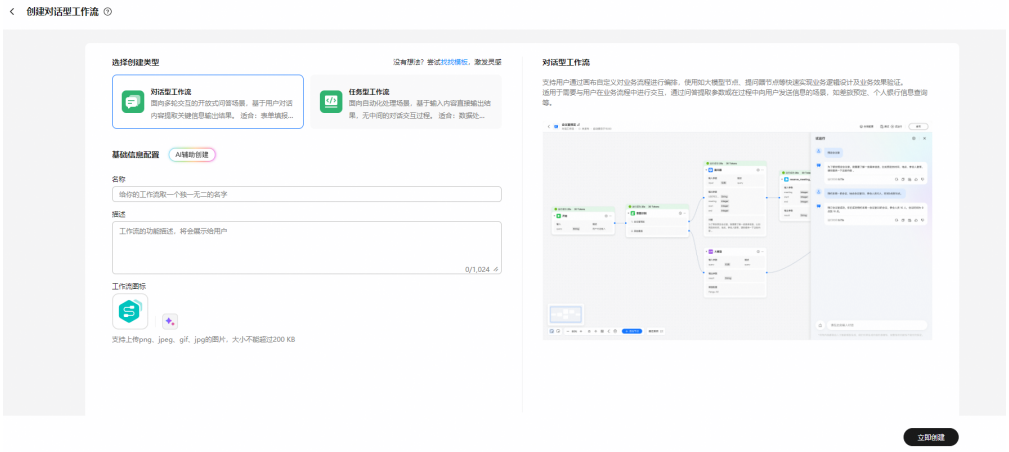
图 6-29 创建应用



步骤3 选择“对话型工作流”或者“任务型工作流”。这两类工作流中均具备插件节点。

- 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。

图 6-30 选择工作流应用



步骤4 填写工作流名称、描述。单击右下角“立即创建”进入到工作流配置页面。

步骤5 在配置页面可以添加插件至工作流中，并配置插件的输入数据。

插件选定后，面板会显示该插件对应的输入参数以及输出参数等信息。可以“引用”上游节点作为插件的输入数据来源。**注意插件的参数类型需要与上游节点的参数类型保持一致。**（如果插件执行报错，请检查插件本身是否可用；上游节点是否确实输出了插件所需的变量；上游输出变量的数据类型（如String、Int、Object）是否与插件定义的输入参数类型完全一致）

如果插件中有不需要变化的固定值参数，直接在输入框中手动填入固定的文本内容即可。

如果插件需要进行鉴权后才可以使用，请按照页面提示配置插件鉴权。

版权所有 © 华为云计算技术有限公司

图 6-31 插件引用与固定输入示例



步骤6 连接插件的下游节点。

插件执行完只是获取了数据，其后可以连接一个大模型节点来处理这些数据并生成回复。并注意在大模型节点的系统提示词中添加插件返回的查询内容。

```
示例：
你是一个天气助手。
用户的问题是：{{query}}
调用天气插件返回的数据是：{{output}}
请根据上述数据，用自然语言回答用户。
```

----结束

6.1.6 管理插件

AgentArts服务支持对插件进行版本管理、各版本历史记录查看、引用查询、删除、导出/导入插件的JSON文件等操作，确保插件全生命周期管理的规范性。核心功能包括：

- 版本管理：发布新插件版本并记录版本信息，按时间顺序查看所有插件版本的发布详情。
- 引用查询：定位使用该插件的智能体与工作流，支持手动更新版本。
- 插件删除：移除不再使用的插件（高危操作，需谨慎）。
- 导出/导入插件JSON文件：实现插件的本地备份或迁移。

前提条件

- 已**创建插件**。
- 已**开通AgentArts服务**。

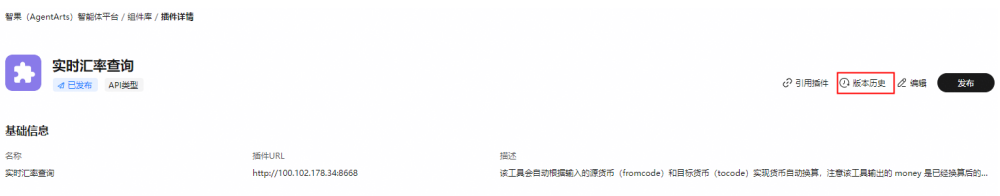
管理插件版本

可以通过发布操作生成插件的新版本，系统会记录版本名称与更新描述，便于后续追溯。

- **查看插件历史版本**

  登录平台后，在左侧导航栏选择“开发中心 > 组件库”，在“插件”页面，单击目标插件，进入插件详情页。可对插件执行版本发布、查看历史版本操作。

图 6-32 管理插件版本



版权所有 © 华为云计算技术有限公司

- **删除插件历史版本**

  在插件详情页面的右上角，单击“版本历史”。鼠标移动到需要删除的历史记录卡片上，单击“删除”，即可删除该发布历史。

  说明

  如果插件已被智能体、工作流使用，此时删除插件是高危操作。删除会导致智能体无法运行，且操作不可撤回。执行前请务必确认。

  图 6-33 删除发布历史



删除插件

登录平台后，在左侧导航栏选择“开发中心 > 组件库”，在“插件”页面，将鼠标悬停在需要删除的插件片上，单击“删除”即可删除该插件。

图 6-34 删除插件



版权所有 © 华为云计算技术有限公司

说明

如果该插件已被引用，删除后引用将自动取消，可能会导致智能体无法运行，且该操作不可撤回，请谨慎操作。

查看插件引用列表

查询使用该插件的智能体与工作流，且支持手动更新插件版本。

为了防止插件版本更新（如参数变更、接口地址修改）导致线上运行的智能体或工作流出现异常，插件不会自动更新。插件发布新版本后，已引用该插件的应用不会自动升级，需由开发者确认无误后手动更新，以确保生产环境的稳定性。

- 单智能体中引用的插件无论是平台预置的插件还是个人创建的插件，都不会自动更新到最新版本。需要手动更新。
  a. 在插件详情界面，单击“引用插件”，可以查看该插件被哪些智能体引用。

  图 6-35 查看插件的引用信息

  b. 在引用插件列表中单击智能体名称，将直接跳转至该智能体的编排页面，您可在此页面更新插件。

  图 6-36 引用插件列表

- 在工作流中使用插件时，无论是平台预置的插件还是个人创建的插件，都不会自动更新到最新版本。这意味着即使插件发布了新版本，工作流仍会继续使用当前指定的版本，确保应用的稳定运行。如果您需要在工作流中使用最新的插件版本，可以在工作流页面根据提示手动升级插件版本，如图6-37所示。

版权所有 © 华为云计算技术有限公司

图 6-37 工作流中插件升级



导出/导入插件 JSON 文件

通过导出/导入插件JSON文件功能实现插件的本地备份或迁移，仅支持对整个插件操作，不支持插件内单个工具的单独导出/导入。

登录平台后，在左侧导航栏选择“开发中心 > 组件库”，在“插件”页面，单击页面右上角“导出”按钮，可以导出所需的插件。单击“导入”按钮可以导入插件的JSON文件。关于JSON文件的介绍请参考**基于JSON文件导入API插件**。

图 6-38 导出/导入



6.2 Skill

6.2.1 Skill 介绍

什么是 Skill

在AgentArts平台上，Skill可以将过去分散在不同文档、代码片段或零散笔记中的隐性知识，以显性化和结构化的方式组织起来，使AI能够按需、稳定地执行专业任务。您可以将Skill看作是开发者的工具箱，将积累的技术经验和最佳实践打包成技能包，使AI成为您专业领域的专家。与一次性使用的提示词不同，Skill是可以保存、复用和优化的任务解决方案，使AI不仅能理解您的指令，更能掌握您的方法。当一个任务的处理流程相对固定、可重复操作，并且需要一致和可靠的输出时，可以将其封装为Skill，供通用智能体使用。

版权所有 © 华为云计算技术有限公司

例如，假设您是一名软件开发工程师，经常需要进行代码审查。您可以将代码审查的最佳实践和常见检查点整理成一个Skill包，包括代码风格检查、逻辑错误检测、性能优化建议等。这样，每次进行代码审查时，AI可以自动应用这些技能，确保审查过程的一致性和可靠性。

说明

本功能仅在贵阳一区域（Region）提供。

当前版本仅支持Skill的导入、管理和浏览，暂不支持在智能体中调用使用。后续版本将逐步开放智能体集成能力。请关注最新动态公告。

Skill 的工作机制

Skill是一种遵循统一规范的文件夹结构，清晰地定义了完成特定任务所需的标准作业流程（SOP）、方法论和参考模板。为了提高对话效率，智能体不会一次性加载Skill的全部内容，而是通过“按需加载”机制，在对话过程中动态判断哪些Skill与当前任务相关，并仅加载完成任务所需的必要信息，从而避免上下文窗口过载。

一个标准的Skill文件夹通常包含以下核心组成部分：

- **SKILL.md**：必需文件
  - **描述**：这是Skill的核心元数据文件和使用说明书，用于向智能体说明该Skill的名称、用途以及操作指南。
  - **功能**：智能体在启动时会优先加载该文件的元数据信息，以便在需要时快速判断是否调用该Skill，并确定如何调用。
- **scripts/目录：**
  - **描述**：存放可执行的代码文件，用于完成具体的计算任务。
  - **功能**：这些脚本文件包含明确的逻辑和算法，使智能体能够执行特定的操作，确保任务的确定性和可重复性。
- **references/目录：**
  - **描述**：存放与Skill相关的参考资料，包括第三方API文档、设计规范、测试报告等。
  - **功能**：这些文档为智能体提供背景信息和上下文支持，帮助其更好地理解任务需求并优化执行过程。
- **assets/目录：**
  - **描述**：存放Skill所需的静态资源和模板文件，例如配置模板、音视频文件等。
  - **功能**：这些资源文件为智能体提供执行任务所需的额外信息和素材，确保任务执行的准确性和高效性。

通过清晰的文件夹结构和按需加载机制，Skill能够帮助智能体在对话场景中灵活扩展功能，从通用助手快速转变为特定领域的专家。

Skill 的制品包规范

Skill制品包为ZIP格式，遵循**Agent Skills Standard**，以SKILL.md为核心清单文件。

1. **目录结构**

   制品包内部的目录结构示例如下：

```
skill-package.zip
└──skill-package
   ├── SKILL.md            # 技能清单文件（必需，YAML frontmatter + Markdown）
```


版权所有 © 华为云计算技术有限公司

```
├── scripts/              # 可执行脚本目录
│   ├── main.py           # 入口文件
│   └── ...
├── references/           # 参考文档（可选）
│   └── REFERENCE.md
└── assets/               # 静态资源（可选，模板、图片等）
```

2. **SKILL.md清单文件格式**：

   SKILL.md完全遵循**Agent Skills Standard**规范。由YAML frontmatter（元数据头）和Markdown（正文描述）两部分组成。必须包含name和description字段，且name不可修改。

```
---
name: skill-creator
description: Guide for creating effective skills. This skill should be used when users want to create a new skill (or update an existing skill) that extends Claude's capabilities with specialized knowledge, workflows, or tool integrations.
license: Complete terms in LICENSE.txt
---

# 此处为 Markdown 正文内容
...
```

3. **元数据字段定义**

   以下为SKILL.md文件中YAML frontmatter的字段详细说明（遵循Agent Skills Standard）：

表 6-14 元数据字段定义

字段	是否必需	说明
name	是	技能唯一标识。 ● 最长64字符（NFKC归一化后）。 ● 仅允许Unicode字母、数字、连字符。 ● 必须全小写；不能以连字符开头或结尾；不允许连续连字符。
description	是	技能描述。 ● 最长1024字符。 ● 需清晰描述技能功能及触发条件。
license	否	许可证名称或文件引用（例如LICENSE.txt）。
allowed-tools	否	允许该技能调用的工具列表。
metadata	否	扩展元数据，支持任意键值对（如author、version等）。
compatibility	否	运行环境要求。 ● 数据类型须为字符串。 ● 最长500字符。

版权所有 © 华为云计算技术有限公司

Skill 与其他资源的区别

在AgentArts中，Skill是一种专为增强智能体能力设计的标准化组件，包含指令文档、元数据配置和执行脚本等要素，能够为智能体注入特定领域的专业能力、流程规范和经验法则。为了更好地理解Skill与其他资源的区别，以下是对Skill与提示词、工作流、插件、MCP和知识库的详细分析：

表 6-15 Skill 与其他资源的区别

对比项	Skill的定义与特点	对比项的定义与特点
提示词	提供特定领域的专业能力与方法，按需触发，用于完成具体任务。	定义智能体的全局人设和工作方式，持续生效，影响输出风格。
工作流	灵活的操作说明书，允许自主决策，不固定执行路径。	预设的固定流程，明确步骤顺序，适用于自动化任务。
插件	封装使用策略，指导何时何地使用插件。	负责调用外部API，处理技术接入。
MCP	指导如何使用MCP提供的接口完成任务。	提供能力调用接口，访问外部服务。
知识库	提供标准操作流程（SOP）与最佳实践，处理明确问题更高效。	存储文档与素材内容，适合信息检索与内容补充类场景。

在AgentArts中，Skill是一种强大的工具，通过将隐性知识显性化、结构化，使智能体能够稳定、高效地执行专业任务。Skill与提示词、工作流、插件、MCP和知识库在功能和应用场景上各有侧重，但它们共同构成了智能体的综合能力体系，使智能体能够在不同场景下灵活、高效地完成任务。

Skill 的类型

根据Skill的提供者和使用范围，Skill可分为以下类型：

表 6-16 Skill 的类型

服务	技能类型	说明
AgentArts	资产广场Skill	由AgentArts开发和维护的编程技能，通过资产广场面向所有用户开放。在对话中，AgentArts会按需自动调用官方Skill完成任务。
	组件库Skill	AgentArts支持用户手动上传的技能，这些技能仅限您个人使用。技能的范畴广泛，不仅限于编程，还包括语言、沟通、管理、技术、创意、商业、个人发展、专业、生活和社交技能等。

版权所有 © 华为云计算技术有限公司

服务	技能类型	说明
		注意 请您注意，"第三方技能"来源于第三方服务商提供，当您使用第三方服务商提供的技能时，您应严格遵守第三方的相关协议。您理解并同意，第三方服务商提供的技能可能存在风险，我们无法对其做出任何形式的保证。请您在使用前审慎考量并评估其风险。 AgentArts支持导入通过第三方工具生成的资源包，或存储在OBS服务中的资源包。

资产广场 Skill

AgentArts通过资产广场Skill页签，与开发者共同构建技能生态系统。Skill使用者则可以根据自身需求，从广场中挑选合适的Skill进行使用。“Skill”页面集中展示了平台精选的各类技能工具，如图6-39所示。这些技能功能丰富，覆盖通用工具、效率提升、自动化处理、设计与可视化、内容创作等多个实用领域，能够有效扩展智能体的应用能力。用户可根据实际业务需求，灵活调用和组合各类技能，以提升工作效率和实现业务目标。

图 6-39 Skill



6.2.2 示例：导入旅游规划 Skill

本文档介绍的旅游规划Skill，旨在根据用户提供的主题内容，智能推荐并导入合适的旅游规划方案。用户可以选择手动指定规划风格，或由系统根据输入内容自动匹配适合的模板，适用于商务考察旅游、休闲度假旅游、自然生态旅游、文化体验、红色旅游等多种应用场景。该Skill能够灵活适配不同需求，提供结构清晰、内容丰富的旅游规划支持，提升整体规划效率与用户体验。本示例以导入压缩包为例。

资源包格式要求

上传的资源包必须满足以下格式要求：

- **文件格式**：
  - 仅支持上传ZIP文件，文件大小≤10MB;
  - 名称格式需要满足条件：
    - 长度在2到64个字符。

版权所有 © 华为云计算技术有限公司

    - 只能包含小写字母（a-z）、数字（0-9）和连字符（-）。
    - 必须以字母或数字开头和结尾，不能以连字符开头和结尾，也不可包含连续连字符（如--）。
  - 解压后需有且仅有1个文件夹，命名与SKILL.md中的Skill名称一致，SKILL.md需位于文件夹内根目录下，详细内容请参考**Skill的制品包规范**。
- **推荐结构**：根目录纯净：解压后，根目录下必须直接包含配置文件（SKILL.md）。

```
travel-planner.skill
├── SKILL.md
├── assets/
│   ├── budget-checklist.md
│   └── itinerary-template.md
└── references/
    ├── destination-research.md
    └── travel-budget-guide.md
```

如果上传的资源包不符合规范，AgentArts会尝试自动适配，并可能通过对话方式向您收集缺失信息。

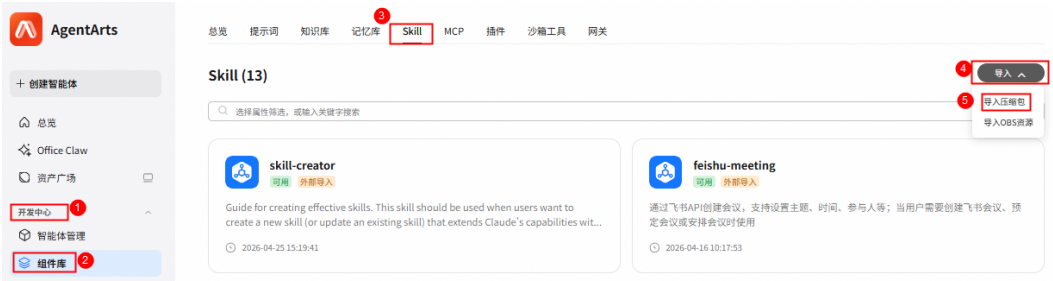
导入旅游规划 Skill

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，在“Skill”页签，单击页面右上角“导入”，并在导入下拉框中选择“导入压缩包”。

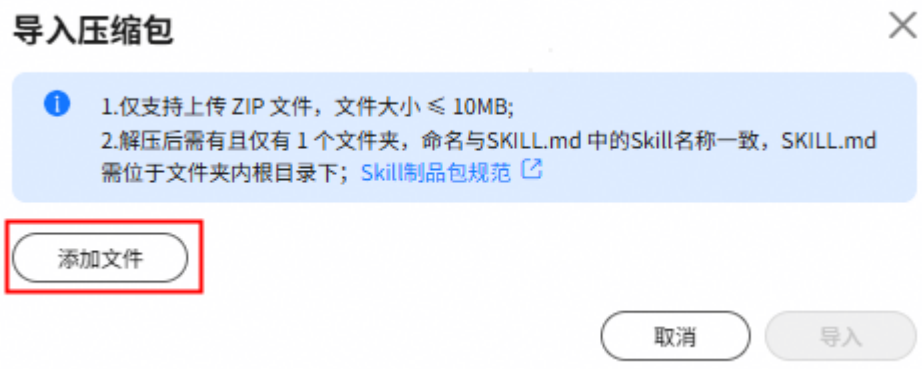
您可以通过**下载Skill压缩包**获取本示例Skill压缩包。

图 6-40 选择导入类型



步骤3 在“导入压缩包”界面单击“添加文件”，并在压缩包存放路径中选择旅游规划压缩包。

图 6-41 导入压缩包



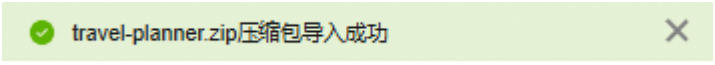
版权所有 © 华为云计算技术有限公司

图 6-42 导入 Skill



步骤4 单击“导入”，即可成功导入压缩包。

图 6-43 导入成功



----结束

Skill 配置说明

1. **SKILL.md文件内容**

```
---
name: travel-planner
description: 提供系统化的旅行规划流程，包括需求分析、景点推荐、路线设计、预算估算和行程优化；当用户需要"制定旅行计划"、"规划行程"、"推荐旅游景点"或"估算旅行费用"时使用
---
# Travel Planner
## 任务目标
- 本 Skill 用于：帮助用户制定完整、合理的旅行计划
- 能力包含：需求收集分析、目的地研究、行程设计、预算估算、行程优化
- 触发条件：用户表达需要规划旅行、制定行程、推荐景点或评估预算时
## 操作步骤
### 步骤 1：需求收集与分析
通过对话收集以下信息：
- 旅行目的地（城市/国家/区域）
- 旅行时间（出发日期、持续天数）
- 旅行人数（成人、儿童、老人）
- 预算范围（总预算或人均预算）
- 兴趣偏好（自然景观、历史文化、美食体验、购物娱乐等）
- 特殊需求（交通限制、身体状况、饮食禁忌等）
- 住宿偏好（酒店档次、位置要求）
### 步骤 2：目的地信息收集
根据用户偏好，整合目的地的关键信息：
- 主要景点与地标（必去景点、特色景点）
- 最佳游览季节与天气特点
```

 版权所有 © 华为云计算技术有限公司

- 交通方式（市内交通、城际交通）
- 特色美食与餐饮推荐
- 住宿区域建议
- 文化习俗与注意事项
参考：references/destination-research.md
步骤 3：行程设计
根据旅行天数和用户偏好，设计每日行程：
- 合理安排景点游览顺序（考虑地理位置、游览时间）
- 平衡活动强度（避免过度紧张或过于松散）
- 预留机动时间（应对突发情况）
- 标注交通方式和时长
- 推荐餐饮安排
步骤 4：预算估算
根据行程规划，提供详细的费用预估：
- 交通费用（往返机票/火车票、市内交通）
- 住宿费用（酒店/民宿费用）
- 餐饮费用（每日餐费标准）
- 门票与娱乐费用
- 购物与其他杂费
参考：references/travel-budget-guide.md
步骤 5：行程优化与调整
根据用户反馈优化行程：
- 检查行程时间是否合理
- 评估预算是否可控
- 调整景点顺序或取舍
- 补充实用建议（购票渠道、最佳游览时间等）
步骤 6：输出行程文档
生成结构化的行程计划文档：
- 行程总览（日期、天数、总预算）
- 每日详细行程（时间、地点、活动、交通、预算）
- 重要提醒与注意事项
- 应急联系方式与建议
使用模板：assets/itinerary-template.md
资源索引
- 预算指南：见 references/travel-budget-guide.md（费用类别与计算方法）
- 目的地研究：见 references/destination-research.md（信息收集清单）
- 行程模板：见 assets/itinerary-template.md（格式化行程表）
- 预算清单：见 assets/budget-checklist.md（预算核对表）
注意事项
- 充分利用智能体的地域知识和推理能力，无需调用外部地图或票务API
- 保持行程的灵活性，提供多个方案供用户选择
- 根据用户的预算范围进行合理推荐，避免超支建议
- 考虑不同用户群体的特殊需求（家庭、情侣、独旅者）
- 提供季节性建议（最佳旅行时间、天气注意事项）
使用示例
示例 1：城市游规划
- 功能说明：为3天的城市短途旅行设计行程
- 执行方式：智能体根据目的地推荐景点、规划路线、估算预算
- 关键要点：合理分配每日活动，平衡观光与休息
示例 2：多城市深度游
- 功能说明：为7天以上的多城市旅行设计详细行程
- 执行方式：智能体规划城市间交通、每个城市的重点景点
- 关键要点：预留足够的城市间移动时间，优化路线顺序
示例 3：主题游规划
- 功能说明：围绕特定主题（美食、历史、自然等）设计行程
- 执行方式：智能体聚焦主题推荐相关景点和体验
- 关键要点：深度挖掘主题相关资源，提供专业建议

2. **assets/budget-checklist.md示例代码**

Budget Checklist
使用说明
本清单用于旅行前和旅行期间的预算核对，确保费用控制在合理范围内。

行前预算核对
Step 1：设定总预算
- [ ] 确定旅行总预算：¥_____ 元
- [ ] 确定人均预算：¥_____ 元/人
Step 2：费用分配

 版权所有 © 华为云计算技术有限公司

```
| 类别 | 预算金额（元） | 占比 |
|------|----------------|------|
| 交通（往返+市内） | ¥_____ | __% |
| 住宿（每晚×晚数） | ¥_____ | __% |
| 餐饮（每日×天数） | ¥_____ | __% |
| 门票与娱乐 | ¥_____ | __% |
| 购物 | ¥_____ | __% |
| 其他（保险、应急） | ¥_____ | __% |
| **总计** | **¥_____** | **100%** |
### Step 3：单项明细确认
#### 交通费用
- [ ] 往返机票：¥_____ 元
- [ ] 往返火车票：¥_____ 元
- [ ] 市内交通预算：¥_____ 元/天 × ____ 天 = ¥_____ 元
- [ ] 其他交通：¥_____ 元
- **小计**：¥_____ 元
#### 住宿费用
- [ ] Day 1：¥_____ 元
- [ ] Day 2：¥_____ 元
- [ ] Day 3：¥_____ 元
- [ ] Day 4：¥_____ 元
- [ ] Day 5：¥_____ 元
- [ ] Day 6：¥_____ 元
- [ ] Day 7：¥_____ 元
- **小计**：¥_____ 元
#### 餐饮费用
- [ ] 早餐预算：¥_____ 元/天 × ____ 天 = ¥_____ 元
- [ ] 午餐预算：¥_____ 元/天 × ____ 天 = ¥_____ 元
- [ ] 晚餐预算：¥_____ 元/天 × ____ 天 = ¥_____ 元
- [ ] 零食饮料：¥_____ 元
- **小计**：¥_____ 元
#### 门票娱乐
- [ ] 景点1：¥_____ 元
- [ ] 景点2：¥_____ 元
- [ ] 景点3：¥_____ 元
- [ ] 景点4：¥_____ 元
- [ ] 景点5：¥_____ 元
- [ ] 娱乐活动：¥_____ 元
- **小计**：¥_____ 元
#### 购物费用
- [ ] 纪念品：¥_____ 元
- [ ] 特产：¥_____ 元
- [ ] 其他：¥_____ 元
- **小计**：¥_____ 元
#### 其他费用
- [ ] 保险：¥_____ 元
- [ ] 签证（如需）：¥_____ 元
- [ ] 应急备用金（总预算10%）：¥_____ 元
- **小计**：¥_____ 元
### Step 4：预算合理性检查
- [ ] 预算总额是否在可承受范围内？是 / 否
- [ ] 各类别预算分配是否合理？是 / 否
- [ ] 是否预留了足够的应急资金？是 / 否
- [ ] 是否有超支风险？低 / 中 / 高
---
## 旅行期间每日记账
### Day [X]：[日期]
#### 收支记录
| 项目 | 类别 | 金额（元） | 备注 |
|------|------|------------|------|
| 早餐 | 餐饮 | ¥_____ | |
| 午餐 | 餐饮 | ¥_____ | |
| 晚餐 | 餐饮 | ¥_____ | |
| 交通 | 交通 | ¥_____ | |
| 景点门票 | 门票娱乐 | ¥_____ | |
| 购物 | 购物 | ¥_____ | |
| 其他 | 其他 | ¥_____ | |
| **今日合计** | | **¥_____** | |
#### 预算对比
```

 版权所有 © 华为云计算技术有限公司

```
- 今日预算：¥_____ 元
- 今日实际：¥_____ 元
- 差额：¥_____ 元（超支/节省）
- 累计已用：¥_____ 元
#### 备注
- 今日支出最高的项目：_____，原因：_____
- 今日节省的项目：_____
- 需要注意的费用：_____
---
## 预算跟踪表
### 每日汇总
| 日期 | 当日预算 | 当日实际 | 差额 | 累计已用 | 剩余预算 | 备注 |
|------|----------|----------|------|----------|----------|------|
| Day 1 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 2 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 3 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 4 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 5 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 6 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| Day 7 | ¥_____ | ¥_____ | ¥_____ | ¥_____ | ¥_____ | |
| **总计** | **¥_____** | **¥_____** | **¥_____** | | | |
---
## 预算控制建议
### 节省策略
- [ ] 提前预订享受折扣
- [ ] 选择性价比高的住宿
- [ ] 避开景区内高价餐厅
- [ ] 购买城市通票
- [ ] 使用公共交通
### 避免超支
- [ ] 每日查看累计支出
- [ ] 大额消费前冷静思考
- [ ] 购物设定上限
- [ ] 避免冲动消费
- [ ] 保留应急备用金
### 超支应对
- 如果连续3天超支：
  - [ ] 检查超支原因
  - [ ] 调整后续行程预算
  - [ ] 减少非必要消费
  - [ ] 优先保障核心体验
---
## 最终结算
### 实际支出汇总
| 类别 | 预算 | 实际 | 差额 | 备注 |
|------|------|------|------|------|
| 交通 | ¥_____ | ¥_____ | ¥_____ | |
| 住宿 | ¥_____ | ¥_____ | ¥_____ | |
| 餐饮 | ¥_____ | ¥_____ | ¥_____ | |
| 门票娱乐 | ¥_____ | ¥_____ | ¥_____ | |
| 购物 | ¥_____ | ¥_____ | ¥_____ | |
| 其他 | ¥_____ | ¥_____ | ¥_____ | |
| **总计** | **¥_____** | **¥_____** | **¥_____** | |
### 结算分析
- **总预算**：¥_____ 元
- **实际支出**：¥_____ 元
- **超支/节省**：¥_____ 元（____%）
- **应急备用金使用**：¥_____ 元
### 经验总结
- **支出最高的类别**：_____
- **最值得的花费**：_____
- **可以节省的部分**：_____
- **下次预算调整建议**：_____
---
## 快速预算计算器
### 按天快速估算
```
每日费用 =（住宿/人）+（餐饮/人）+（交通/人）+（门票/人）
示例：舒适型5天旅行
```

 版权所有 © 华为云计算技术有限公司 
```

```
- 住宿：400元/人/天
- 餐饮：200元/人/天
- 交通：50元/人/天
- 门票：100元/人/天
每日费用：750元/人
5天总费用：3750元/人
```
按比例快速分配
- 总预算：¥_____ 元
- 交通（30%）：¥_____ 元
- 住宿（30%）：¥_____ 元
- 餐饮（20%）：¥_____ 元
- 门票娱乐（15%）：¥_____ 元
- 其他（5%）：¥_____ 元

3. **assets/itinerary-template.md示例**

```
# Travel Itinerary Template
## 行程总览
### 基本信息
- **目的地**：[城市/国家名称]
- **旅行日期**：[YYYY-MM-DD] 至 [YYYY-MM-DD]
- **旅行天数**：[X] 天
- **旅行人数**：[X] 人（[成人] 成人，[儿童] 儿童，[老人] 老人）
- **总预算**：[金额] 元
- **旅行主题**：[休闲度假/文化探索/美食之旅/自然风光等]
### 行程亮点
1. [亮点1]
2. [亮点2]
3. [亮点3]
---
## 每日详细行程
### Day 1：[日期] - [主题]
#### 上午（08:00 - 12:00）
| 时间 | 活动内容 | 地点 | 交通 | 预计费用 | 备注 |
|------|----------|------|------|----------|------|
| 08:00-09:00 | 早餐 | [餐厅名称/地点] | 步行/公交 | [金额]元 | |
| 09:00-10:30 | [景点1] | [具体地址] | [交通方式] | [金额]元 | 预计游览时长 |
| 10:30-12:00 | [景点2] | [具体地址] | [交通方式] | [金额]元 | |
#### 中午（12:00 - 13:30）
| 时间 | 活动内容 | 地点 | 交通 | 预计费用 | 备注 |
|------|----------|------|------|----------|------|
| 12:00-13:30 | 午餐 | [餐厅名称] | | [金额]元 | 推荐菜品： |
#### 下午（13:30 - 18:00）
| 时间 | 活动内容 | 地点 | 交通 | 预计费用 | 备注 |
|------|----------|------|------|----------|------|
| 13:30-14:00 | 交通 | 从[地点A]到[地点B] | [交通方式] | [金额]元 | |
| 14:00-16:00 | [景点3] | [具体地址] | | [金额]元 | |
| 16:00-18:00 | [景点4/自由活动] | [具体地址] | [交通方式] | [金额]元 | |
#### 晚上（18:00 - 22:00）
| 时间 | 活动内容 | 地点 | 交通 | 预计费用 | 备注 |
|------|----------|------|------|----------|------|
| 18:00-19:30 | 晚餐 | [餐厅名称] | | [金额]元 | |
| 19:30-22:00 | [夜游/休息/娱乐] | [具体地点] | [交通方式] | [金额]元 | |
#### 今日小结
- **今日费用**：[金额] 元
- **今日亮点**：
- **注意事项**：
---
### Day 2：[日期] - [主题]
[按相同格式填写...]
---
### Day 3：[日期] - [主题]
[按相同格式填写...]
---
## 预算明细
### 费用分类
| 类别 | 明细 | 金额（元） | 备注 |
|------|------|------------|------|
| **交通** | | | |
| | 往返机票/火车票 | | |
```

 版权所有 © 华为云计算技术有限公司

```
| | 市内交通 | | |
| | 城市间交通 | | |
| | **小计** | | |
| **住宿** | | | |
| | Day 1：[酒店名] | | |
| | Day 2：[酒店名] | | |
| | ... | | |
| | **小计** | | |
| **餐饮** | | | |
| | 早餐 | | |
| | 午餐 | | |
| | 晚餐 | | |
| | 零食饮料 | | |
| | **小计** | | |
| **门票娱乐** | | | |
| | 景点门票 | | |
| | 娱乐活动 | | |
| | **小计** | | |
| **购物** | | | |
| | 纪念品 | | |
| | 特产 | | |
| | **小计** | | |
| **其他** | | | |
| | 保险 | | |
| | 应急备用金 | | |
| | **小计** | | |
| **总计** | | | |
### 预算控制
- **预算上限**：[金额] 元
- **预计总费用**：[金额] 元
- **余额**：[金额] 元
- **超支风险**：[低/中/高]
---
## 实用信息
### 交通指南
- **机场/火车站到市区**：
  - 方式1：[交通方式]，约[X]分钟，[金额]元
  - 方式2：[交通方式]，约[X]分钟，[金额]元
- **市内交通**：
  - 一日票价格：[金额]元
  - 网约车：[App名称]
  - 出租车：起步价[金额]元
### 住宿信息
- **酒店名称**：[名称]
- **地址**：[详细地址]
- **电话**：[电话号码]
- **入住/退房时间**：[时间]
- **周边交通**：[最近的地铁站/公交站]
### 重要联系方式
- **紧急电话**：报警[110/当地号码]、急救[120/当地号码]
- **中国领事馆**：[电话号码]
- **酒店前台**：[电话号码]
- **出租车/网约车**：[电话号码]
- **航空公司/铁路客服**：[电话号码]
### 网络与通讯
- **当地SIM卡**：[运营商名称]，[套餐价格]
- **WiFi**：酒店免费WiFi / 公共WiFi热点
- **翻译工具**：[推荐App]
---
## 注意事项
### 行前准备
- [ ] 确认护照/身份证有效期
- [ ] 办理签证（如需要）
- [ ] 购买旅行保险
- [ ] 预订机票/火车票
- [ ] 预订酒店
- [ ] 准备现金和银行卡
- [ ] 兑换当地货币
- [ ] 下载必要App（地图、翻译、交通）
```

 版权所有 © 华为云计算技术有限公司

```
- [ ] 查询天气预报，准备衣物
- [ ] 准备常用药品
### 旅行期间
- [ ] 保持联系畅通
- [ ] 注意财物安全
- [ ] 遵守当地法律法规
- [ ] 尊重当地文化习俗
- [ ] 保持环境卫生
- [ ] 定期记录支出
### 应急预案
- **遗失证件**:
  - 立即联系领事馆
  - 报案并获取证明
  - 办理临时证件
- **生病就医**:
  - 联系酒店前台推荐医院
  - 使用保险报销
- **航班延误/取消**:
  - 联系航空公司改签
  - 联系酒店调整入住
---
## 行程调整记录
| 日期 | 调整内容 | 原因 | 新安排 | 费用变化 |
|------|----------|------|--------|----------|
| [日期] | [原计划] → [新计划] | [原因] | [新安排] | [+/-金额]元 |
---
## 旅行反馈
### 满意度评分
- **整体满意度**: [1-5分]
- **景点推荐**: [推荐/一般/不推荐]
- **住宿体验**: [推荐/一般/不推荐]
- **餐饮体验**: [推荐/一般/不推荐]
### 经验总结
- **最满意的安排**:
- **最失望的安排**:
- **值得改进的地方**:
- **下次建议**:
### 费用复盘
- **实际总费用**: [金额] 元
- **与预算对比**: [超支/节省] [金额] 元
- **超支/节省原因**:
```

4. **references/destination-research.md示例**

```
# Destination Research Checklist
## 概览
本清单提供目的地研究的关键信息框架，帮助系统化收集旅行规划所需的核心信息。
## 目录
- [基础信息](#基础信息)
- [交通信息](#交通信息)
- [景点信息](#景点信息)
- [住宿信息](#住宿信息)
- [餐饮信息](#餐饮信息)
- [文化习俗](#文化习俗)
- [实用信息](#实用信息)
## 基础信息
### 地理与气候
- 地理位置（国家、地区、城市）
- 气候特点（温度、湿度、降雨）
- 最佳旅行季节（月份、天气状况）
- 当前季节的天气特点
### 时区与语言
- 时区信息（与出发地的时差）
- 官方语言
- 英语普及程度
- 常用短语（问候、感谢、求助）
### 货币与支付
- 当地货币名称与汇率
- 支付方式（现金、信用卡、移动支付）
- 小费习惯（是否需要、标准金额）
- ATM分布情况
```

 版权所有 © 华为云计算技术有限公司

```
## 交通信息
### 到达方式
- 航班：机场名称、与市区的距离、交通方式
- 火车：主要火车站、班次频率
- 长途汽车：主要汽车站、运营时间
- 自驾：主要高速公路、停车情况
### 市内交通
- 地铁/轻轨：线路覆盖、票价、运营时间
- 公交：线路网络、票价、运营时间
- 出租车/网约车：起步价、叫车方式
- 共享单车/电动车：使用方法、收费标准
- 步行友好度：是否适合步行游览
### 城市间交通
- 高速铁路：班次、时间、票价
- 长途巴士：班次、时间、票价
- 飞机：航班、时间、票价
## 景点信息
### 必去景点
- 标志性景点（地标、建筑）
- 文化历史景点（博物馆、古迹）
- 自然景观（公园、山脉、海滩）
- 特色景点（独特体验、网红打卡点）
### 景点详情（每个景点）
- 开放时间
- 门票价格（成人、学生、儿童）
- 建议游览时长
- 最佳游览时间
- 是否需要预约
- 拍照限制
- 无障碍设施
### 景点分布
- 地理位置分布（按区域划分）
- 景点之间的距离
- 推荐游览路线（一日游/半日游）
- 避免拥挤的时间段
## 住宿信息
### 住宿区域选择
- 市中心（交通便利，价格较高）
- 景点附近（游览方便，性价比高）
- 机场/火车站附近（适合中转）
- 特色街区（体验当地生活）
### 住宿类型
- 酒店：星级、设施、价格范围
- 民宿：特色、环境、预订方式
- 青年旅社：床位价格、公共设施
- 公寓：整租、设施、价格
### 住宿建议
- 预订时机（提前多久）
- 性价比推荐
- 避雷指南（避免选择）
## 餐饮信息
### 当地特色美食
- 必尝菜品（代表性菜肴）
- 特色小吃（街头美食）
- 传统餐厅（老字号、特色店）
- 夜市美食（夜市推荐）
### 餐厅类型
- 高端餐厅：人均消费、特色
- 中档餐厅：人均消费、推荐
- 经济餐厅：人均消费、特色
- 快餐连锁：品牌、价格
### 餐饮建议
- 用餐时间（当地用餐习惯）
- 排队攻略（热门餐厅）
- 饮食禁忌（需要注意的）
- 安全饮食（避免生冷、街边摊）
## 文化习俗
### 当地文化
- 历史背景（简要了解）
```

 版权所有 © 华为云计算技术有限公司

- 文化特色（节日、习俗）
- 宗教信仰（如有）
行为规范
- 礼仪习惯（问候、礼貌用语）
- 着装要求（宗教场所、正式场合）
- 禁忌事项（避免冒犯）
- 拍照礼仪（何时可以拍）
购物文化
- 特色商品（值得购买）
- 购物场所（商场、市场、集市）
- 砍价习惯（是否需要）
- 消费水平（物价水平）
实用信息
通讯与网络
- SIM卡购买（运营商、套餐）
- WiFi覆盖（公共区域、酒店）
- 网络稳定性（4G/5G）
安全提示
- 治安状况（安全程度）
- 常见骗局（需要防范）
- 应急电话（报警、医疗）
- 重要物品保管（护照、现金）
医疗与健康
- 医院分布（主要医院）
- 药店位置（24小时药店）
- 常见疾病预防
- 疫苗接种要求
- 保险建议
其他建议
- 电源插座（是否需要转换器）
- 语言障碍（翻译工具）
- 天气准备（衣物、雨具）
- 行李建议（必备物品）
研究方法
信息来源
- 官方旅游网站（旅游局官网）
- 在线旅游平台（携程、马蜂窝、TripAdvisor）
- 社交媒体（小红书、微博、Instagram）
- 当地论坛和社区
- 旅游指南书籍（孤独星球等）
验证方法
- 对比多个来源的信息
- 查看最新评论（最近1-3个月）
- 关注季节性变化
- 确认价格时效性
研究示例
示例：研究京都旅游
- ****基础信息****：日本古都，气候四季分明，春秋最佳
- ****交通****：地铁发达，一日票500日元
- ****景点****：清水寺、金阁寺、伏见稻荷大社，门票200-600日元
- ****住宿****：祇园区域特色民宿，800-1500元/晚
- ****餐饮****：怀石料理、拉面、抹茶甜品，人均100-300元
- ****文化****：保持安静，寺庙内禁止拍照
- ****实用****：JR Pass适用于城际交通，7天通票约1800元

5. **references/travel-budget-guide.md示例**

Travel Budget Guide
概览
本指南提供旅行预算估算的完整框架，包括常见费用类别、计算方法和实用建议。
目录
- [费用类别](#费用类别)
- [费用估算方法](#费用估算方法)
- [预算优化建议](#预算优化建议)
费用类别
1. 交通费用
- ****国际/国内长途交通****
  - 机票：往返价格（含税费）
  - 火车：高铁/普铁票价
  - 长途汽车：大巴票价
  - 自驾：油费、过路费、停车费

 版权所有 © 华为云计算技术有限公司

- **市内交通**
  - 地铁/公交：单程票价
  - 出租车/网约车：平均费用
  - 自行车租赁：日租费用
  - 步行：免费（节省成本）
2. 住宿费用
- **酒店**
  - 经济型：200-400元/晚
  - 舒适型：400-800元/晚
  - 高端型：800-2000元/晚
  - 豪华型：2000元+/晚
- **民宿/青年旅社**
  - 青年旅社床位：50-150元/晚
  - 民宿单间：200-500元/晚
  - 整租公寓：500-1500元/晚
3. 餐饮费用
- **每日餐费标准（人均）**
  - 经济型：80-150元/天
  - 舒适型：150-300元/天
  - 高端型：300-600元/天
- **费用构成**
  - 早餐：20-50元
  - 午餐：30-80元
  - 晚餐：50-150元
  - 零食饮料：20-50元
4. 门票与娱乐
- **景点门票**
  - 免费景点：公园、广场、步行街
  - 低价景点：50-100元
  - 中价景点：100-200元
  - 高价景点：200-500元
- **娱乐活动**
  - 演出/演出票：100-500元
  - 体验活动（潜水、滑雪等）：200-1000元
  - 夜生活：酒吧、夜市等100-300元
5. 购物费用
- 纪念品：200-500元
- 当地特产：300-1000元
- 服装/配饰：500-2000元
- 其他购物：根据需求
6. 其他费用
- 保险：50-200元
- 通讯：SIM卡/漫游费：50-200元
- 签证（如需要）：200-1000元
- 应急备用金：总预算的10-15%
费用估算方法
方法 1：按天数估算
```
*总预算 =（每日平均费用 × 天数）× 人数*
```
示例：
- 5天旅行，舒适型标准
- 每日费用：住宿400 + 餐饮200 + 交通50 + 门票100 = 750元/人
- 总预算：750 × 5 × 2人 = 7500元
方法 2：按类别累计
```
*总预算 = 交通 + 住宿 + 餐饮 + 门票 + 购物 + 其他*
```
示例：
- 交通（往返）：2000元
- 住宿（4晚）：1600元
- 餐饮（5天）：1000元
- 门票：500元
- 购物：500元
- 其他：200元
- 总预算：5800元
方法 3：按预算分配
根据总预算分配到各类别：
- 交通：30%

 版权所有 © 华为云计算技术有限公司

```
- 住宿：30%
- 餐饮：20%
- 门票娱乐：15%
- 其他：5%
## 预算优化建议
### 省钱技巧
1. **交通**
   - 提前预订享受折扣
   - 选择淡季出行
   - 使用公共交通替代出租车
   - 比较不同交通方式的性价比
2. **住宿**
   - 预订享受早鸟优惠
   - 选择地理位置稍偏的酒店
   - 考虑民宿或青年旅社
   - 与朋友拼房分摊费用
3. **餐饮**
   - 避免景区内高价餐厅
   - 尝试当地小吃街
   - 自己购买食材（如有厨房）
   - 预订餐厅享受团购价
4. **门票**
   - 提前在线购票享受折扣
   - 购买城市通票（一日票/多日票）
   - 关注免费开放日
   - 选择部分付费景点，搭配免费景点
### 避免超支
- 预先设定各项费用的上限
- 记录每日支出，实时监控
- 预留10-15%的应急备用金
- 购物前冷静思考，避免冲动消费
- 使用预算清单核对：[assets/budget-checklist.md](assets/budget-checklist.md)
## 示例计算
### 示例 1：3天北京游（2人，舒适型）
| 类别 | 明细 | 金额 |
|------|------|------|
| 交通 | 往返高铁（2人） | 1200元 |
| 住宿 | 酒店2晚（舒适型） | 800元 |
| 餐饮 | 3天（2人） | 900元 |
| 门票 | 故宫、长城等 | 400元 |
| 市内交通 | 地铁/公交 | 200元 |
| 其他 | 应急备用金 | 350元 |
| **总计** | | **3850元** |
### 示例 2：7天云南游（2人，经济型）
| 类别 | 明细 | 金额 |
|------|------|------|
| 交通 | 往返机票（2人） | 3000元 |
| 住宿 | 酒店6晚（经济型） | 1800元 |
| 餐饮 | 7天（2人） | 2100元 |
| 门票 | 丽江、大理等 | 600元 |
| 市内交通 | 巴士/包车 | 800元 |
| 其他 | 应急备用金 | 830元 |
| **总计** | | **9130元** |
```

6.2.3 导入 Skill

AgentArts支持导入压缩包和导入OBS资源。

选型建议：如果您的资源包存储在本地且大小不超过10MB，推荐使用“导入压缩包”；如果资源包已存储在OBS中，或体积较大、或需要通过CI/CD流水线集成，推荐使用“导入OBS资源”。

- **导入压缩包：**在AgentArts平台中，用户可以通过“导入压缩包”功能，将本地开发完成的资源包上传至平台，平台会自动解析、检测、适配并部署到对话系统中使用。该功能适用于以下场景：
  - 已通过第三方工具生成资源包。

 版权所有 © 华为云计算技术有限公司

  - 在本地编写了完整的资源文件。
  - 想要将已有Skill或资源快速集成到AgentArts平台中。
- **导入OBS资源：**如果您在对象存储服务（OBS）中已存有开发完成的资源包，可以通过AgentArts工作台的“导入OBS资源”功能，将其快速集成到平台中。AgentArts智能引擎会自动拉取远程资源、解析包结构、执行深度安全性检测、补全元数据信息，并重新打包，以便在Agent编排或对话中直接调用。详细对象存储服务（OBS）资源获取操作详见**分享对象**。

  该功能特别适用于以下场景：
  - 已在OBS桶中存储了构建好的资源文件，需快速引入平台。
  - 资源包体积较大，通过本地上传耗时较长或易受网络限制。
  - CI/CD自动化流水线将构建产物直接输出至OBS，需实现无缝集成。

资源包格式要求

上传的资源包必须满足以下格式要求：

1. **导入压缩包：**
   - 仅支持上传ZIP文件，文件大小≤10MB;
   - 名称格式需要满足条件：
     - 长度在1到64个字符。
     - 只能包含小写字母（a-z）、数字（0-9）和连字符（-）。
     - 必须以字母或数字开头和结尾，不能以连字符开头和结尾，也不可包含连续连字符（如--）。
   - 解压后需有且仅有1个文件夹，命名与SKILL.md中的Skill名称一致，SKILL.md需位于文件夹内根目录下，详细内容请参考**Skill的制品包规范**。
2. **导入OBS资源：**将Skill ZIP文件上传至对象存储OBS服务，获取文件URL并在此传入，文件大小≤10MB。

说明

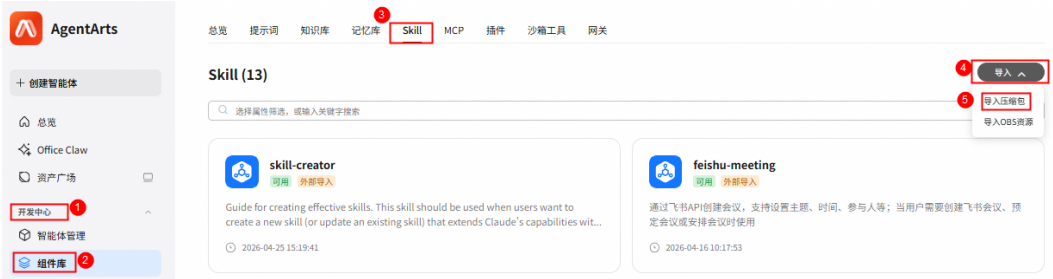
如果上传的资源包不符合规范，AgentArts会尝试自动适配，并可能通过对话方式向您收集缺失信息。

导入 Skill 步骤

步骤1 登录**AgentArts智能体平台**。

步骤2 选择左侧导航栏“开发中心 > 组件库 ”，单击左上角“Skill”页签并单击“导入”下拉框，选择“导入压缩包”或“导入OBS资源”。本文以导入压缩包为例：

图 6-44 选择导入类型



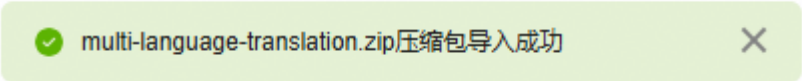
 版权所有 © 华为云计算技术有限公司

图 6-45 导入压缩包



步骤3 单击“导入”，即可成功导入压缩包。

图 6-46 导入成功



----结束

常见问题

- 问题：上传失败，提示“文件格式不支持”。
  原因：上传的文件不是.zip格式。
  解决方案：
  - 确保文件扩展名为.zip。
  - 如果是.rar或其他格式，请先转换为.zip。
  - 使用压缩工具（如WinRAR、7-Zip）将文件打包为.zip格式。
- **问题：上传后提示“Skill包结构不完整”**
  原因：Skill包缺少必要文件，如SKILL.md。
  解决方案：
  - 确保Skill包中包含以下文件：
    - SKILL.md：技能说明文件（必填）。
    - scripts/目录：存放可执行的代码文件。
    - references/目录：存放与Skill相关的参考资料。

 版权所有 © 华为云计算技术有限公司

    - assets/目录：存放Skill所需的静态资源和模板文件。
  - 检查文件路径是否正确，避免嵌套过深或路径错误。
- **问题：Skill无法解析或执行失败**

  原因：Skill包中的代码存在语法错误、依赖缺失或环境配置错误。

  解决方案：
  - 在本地环境中测试Skill脚本，确保其可以正常运行。
  - 检查代码中是否使用了平台不支持的库或语法。
  - 确保所有依赖项已正确安装（如Python 3.x、Node.js等）。

6.2.4 管理 Skill

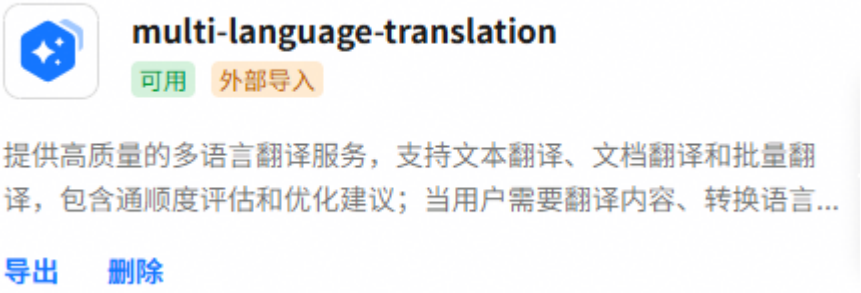
AgentArts服务支持对Skill进行引用查询、删除、导出、导入Skill文件等操作，确保Skill全生命周期管理的规范性。核心功能包括：

- 引用查询：定位使用该Skill的智能体。
- Skill删除：移除不再使用的Skill（高危操作，需谨慎）。
- 导出/导入Skill文件：实现Skill本地备份等。

删除 Skill

登录平台后，在左侧导航栏选择 “开发中心 > 组件库”，在“Skill”页面，鼠标指向目标Skill可以执行Skill删除操作。

图 6-47 删除 Skill



导出 Skill

通过导出/导入Skill功能实现Skill的本地备份。

登录平台后，在左侧导航栏选择 “开发中心 > 组件库”，在“Skill”页面，选择目标Skill，鼠标移至目标Skill卡片，单击“导出”按钮，可以导出ZIP格式压缩包。

版权所有 © 华为云计算技术有限公司

图 6-48 导出 Skill



导入 Skill

AgentArts平台支持导入用户本地自定义的Skill，以及第三方Skill，详情请参见**导入Skill**。

6.2.5 Skill 常见问题

本文档汇总了用户在使用AgentArts平台上传和使用Skill过程中可能遇到的常见问题及解决方案，帮助用户快速定位并解决问题。

Skill 上传失败相关问题

1. **问题：文件格式校验不通过**

   原因：
   - 提交的文件不是ZIP格式。
   - ZIP文件解压后包含多个文件夹。
   - 文件夹名称与SKILL.md中定义的Skill名称不一致。
   - SKILL.md文件未放置在文件夹根目录。

   解决方案：
   - 提交的文件必须为标准ZIP格式。
   - ZIP文件解压后应仅包含一个文件夹，且该文件夹名称需与SKILL.md中定义的Skill名称完全一致。
   - 确保SKILL.md文件位于该文件夹的根目录下，不可嵌套在子目录中。

2. **问题：文件大小校验不通过**

   原因：Skill包大小超过平台限制（最大10MB）。

   解决方案：
   - 压缩非必要文件或优化代码结构，确保Skill包大小不超过10MB。
   - 使用压缩工具进行无损压缩，减少文件体积。

3. **问题：上传Skill包失败，提示“缺少必要文件”**

   原因：Skill包缺少必要文件，如SKILL.md。

   解决方案：
   - 确保Skill包中包含以下文件：
     - SKILL.md：技能说明文件（必填）。

版权所有 © 华为云计算技术有限公司

- scripts/目录：存放可执行的代码文件。
- references/目录：存放与Skill相关的参考资料。
- assets/目录：存放Skill所需的静态资源和模板文件。

– 检查文件路径是否正确，避免嵌套过深或路径错误。

4. **问题：frontmatter字段白名单未通过校验**

   原因：

   – SKILL.md中包含不在白名单内的字段。
   – 必填字段（如name、description、license、allowed-tools、metadata、compatibility）缺失或格式错误。

   解决方案：

   – 仅允许使用白名单中的字段，禁止添加未授权字段。
   – 确保必填字段存在且格式正确，符合字段定义规范。

5. **问题：name字段未通过校验**

   原因：

   – name字段为空或长度不符合要求。
   – name字段包含非法字符。
   – name字段以连字符（-）开头或结尾。
   – name字段中包含连续连字符。
   – 制品包目录名与name字段不一致。

   解决方案：

   – name字段长度应在1-64字符之间。
   – 仅允许使用小写字母、数字和连字符（-）。
   – name字段不可以连字符开头或结尾，也不可包含连续连字符（如--）。
   – 制品包目录名必须与name字段完全一致。

6. **问题：description字段未通过校验**

   原因：

   – description字段为空或长度不符合要求。
   – description字段中包含XML标签。

   解决方案：

   – description字段长度应在1-1024字符之间。
   – 禁止在description中使用XML标签，确保内容为纯文本。

7. **问题：compatibility字段未通过校验**

   原因：

   – compatibility字段格式不正确。
   – compatibility字段长度超过500字符。

   解决方案：

   – compatibility字段必须为字符串类型。
   – 字段长度不得超过500字符，确保内容简洁明了。

 版权所有 © 华为云计算技术有限公司

6.3 MCP

6.3.1 MCP 介绍

MCP（Model Context Protocol）是一个开放协议，旨在打通大语言模型（LLM）应用与外部数据源、工具之间的交互壁垒。在传统的开发场景中，由于每个数据源、工具或服务都有独立的格式规范、对接协议和认证体系，开发者往往需要为每个API单独编写代码、处理文档、配置认证方式和错误处理，这不仅效率低下，也大大增加了开发和维护的成本。

而MCP的出现，如同在AI模型与外部世界之间搭建了一座标准化的桥梁。MCP以通用的“标准语言”把工具、数据通过“MCP服务”的方式供给（一次开发、无限连接），可以更高效、更便捷地实现Agent应用与成千上万的外部工具与数据的互通，极大提升了开发效率和灵活性。如需了解更多关于MCP的详细信息，请参考**MCP官方文档**。

MCP 分类

AgentArts服务提供了以下几种MCP创建方式，均支持在平台内便捷配置调用。您可以根据下表快速选择适合您的创建方式：

表 6-17 MCP 创建方式对比

创建方式	适用场景	特点	相关操作
自定义接入MCP	适用于接入开源社区的MCP或自主开发的MCP服务。当平台提供的MCP服务无法满足业务需求时使用。	灵活性高，需自行开发或维护。	接入方法请参考**自定义接入MCP**。
平台精选MCP	平台官方上架到资产广场的标准化MCP服务，可直接添加使用。	开箱即用，无需开发（部分需配置鉴权）。	接入方法请参考**基于模板创建MCP**。

- **自定义接入MCP**（空白创建）：适用于接入开源社区的MCP或自主开发的MCP服务。当平台精选MCP服务无法满足业务需求时，可通过自定义MCP服务接入。接入方法请参考**自定义接入MCP**。
- **平台精选MCP**：平台官方上架到资产广场的标准化MCP服务，用户可直接添加使用（部分MCP需要配置鉴权参数后才可使用），无需额外配置开发，实现快速接入调用。接入方法请参考**基于模板创建MCP**。

MCP 与插件的区别

AgentArts服务同时支持通过插件和MCP两种方式为智能体扩展工具能力，两者的核心区别如下：

版权所有 © 华为云计算技术有限公司

表 6-18 MCP 与插件的区别

对比维度	插件	MCP
接入协议	基于OpenAPI/自定义HTTP接口	基于MCP标准协议
配置方式	需逐个配置每个接口的地址、参数、响应	只需配置MCP服务地址，工具自动发现
工具使用	手动定义每个工具	MCP服务自动暴露可用工具列表
使用场景	对接已有的REST API接口	对接符合MCP协议的标准化服务

创建 MCP 前的准备

在开始创建之前，建议先确认以下信息：

- **明确需求：**智能体需要什么样的外部能力？（如网页抓取、联网搜索、数据处理等）
- **选择创建方式：**
  - 有现成的MCP服务地址（开源或自研）：**自定义接入MCP**
  - 平台精选中有满足需求的MCP：**基于模板创建MCP**
- **准备鉴权信息（如需要）：**部分MCP服务需要API Key、OAuth2.0等鉴权信息才能正常使用。

6.3.2 示例：使用模板创建搜索 MCP

本示例将介绍如何使用资产广场的预置MCP，并以必应搜索MCP为例介绍预置MCP的安装和使用方法。

前提条件

已**开通AgentArts服务**。

步骤一：安装预置 MCP

步骤1 登录**AgentArts智能体平台**。

步骤2 单击“资产广场”并进入“MCP”页签，搜索并找到一个必应搜索MCP服务（Bing）。

版权所有 © 华为云计算技术有限公司

图 6-49 获取搜索 MCP 服务



步骤3 鼠标指向MCP，单击“安装”，在弹出的MCP信息中**检查是否需要服务鉴权**：

一般平台展示的MCP配置，无论是表单编辑还是JSON编辑，均不需要改动，只需要关注是否填写鉴权信息即可。MCP是否需要鉴权取决于其所对接的后端服务本身的要求。

- 不需要鉴权：有些MCP服务本身是公开开放的，任何人都可以直接调用，自然无需鉴权，可直接安装使用。
- 需要鉴权：有些MCP服务涉及私有数据、权限控制，服务端要求调用方必须证明身份后才能访问，MCP就需要配置对应的鉴权信息，不配置则无法正常使用。需要鉴权的MCP服务要按照页面提示获取鉴权信息，并填写鉴权参数和取值。

版权所有 © 华为云计算技术有限公司

图 6-50 需要鉴权的 MCP 服务



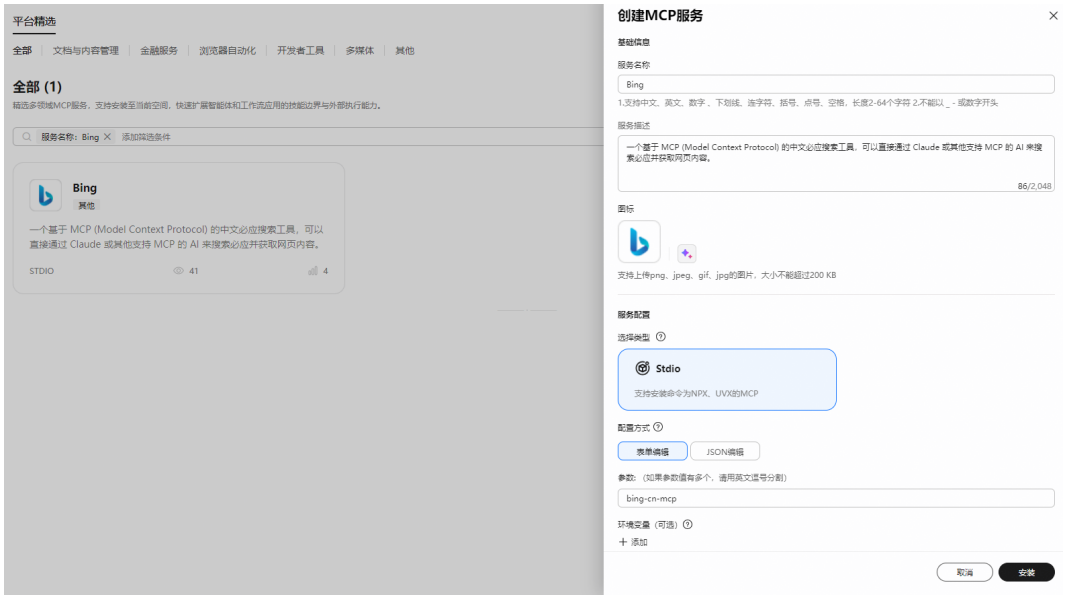
图 6-51 不需要鉴权的 MCP 服务



步骤4 本示例中所选取的必应搜索MCP无需鉴权，直接安装即可。

版权所有 © 华为云计算技术有限公司

图 6-52 安装必应搜索 MCP



步骤5 返回“组件库 > MCP”页面，可以查看已部署的MCP。

步骤6 单击部署好的MCP，进入到详情页，可查看该MCP下属的工具信息。单击“工具列表”中操作列的“调试”，根据MCP服务定义的参数结构填写调试用例。单击“运行”进行功能验证。

----结束

步骤二：调试 MCP

单击部署好的MCP，进入到详情页，可查看该MCP下属的工具信息。单击“工具列表”中操作列的“调试”，根据MCP服务定义的参数结构填写调试用例。单击“运行”进行功能验证。

步骤三：在单智能体中使用 MCP

步骤1 在“智能体管理 > 单智能体”页面单击“创建单智能体”。填写单智能体名称、描述为“问答助手”。

步骤2 设置提示词为**“回答问题前请先使用Bing MCP搜索用户问题”**（该提示词用于引导模型在回答问题前优先调用搜索MCP，以确保获取最新信息）；选择一个可用的模型；选择已经安装好的Bing MCP。

步骤3 在页面右侧输入测试问题“最近有什么热门新闻”等待智能体输出结果。

根据回答的内容即可判断出有调用MCP（如回答中是否包含实时信息），或单击“调试”按钮，选择近期会话，在“调用详情”中也可以看到调用了MCP。

如果智能体未调用搜索工具，请尝试调整提示词。

版权所有 © 华为云计算技术有限公司

图 6-53 查看调用详情



----结束

6.3.3 示例：快速接入 ModelScope 魔塔社区 MCP 工具集

本示例将介绍如何在AgentArts智能体平台中接入来自ModelScope魔塔社区的开源MCP工具集，通过接入MCP工具集，智能体可以调用外部工具能力（如网页抓取、联网搜索等），扩展智能体的功能边界。

前提条件

已**开通AgentArts服务**。

MCP 安装方式

AgentArts智能体平台支持Stdio（本地进程，含NPX、UVX）和HTTP（远程网络，含Streamable HTTP、SSE）两类连接模式。

图 6-54 MCP 安装方式



版权所有 © 华为云计算技术有限公司

Stdio（本地进程，含NPX、UVX）

这里的本地指的是AgentArts服务提供的运行环境，当选择**Stdio**方式安装MCP时，平台会在内部下载并启动这个MCP。

NPX：基于Node.js生态的MCP服务。

UVX：基于Python生态的MCP服务。

HTTP（远程网络，含Streamable HTTP、SSE）

远程指的是用户自己搭建的服务器MCP服务或者第三方提供的MCP服务。MCP服务代码并没有运行在AgentArts服务中，而是运行在别处（如通过本地电脑内网穿透暴露出来的地址、第三方MCP厂商提供的MCP API接口），平台仅作为客户端通过网络协议发起连接。

Streamable HTTP类型：基于标准HTTP协议的MCP连接方式，支持流式传输，是MCP协议推荐的远程连接方式，只需提供MCP服务的HTTP接口地址即可接入。

SSE类型：基于Server-Sent Events（服务器推送事件）协议的MCP连接方式，是较早期的远程连接方式，通过服务端主动向客户端推送数据实现实时通信，只需提供MCP服务的SSE接口地址即可接入。

接入MCP服务时，MCP提供方基本均会标注出MCP的安装方式，如果未明确写出，也可以从MCP的配置脚本中判断出来使用哪种方式进行安装。请选择合适的MCP安装方式。

图 6-55 MCP 服务安装方式示例



步骤一：接入 ModelScope 魔塔社区 MCP 工具集

步骤1 注册并登录**ModelScope魔塔社区**，进入“MCP广场”。选择“浏览器自动化”中的“Fetch网页内容抓取”MCP。

图 6-56 ModelScope MCP 广场



版权所有 © 华为云计算技术有限公司

步骤2 单击MCP名称，在详情页面可以查看MCP的配置信息。页面显示“Fetch网页内容抓取”MCP无需鉴权即可使用；对于需要鉴权的MCP，请按照提示获取对应的API Key。

图 6-57 查看 MCP 配置信息



如果使用Streamable HTTP方式接入MCP，单击“连接”按钮即可获取到配置脚本；Stdio方式同理，也可获取对应的配置脚本。

推荐优先选择Streamable HTTP方式，这是MCP协议推荐的远程连接方式，兼容性更好。仅当MCP提供方仅提供SSE端点时，才选择SSE方式。

图 6-58 获取 MCP 配置脚本

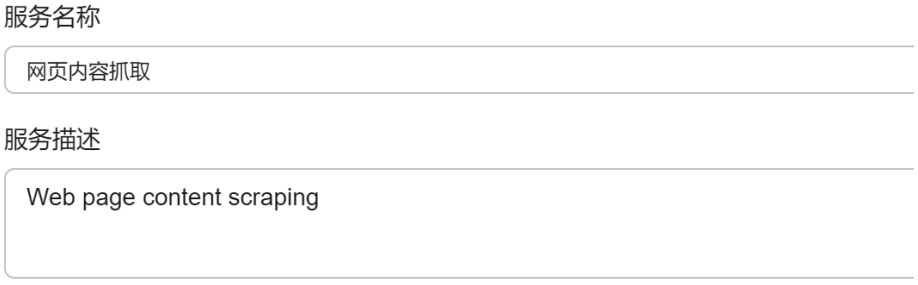


步骤3 登录**AgentArts智能体平台**。

步骤4 在“组件库 > MCP”页签，单击“创建MCP > 空白创建”。

步骤5 填写服务名称和服务描述（网页内容抓取、Web page content scraping）。

图 6-59 填写 MCP 名称与描述

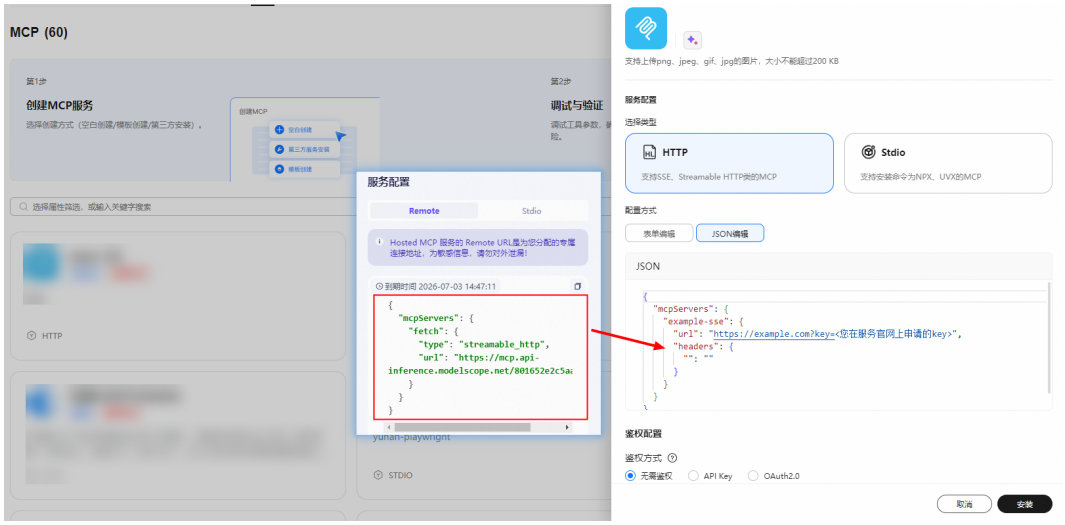


版权所有 © 华为云计算技术有限公司

步骤6 删除页面中的示例，安装方式根据复制的MCP配置脚本进行选择。本示例选择“HTTP”。配置录入方式选择“JSON编辑”。

步骤7 复制ModelScope魔塔社区获取的配置信息，填入至AgentArts平台中。单击“安装”等待MCP安装成功。

安装无问题后，MCP状态将显示为“部署成功”。



注意

如果MCP安装失败，多数原因是MCP本身的问题，如开源社区的MCP维护不及时或后端服务失效。遇到安装失败的情况，请尝试更换MCP来源。

----结束

步骤二：调试 MCP

步骤1 单击部署好的MCP，进入到详情页，可查看该MCP下属的工具信息。单击“工具列表”中操作列的“调试”，根据MCP服务定义的参数结构填写调试用例。单击“运行”进行功能验证。

由于本案例使用的是“Fetch网页内容抓取”MCP，调试时可以输入一个网址，调试成功后，将显示出具体的网页内容。

如果调试返回空内容，对应的网址可能设置了反爬虫机制，导致无法抓取到具体内容。

如果MCP工具解析失败，多数原因是MCP本身的问题，如开源社区的MCP维护不及时或后端服务失效。遇到工具解析失败的情况，请尝试更换MCP来源。

版权所有 © 华为云计算技术有限公司

图 6-60 调试 MCP



注意

- 在调试MCP时，请严格遵循参数格式要求，以避免出现错误。
- 使用“网页内容抓取”类MCP获取网页内容时，如目标网站设置了反爬虫防护（如robots限制、IP封禁、验证码等），可能导致抓取失败，这属于常见且正常的技术限制。

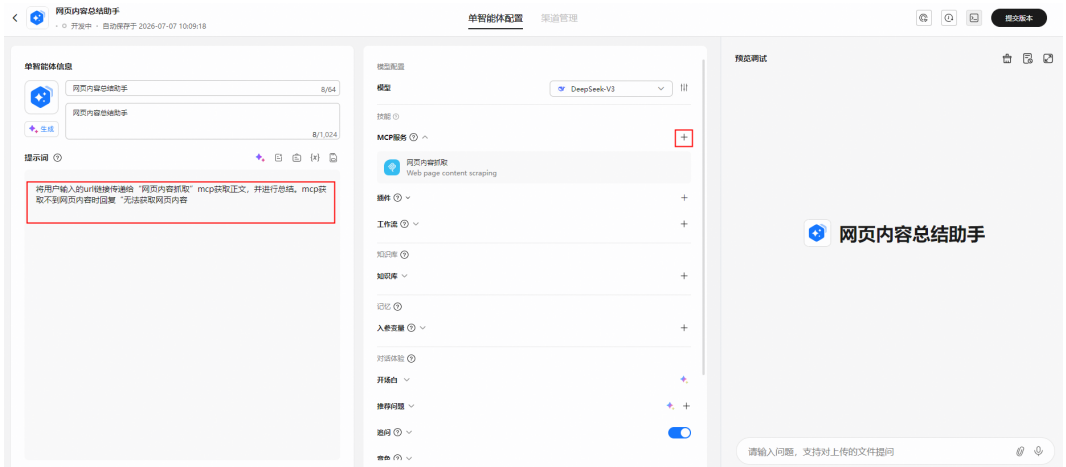
----结束

步骤三：在单智能体中使用 MCP

步骤1 在“智能体管理 > 单智能体”页面单击“创建单智能体”。填写单智能体名称、描述为“网页内容总结助手”。

步骤2 设置提示词为**将用户输入的url链接传递给“网页内容抓取”mcp获取正文，并进行总结。mcp获取不到网页内容时回复“无法获取网页内容”**；选择一个可用的模型；选择已经安装好的网页内容抓取MCP。

图 6-61 配置单智能体



版权所有 © 华为云计算技术有限公司

步骤3 在页面右侧输入测试链接，等待智能体输出结果。

图 6-62 测试单智能体



由于本案例使用的是“Fetch网页内容抓取”MCP，输入网址后，如果模型回复中包含了网页实际内容，即可判断出有调用MCP，或单击“调试”按钮，选择近期会话，在“调用详情”中也可以看到调用了MCP。

图 6-63 查看调用详情



----结束

6.3.4 创建 MCP

6.3.4.1 自定义接入 MCP

自定义接入是最灵活的MCP创建方式。您可以通过填写MCP服务的连接地址，将任意MCP服务接入平台供智能体调用。

该方式支持接入开源社区中的MCP（如GitHub、ModelScope上的开源MCP），也支持接入用户自主开发的MCP服务，适用于有特定业务系统对接需求、或需要使用社区开源MCP能力的场景。操作示例请参见**示例：快速接入ModelScope魔塔社区MCP工具集**。

费用说明

- 创建自定义MCP服务不涉及计费。
- 在调用自定义MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

版权所有 © 华为云计算技术有限公司

- 如果使用KMS对密钥进行加密存储，超出20,000次免费调用额度后按实际调用次数计费，详见**密码安全中心计费说明**。

前提条件

- 已**开通AgentArts服务**。
- 如果需要接入自主开发或私有部署的MCP服务，请确保该服务已完成部署并**可通过公网地址访问**。由于AgentArts为公有云服务，平台需要通过公网与MCP服务建立连接，无法直接访问内网或本地环境中的服务。
- 如果MCP服务需要鉴权，请提前准备好所需的鉴权信息（如API Key、OAuth 2.0的客户端ID和密钥等）。

操作步骤

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏选择“开发中心 > 组件库”，单击左上角“MCP”页签，进入MCP管理界面。

步骤3 在MCP管理界面，单击右上角的“**创建MCP > 空白创建**”创建MCP服务。

步骤4 在“空白创建”的弹框中，输入MCP服务的配置信息，参数说明请参考**表6-19**。

表 6-19 MCP 服务参数说明

参数	说明	示例
服务名称	MCP服务名称，用于区分不同的MCP服务实例，不会对大模型的判断和调用产生影响。 • 命名要求：仅支持以**中文**或**英文字母**开头。 • 支持字符：中文、英文、数字、中划线（-）、下划线（_）。 • 长度限制：2~64个字符。	网页内容抓取
服务描述	MCP服务的描述信息，帮助用户理解服务功能。例如，一个强大的MCP服务器，可以轻松地将网页内容抓取并转换为各种格式（HTML、JSON、Markdown、纯文本）。	该服务器使大型语言模型能够检索和处理网页内容，将HTML转换为markdown格式，以便于更轻松地使用。
服务图标	系统默认提供MCP服务图标。输入名称和描述后，单击 可以自动生成头像。用户也可以自定义图标，自定义图标步骤如下： 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。支持png、jpeg、gif、jpg格式图片，且图片大小不大于200KB。	默认

版权所有 © 华为云计算技术有限公司

参数	说明	示例

选择类型	选择安装MCP服务的类型。平台支持以下两类、四种MCP连接模式：
	如果您获取的是MCP源码，推荐使用Stdio模式；如果您获取的是已部署服务的API地址，请选择HTTP模式。
	● **Stdio（含NPX、UVX）**
	选择该方式时，平台会在内部自动下载并启动MCP服务，无需用户自行部署。其中NPX适用于Node.js生态的MCP服务，UVX适用于Python生态的MCP服务。
	● **HTTP（含Streamable HTTP、SSE）**
	选择该方式时，MCP服务运行在外部环境中（如用户自建服务器、第三方MCP厂商提供的接口等），平台作为客户端通过网络地址发起连接。其中Streamable HTTP是MCP协议推荐的远程连接方式，SSE是较早期的远程连接方式，两者均只需提供MCP服务的接口地址即可接入。
	平台提供两种配置录入方式，选择其中一种填写即可：
	● **表单编辑：**直接在表单中填写MCP服务的连接地址，适合配置信息简单的场景。在表单编辑模式下，除了填写连接地址外或参数，还需要根据安装方式的不同配置以下信息。
	环境变量：用于设置MCP服务运行时所需的变量参数，通常用来存放API Key、Token等敏感信息。
	请求头：用于在平台向MCP服务发起请求时附加自定义的HTTP头部信息。部分MCP服务要求在请求头中传递鉴权凭证或其他标识参数。
	● **JSON编辑**：以标准JSON格式填写完整的MCP配置。**多数MCP厂商会提供完整的JSON配置，可以直接复制使用，无需额外再转换成表单。**
	接入MCP服务时，MCP提供方通常会标注出MCP的安装方式，如果未明确写出，也可以从MCP的配置脚本中判断出来使用哪种方式进行安装。请选择合适的MCP安装方式。
		使用HTTP，JSON编辑模式。
		{
		"mcpServers": {
		"example-sse": {
		"url": "https://example.com?key=<您在服务官网上申请的key>"
		}
		}
		}

版权所有 © 华为云计算技术有限公司

参数	说明	示例
	图 6-64 各类 MCP 服务安装方式示例 	

版权所有 © 华为云计算技术有限公司

参数	说明	示例
鉴权配置	选择调用MCP时是否需要鉴权。 仅选择HTTP类型时需要配置鉴权信息。 ● **无需鉴权**：API可以公开访问，不需要任何形式的身份认证或授权。 ● **API Key**：在调用API时提供一个唯一的API Key进行鉴权。需配置以下信息：需填写密钥位置、密钥参数名称和取值。智能体在发起请求时，会自动把这个API Key添加到请求的Header（请求头）或Query（URL参数）中。 – 密钥位置：密钥是从Header中读取还是从Query中读取。 – 参数名称：API Key的鉴权参数名称。 – 参数值：API Key的具体值。 ● **OAuth2.0**：OAuth 2.0是一个开放授权协议，允许第三方应用在无需获取您账号密码的情况下，安全地获取您在其他服务中的资源访问权限。 – **授权服务器地址URL**：授权服务器的端点URL，用于发送授权请求和接收响应。平台会向该地址发送请求以获取Access Token。此地址由您所对接的API服务商提供 – **客户端id**：由API服务商分配的应用唯一标识，用来区分和识别不同的应用，通常在服务商的开发者后台创建应用后获得。 – **客户端密钥**：与客户端id配对的密钥，用于验证应用身份的合法性，确保只有被授权的应用可以使用。 – **请求的权限范围**：声明本次授权需要访问的资源范围，用于限制应用的访问权限。 **关于密钥加密：**当服务配置类型为**HTTP**时，鼠标单击“鉴权配置”右侧的 上，可查看密钥的加密方式。平台默认使用**KMS加密+默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 ● 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 ● 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持	无需鉴权

版权所有 © 华为云计算技术有限公司

参数	说明	示例
	缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	

步骤5 配置完成后，单击“安装”提交MCP部署任务。您可在“MCP 列表”中查看服务状态，如果显示“部署失败”，请单击“查看原因”排查问题后重试。

图 6-65 查看失败原因



----结束

6.3.4.2 基于模板创建 MCP

如果您不想手动配置MCP服务的连接地址和参数，可以直接使用平台精选的MCP模板。平台精选MCP是由官方筛选并预置到资产广场的标准化MCP服务，已完成适配和验证，您只需安装即可使用（部分MCP需要填写鉴权信息后方可正常调用），无需额外的配置开发。当您的需求恰好在平台精选的覆盖范围内，这是最快捷的接入方式。操作示例请参见**示例：使用模板创建搜索MCP**。

费用说明

- 安装平台预置MCP服务不涉及计费。
- 在调用预置MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

操作步骤

步骤1 登录**AgentArts智能体平台**。

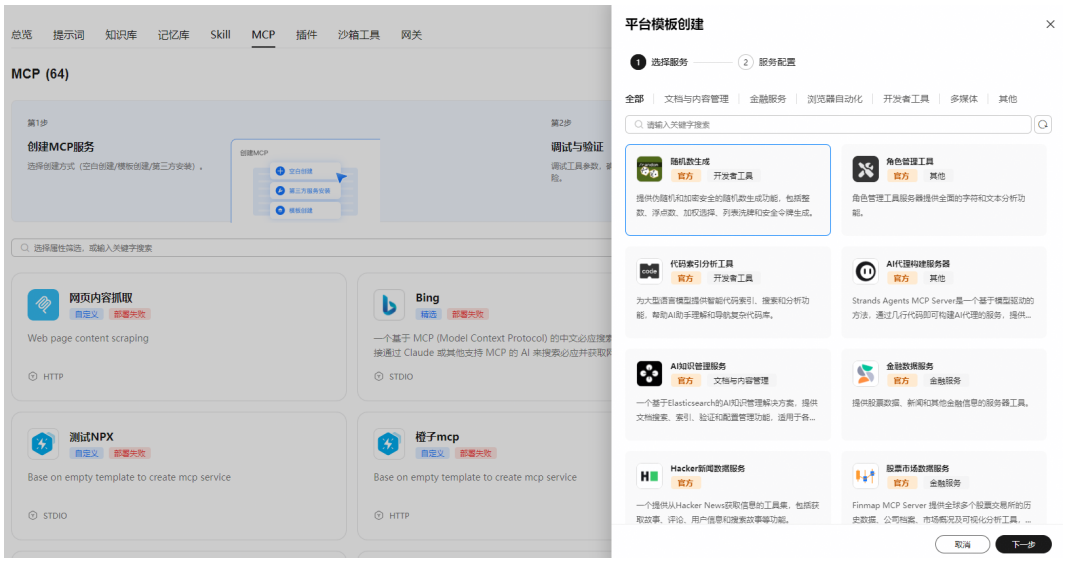
步骤2 在左侧导航栏选择“开发中心 > 组件库”，单击左上角“MCP”页签，进入MCP管理界面。

步骤3 在MCP管理界面，单击右上角的“**创建MCP > 平台模板创建**”创建MCP服务。

步骤4 平台提供了多种预置MCP服务，选择后单击“下一步”配置MCP信息。

版权所有 © 华为云计算技术有限公司

图 6-66 选择 MCP



步骤5 在MCP服务配置页签可以查看预置MCP的介绍，并对其进行修改。需要**重点关注MCP是否需要鉴权**。

一般平台展示的MCP配置，无论是表单编辑还是JSON编辑，均不需要改动，只需要关注是否填写鉴权信息即可。MCP是否需要鉴权取决于其所对接的后端服务本身的要求。如果鉴权信息填写错误，可能导致调用失败，请务必核对参数有效性。

- 不需要鉴权：有些MCP服务本身是公开开放的，任何人都可以直接调用，自然无需鉴权，可直接安装使用。
- 需要鉴权：有些MCP服务涉及私有数据、权限控制，服务端要求调用方必须证明身份后才能访问，MCP就需要配置对应的鉴权信息，不配置则无法正常使用。需要鉴权的MCP服务要按照提示要求获取鉴权信息，并填写鉴权参数和取值。

图 6-67 需要鉴权的 MCP 服务



版权所有 © 华为云计算技术有限公司

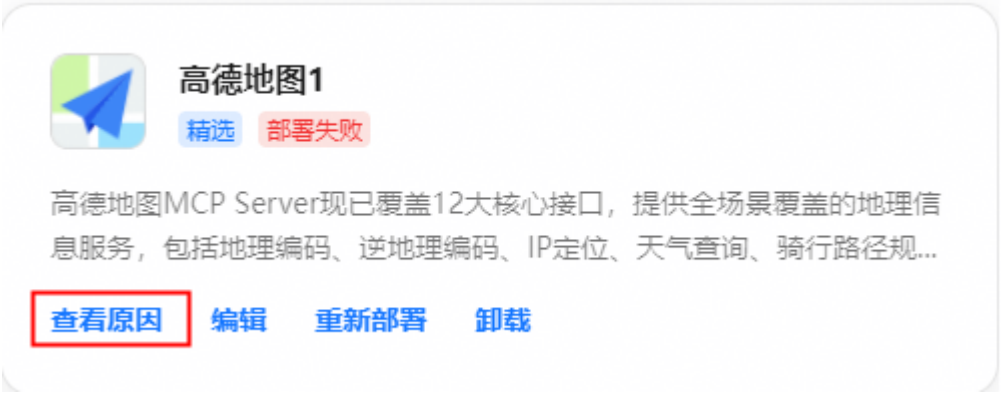
图 6-68 不需要鉴权的 MCP 服务



步骤6 配置信息填写并确认无误后，单击“安装”。

安装完成后，您可以在MCP列表中查看到新创建的MCP服务，状态显示为“部署成功”。如果部署失败可以单击“查看原因”，查看部署失败的原因。

图 6-69 查看原因



----结束

6.3.5 调试 MCP

MCP创建完成后，建议先进行调试验证，确认MCP服务连接正常、工具已正确识别、调用结果符合预期后，再将其关联到智能体中使用。调试阶段是发现和排查问题的最佳时机，直接跳过调试，将问题带入智能体运行环节，会大幅增加排查难度。

费用说明

调用和调试MCP服务可能产生费用。

- 平台预置的第三方MCP服务（如高德地图）可能需要您在对应平台申请API Key并承担相关调用费用，请确保您的账户额度充足。
- 智能体调用MCP工具会产生大模型Token消耗。

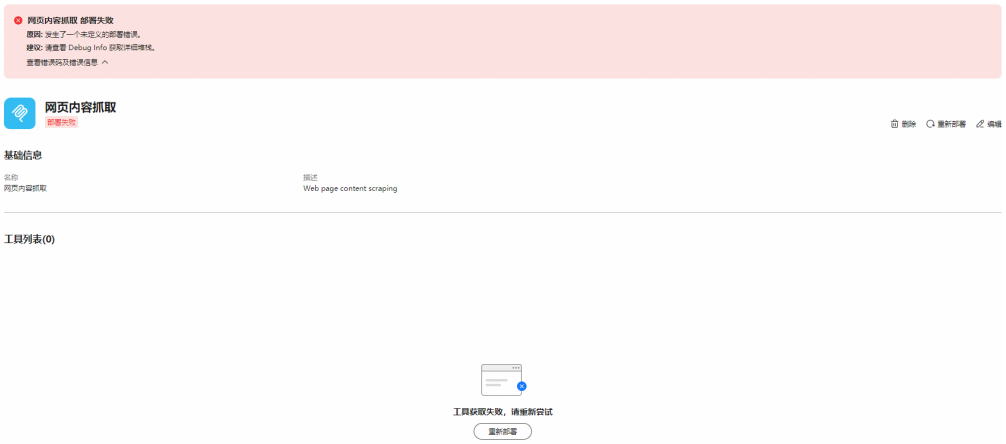
调试前确认

在开始调试之前，请先检查以下几项：

版权所有 © 华为云计算技术有限公司

- **工具是否已识别：**进入MCP详情页，在“工具列表”中确认工具已正确加载。如果工具列表为空，通常是MCP配置填写错误，或MCP服务本身异常（如开源社区的MCP维护不及时或后端服务失效）。遇到失败的情况，请尝试更换MCP来源。

  图 6-70 工具识别异常示例



- **鉴权信息是否已配置：**如果MCP服务需要鉴权，请确认已在创建时正确填写鉴权参数，否则调试时会返回鉴权失败的错误。
- **服务是否可正常访问：**对于Streamable HTTP、SSE类型的MCP，请确认服务地址可通过公网正常访问。

调试 MCP

步骤1 MCP部署成功后，单击MCP名称并进入详情页，可以查看MCP具有的工具，并进行调试。

调试前先查看每个工具的名称、描述和参数定义是否完整且准确。这些信息直接影响智能体后续能否正确理解和调用该工具。

如果发现工具描述缺失或参数定义不完整，需要在MCP服务端进行修正（或尝试重新部署MCP）。

参数格式不正确是调试中最常见的报错原因。调试时请仔细查看每个参数的类型和格式说明，按照要求填写。详细格式问题请参考**MCP调试故障**，常见格式问题示例：

错误写法：
```
url: example.com
date: 2025年1月15日
tags: 科技,金融
age: "25"
enabled: "true"
filter: {name: "张三"}
language: ZH（可选值为 zh/en/ja）
```

正确写法：
```
url: https://www.example.com
date: 2025-01-15
tags: ["科技", "金融"]
age: 25
enabled: true
filter: {"name": "张三"}
language: zh
```


版权所有 © 华为云计算技术有限公司

图 6-71 调试 MCP



----结束

6.3.6 使用 MCP

6.3.6.1 在单智能体中使用 MCP

MCP创建并调试通过后，即可将其添加到智能体中使用。平台支持在单智能体和工作流两种模式下使用MCP，工作流中使用方法请参见**在工作流中使用MCP**。两者的核心差异在于工具的调用方式不同：

表 6-20 单智能体、工作流使用 MCP 差异

对比维度	在单智能体中使用	在工作流中使用
调用决策	由大模型自主判断是否调用、调用哪个工具	在指定的流程节点中固定调用
参数来源	大模型从对话上下文中自动提取参数	在节点中预先配置，或引用上游节点的变量
执行顺序	不固定，由大模型根据对话动态决定	固定，按照编排好的流程顺序执行
使用场景	用户意图多样、需要灵活响应的对话场景	流程明确、步骤固定、需要精确控制的场景

简单理解：单智能体模式下，MCP工具是大模型的“自选动作”；工作流模式下，MCP工具是流程中的“规定动作”。

费用说明

调用MCP服务可能产生费用。

- 平台预置的第三方MCP服务（如高德地图）可能需要您在对应平台申请API Key并承担相关调用费用，请确保您的账户额度充足。

版权所有 © 华为云计算技术有限公司

- 智能体调用MCP工具会产生大模型Token消耗。

注意事项

在单智能体模式下，MCP的调用完全由大模型自主决策。大模型会根据用户的输入内容，结合工具的名称和描述信息，自行判断是否需要调用工具、调用哪个工具、以及从对话上下文中提取什么参数。整个过程无需人为干预，但也意味着工具能否被准确调用，高度依赖于配置细节。以下是关键注意事项：

- **在提示词中明确工具的使用场景和规则**

  这是提升MCP工具调用准确率最有效的方式。不要假设大模型能自动理解何时该用哪个工具，而是在提示词中明确写出每个工具的使用时机、调用条件和使用规则：

  ```
  示例：
  # MCP使用规则
  1.【天气查询MCP】
    - 使用时机：当用户询问某个城市的天气、气温、是否下雨等天气相关问题时
    - 注意事项：需要从用户对话中提取城市名称作为参数，如果用户未提及具体城市，先追问
  2.【网页抓取MCP】
    - 使用时机：当用户提供了一个网页链接并要求你获取、分析或总结该网页内容时
    - 注意事项：用户必须提供完整的 URL（含 https://），如果用户只给了域名，先引导补全
  重要：
  - 如果用户的问题可以直接回答，不要调用MCP
  - 不确定使用哪个MCP时，优先通过对话澄清用户需求
  - 每次只调用一个MCP，等返回结果后再决定下一步
  ```

- **确保MCP描述清晰准确**

  大模型除了参考提示词外，还会依赖MCP自身的名称和描述来判断用途。如果MCP服务提供的描述过于简略或含糊，可能导致大模型理解偏差。

- **引导大模型在调用前做信息确认**

  MCP调用的参数由大模型从对话中自动提取，如果用户表述模糊或信息不全，可能导致参数提取错误。建议在提示词中要求大模型在调用MCP前先确认关键信息：

  ```
  # 调用MCP前的确认规则
  - 调用MCP前，先确认已从用户对话中获取到所有必填参数
  - 如果用户提供的信息不完整或有歧义，先通过对话追问确认，不要猜测填写
  ```

- **控制MCP数量**

  挂载的MCP过多会增加大模型的选择难度，容易出现选错MCP或犹豫不决的情况。建议单个智能体关联的MCP总数不超过5个。如果业务确实需要更多MCP，可考虑拆分为多智能体架构，让不同子智能体各自负责特定领域的工具。

- **调试验证MCP调用效果**

  单智能体配置完成后，建议用不同类型的测试问题验证工具调用是否符合预期：

  ```
  测试要点：
  1 应该触发MCP调用的问题 → 是否正确调用了对应的MCP?
  2 不应该触发MCP调用的问题 → 是否没有误调用?
  3 信息不全的问题 → 是否先追问而不是直接调用?
  4 多个MCP都可能匹配的问题 → 是否选择了正确的那个?
  ```

在单智能体中使用 MCP

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏选择“开发中心 > 智能体管理”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

 版权所有 © 华为云计算技术有限公司

步骤3 单击目标单智能体应用，在“单智能体配置”页签中，单击“MCP服务”右侧的+。

步骤4 在“添加MCP服务”弹框中，选择需要的MCP服务，单击“确定”。

步骤5 在提示词区域，设置MCP的调用规则。并在右侧调试区域输入问题验证MCP是否触发成功。

如果调试区域返回结果中包含工具调用的过程信息或正确结果，说明配置成功。如果在单智能体中调用失败，请参考**故障排除**。

图 6-72 在单智能体中使用 MCP



----结束

6.3.6.2 在工作流中使用 MCP

MCP创建并调试通过后，即可将其添加到工作流使用。平台支持在单智能体和工作流两种模式下使用MCP，单智能体中使用方法请参见**在单智能体中使用MCP**。两者的核心差异在于工具的调用方式不同：

表 6-21 单智能体、工作流使用 MCP 差异

对比维度	在单智能体中使用	在工作流中使用
调用决策	由大模型自主判断是否调用、调用哪个工具	在指定的流程节点中固定调用
参数来源	大模型从对话上下文中自动提取参数	在节点中预先配置，或引用上游节点的变量
执行顺序	不固定，由大模型根据对话动态决定	固定，按照编排好的流程顺序执行
使用场景	用户意图多样、需要灵活响应的对话场景	流程明确、步骤固定、需要精确控制的场景

简单理解：单智能体模式下，MCP工具是大模型的“自选动作”；工作流模式下，MCP工具是流程中的“规定动作”。

费用说明

调用MCP服务可能产生费用。

版权所有 © 华为云计算技术有限公司

- 平台预置的第三方MCP服务（如高德地图）可能需要您在对应平台申请API Key并承担相关调用费用，请确保您的账户额度充足。
- 智能体调用MCP工具会产生大模型Token消耗。

注意事项

在工作流模式下，MCP作为流程中的一个固定节点被调用，何时调用、传什么参数都在编排时预先配置好，不依赖大模型的自主判断。相比单智能体模式，工作流的工具调用更加确定和可控，但需要在搭建时注意以下几点：

- **参数配置要正确：**
  MCP的输入参数需要在节点中手动配置，可以填写固定值，也可以引用上游节点的变量。请确保参数的类型和格式与MCP要求一致（如MCP要求输入为String类型，被引用的节点参数也要是String类型，并且格式也要保持一致）。
- **注意节点间的数据传递：**
  MCP节点的输出结果如需在后续节点中使用，需确认变量引用关系正确，数据格式能够被下游节点正常解析。
- **做好异常分支处理：**
  MCP调用可能因网络超时、参数错误、服务异常等原因失败，建议在工作流中配置异常处理分支，避免整个流程因单个节点报错而中断。

在工作流中使用 MCP

步骤1 登录AgentArts智能体平台。

步骤2 在左侧导航栏选择“开发中心 > 智能体管理”，单击“工作流”页签，进入工作流应用管理界面。

步骤3 单击目标工作流应用，进入工作流编排界面。

步骤4 在工作流编排界面中，单击“添加节点”并选择“MCP服务”节点。将MCP节点添加到工作流中。

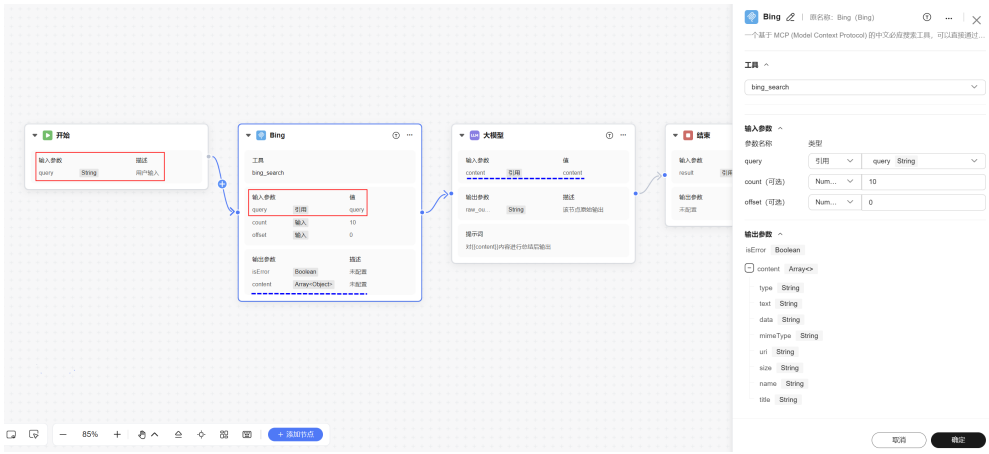
图 6-73 添加 MCP 节点



步骤5 在节点配置面板的“输入参数”区域，根据业务需求填写固定值，或选择引用变量并从下拉列表中选择上游节点的输出，设置MCP后续节点的输入，将MCP的结果传递给后续节点。

版权所有 © 华为云计算技术有限公司

图 6-74 设置 MCP 节点参数



步骤6 单击页面右上角“试运行”验证工作流是否可以正常运行。

如果试运行时返回结果中包含工具调用的过程信息或正确结果，说明配置成功。如果在工作流中调用失败，请参考**故障排除**。

----结束

6.3.7 管理 MCP

MCP服务创建完成后，您可以在组件库中对其进行全生命周期管理。例如，当服务配置变更时需编辑配置，或重新部署。您可以通过服务名称或搜索关键字功能来查找MCP服务。也可根据需要执行**表6-22**所示的操作。

表 6-22 更多操作

操作	步骤
查看原因	如果MCP部署失败，可将鼠标悬停在MCP服务卡片上并单击“查看原因”，查看具体的部署失败原因。
编辑MCP配置信息	1. 在“MCP”页签中，单击目标MCP服务卡片进入详情页。 2. 单击页面右上角的“编辑”，即可修改MCP服务的配置信息。 3. 修改完成后，单击“安装”重新安装MCP服务。安装成功后，该MCP服务卡片状态将变为“部署成功”。 **说明** 仅MCP服务部署失败时，支持编辑配置信息。
重新部署	如果MCP部署失败需要重新部署，将鼠标悬停在MCP服务卡片上，并单击“重新部署”，即可重新部署MCP服务。部署成功后，该MCP服务卡片状态将变为“部署成功”。

版权所有 © 华为云计算技术有限公司

操作	步骤
删除MCP服务	删除后，MCP服务将被移除，请谨慎操作。 1. 在“MCP”页签中，将鼠标悬停在需要卸载的MCP服务卡片上。 2. 单击“删除”，确认删除后，该MCP服务卡片将从列表中移除。 **注意** 如果该MCP服务已被引用，删除后将自动取消引用关系，这可能会导致智能体无法运行，且该操作不可撤销，请谨慎操作。

6.3.8 故障排除

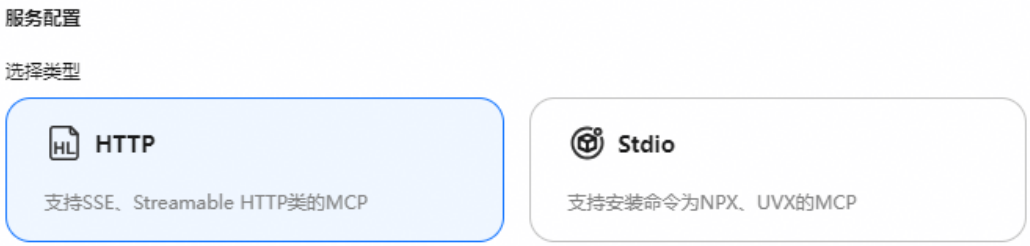
当MCP安装失败时，平台会返回特定的错误码以及错误信息提示。针对服务创建过程中可能遇到的网络连通性、环境配置冲突及协议错误，本章节提供了典型错误码的诊断流程与解决方案。

前置检查：确认 MCP 安装方式

多数MCP服务安装失败源于MCP的安装方式选择错误。在您深入排查网络问题、配置问题前，请先确认MCP的安装方式是否选择正确。

平台支持Stdio（本地进程，含NPX、UVX）和HTTP（远程网络，含Streamable HTTP、SSE）两类连接模式。

图 6-75 MCP 安装方式



Stdio（本地进程，含NPX、UVX）

这里的本地指的是AgentArts服务提供的运行环境，当选择NPX、UVX方式安装MCP时，平台会在内部自动从npm仓库/PyPI下载并启动MCP服务。

表 6-23 NPX、UVX 安装 MCP 差异

方式	适用生态	MCP来源
NPX	Node.js生态的MCP服务	npm仓库自动下载
UVX	Python生态的MCP服务	PyPI自动下载

HTTP（远程网络，含Streamable HTTP、SSE）

版权所有 © 华为云计算技术有限公司

远程是指MCP服务未部署在AgentArts平台内，而是运行在外部环境（如您的本地服务器、第三方云服务）中，AgentArts平台通过网络协议远程连接这些服务，平台仅作为客户端通过网络协议发起连接，MCP服务需可通过公网访问。

表 6-24 Streamable HTTP、SSE 安装 MCP

方式	说明	推荐程度
Streamable HTTP	基于标准HTTP协议，支持流式传输，是MCP协议官方推荐的远程连接方式。只需提供MCP服务的HTTP接口地址即可接入。	推荐（新建服务首选）
SSE	基于Server-Sent Events（服务器推送事件）协议，是较早期的远程连接方式，通过服务端主动向客户端推送数据实现实时通信。只需提供MCP服务的SSE接口地址即可接入。	仅兼容存量旧版MCP服务，新建服务不推荐

接入MCP服务时，MCP提供方通常均会标注出MCP的安装方式，如果未明确写出，也可以从MCP的配置脚本中判断出来使用哪种方式进行安装。请选择合适的MCP安装方式。

图 6-76 各类 MCP 服务安装方式示例



MCP 调试故障

调试MCP工具时，请严格按照参数的格式要求填写输入值。参数格式不正确是调试中最常见的报错原因。建议在调试前先查看每个参数的类型、格式说明及示例，确保输入值符合要求。

以下是几类常见的参数格式错误及正确写法：

示例一：日期/时间参数

部分MCP工具要求传入日期或时间参数，通常需要遵循特定格式：

错误写法：
date: 2025年1月15日
date: 01-15-2025
date: 2025/01/15

版权所有 © 华为云计算技术有限公司

```
正确写法（需要根据MCP的实际要求填写）：
date: 2025-01-15              （日期格式：YYYY-MM-DD）
datetime: 2025-01-15T10:30:00Z （日期时间格式：ISO 8601）
```

示例二：数组/列表参数

部分工具的参数需要传入多个值（数组格式）：

```
错误写法：
tags: 科技,金融,医疗
tags: 科技、金融、医疗
tags: "科技" "金融" "医疗"

正确写法：
tags: ["科技", "金融", "医疗"]
```

常见报错：Expected array、got string或参数类型不匹配。

原因：数组参数需要使用JSON数组格式 ["值1", "值2"]，而非逗号分隔的普通文本。

示例三：JSON对象参数

部分工具需要传入结构化的JSON对象：

```
错误写法：
filter: name=张三, age=25
filter: {name: 张三, age: 25}

正确写法：
filter: {"name": "张三", "age": 25}
```

常见报错：JSON parse error或Invalid JSON format。

原因：JSON格式要求键名必须用英文双引号包裹，字符串值也需要双引号，不能使用单引号或省略引号。

示例四：枚举/固定可选值参数

部分参数只接受预定义的几个固定值：

```
参数说明：language（语言），可选值：zh / en / ja

错误写法：
language: 中文
language: chinese
language: ZH

正确写法：
language: zh
```

常见报错：Invalid enum value或参数值不在允许范围内。

原因：枚举参数必须严格使用规定的可选值，注意大小写通常也需要完全匹配。

快速排查清单：

```
参数值的类型是否正确（字符串/数字/布尔值/数组/对象）？
字符串值是否需要加引号?
日期格式是否符合 ISO 8601 标准（YYYY-MM-DD）？
数组参数是否使用了 JSON 数组格式 ["值1", "值2"]?
JSON 对象的键名是否用了英文双引号?
枚举参数的值是否与可选项完全一致（含大小写）？
必填参数是否全部填写?
参数值中是否包含多余的空格或特殊字符?
```

智能体中使用 MCP 故障

以下故障通常由智能体配置不当引起，用户可通过调整提示词、MCP描述解决。

 版权所有 © 华为云计算技术有限公司

表 6-25 故障排查说明

故障现象	可能原因	排查方法
用户明确提出需求后智能体未调用MCP	MCP描述不够清晰，大模型无法关联。	优化MCP的名称和描述，使其更明确。例如填写：获取天气信息，根据城市名称查询当前天气信息，包括温度、湿度、风力等。
	提示词中缺少MCP使用引导。	在提示词中明确说明MCP的使用场景和时机。例如添加：当用户询问天气信息时，请使用获取天气信息MCP查询。
	MCP未正确关联到智能体。	检查智能体配置中是否已添加对应的MCP。如果已添加优先检查提示词中是否指明了MCP的应用场景。如果仍然无法触发，检查MCP本身是否可以正常使用，详见**调试MCP**。
智能体选错了MCP	多个MCP描述相似，大模型混淆。	差异化各MCP的描述，突出各自的使用场景。
	挂载的MCP过多。	减少MCP数量，建议不超过5个。挂载MCP过多会导致大模型在工具选择时产生混淆，降低调用准确率。如需使用更多工具，建议合并功能相近的MCP服务。
	提示词中未区分不同MCP的使用条件。	在提示词中为每个MCP写明具体的使用场景。
智能体对MCP返回结果的解读有误	返回数据字段含义不明确。	在提示词中说明如何解读MCP返回的数据。例如：MCP返回的JSON中，temp字段表示摄氏温度，humidity字段表示湿度百分比。

MCP 安装故障

当MCP安装失败时，平台会返回特定的错误码与错误信息。请按照思路进行初步诊断：

AgentArts内置了一套智能侦测机制。了解这个机制，能帮助您在面对复杂报错（如“超时”或“冲突”）时，快速判断问题是出在物理网络层还是应用配置层。

系统在连接您的MCP服务时，会依次执行两步探测：

- 第一步 (Scout侦察)：使用TCP协议尝试与MCP服务建立连接。
- 第二步 (SDK握手)：连接建立后，使用HTTP协议发送数据。

基于此机制，各错误码对应的失败阶段如下：

版权所有 © 华为云计算技术有限公司

表 6-26 失败阶段说明

失败阶段	对应错误码	错误特征
第一步（TCP侦察）失败	1173、1161、1162	连接尚未建立即报错，问题在网络层
第二步（HTTP/SDK握手）失败	1164、1150、1163	TCP通但HTTP层被拦截或证书异常，问题在应用层

在排查安装故障前，请先了解报错的处置归属：

- 用户可自行修复：配置填写错误、URL路径错误这类配置类问题。
- 需更换MCP服务：第三方MCP封禁了华为云IP段、WAF拦截等，用户无法干预对方服务策略。
- 需联系服务技术支持解决：平台侦测机制异常、错误码无法匹配已知原因等平台侧问题。

不同场景需要重点关注的错误码不同，问题解决后重新创建MCP，如果MCP显示“部署成功”则表示问题已解决：

表 6-27 错误排查指引

场景	重点关注的错误码	首要排查方向
使用第三方MCP厂商服务	1161、1164	网络连通性、WAF/CDN拦截
本地自建MCP通过内网穿透暴露	1173、1161、1162	域名解析、公网可达性、端口监听
公司内网部署MCP	1164、1150	代理策略拦截、证书问题

版权所有 © 华为云计算技术有限公司

表 6-28 MCP 安装错误码说明

错误码	报错场景	报错信息	处置归属	排查与解决思路
AgentArts.02401173	域名错误，无法解析目标主机地址	AgentStudioException: Unknown Host: mcp-unknown-host.local	用户可修复/需更换MCP	**DNS解析被拦截** • **现象：**错误信息显示Unknown Host。 • **原因：**系统在解析MCP的域名IP阶段就失败了，根本没发出网络包。 • **解决思路：** – 创建MCP服务时：检查在AgentArts填写的MCP地址是否正确，URL是否存在拼写错误。  – 连接第三方MCP时：对方可能封禁了云厂商的IP段（地理围栏）。建议更换相同功能的MCP。 – 连接本地自建MCP时：自建MCP服务必须绑定公网IP(EIP) 或使用内网穿透域名。本地MCP服务器的安全组/系统防火墙没有放行端口，或者您没有配置公网IP，或者是否填写了.local或k8s-svc等内部域名？注意，AgentArts使用公网DNS，无法解析私有域名，导致AgentArts无法连通MCP，请勿在MCP地址中填写仅内网可用的域名。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	处置归属	排查与解决思路
AgentArts.02401161	连接超时，网络物理隔离或防火墙丢包	Connect timed out (2s limit) 或 TimeoutException... within 120000ms	用户可修复/需更换MCP	**物理网络直接断连** ● **现象：**错误信息显示Connect timed out (2s limit) ● **原因：**系统连TCP握手都无法完成，报错连接超时。 ● **解决思路：** – 连接第三方MCP时：对方可能封禁了云厂商的IP段（地理围栏）。建议更换相同功能的MCP。 – 连接本地自建MCP时：MCP服务器是否因为安全组/系统防火墙原未放行端口，或者没有配置公网IP。
AgentArts.02401162	连接被拒绝，IP可达但端口不通	Connection Refused: Connection refused: no further information	自建MCP问题，需联系供应方	**连接被拒绝** ● **现象：**错误信息显示Connection refused ● **原因：**系统成功触达了MCP服务，但请求的URL路径在服务端不存在。 ● **解决思路：** – 连接本地自建MCP时：检查代码中是否固定监听地址，导致只接受服务器内部请求，拒绝外部AgentArts请求，需要将地址改为公网IP或内网穿透域名。 – 创建MCP服务时：检查协议端口是否匹配，HTTP默认端口80，HTTPS默认端口443。请勿将HTTPS协议强行指定到非加密端口。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	处置归属	排查与解决思路
AgentArts.02401163	路径错误 (404) 服务通但路径不对	HTTP 404 Not Found: http://.../wrong-path	自建MCP问题，需联系供应方	**配置错误** ● **现象：**错误信息显示HTTP 404 Not Found。 ● **原因：**系统成功触达了MCP服务，但MCP显式地反馈了拒绝信息。 ● **解决思路：** – 检查在AgentArts填写的MCP地址是否正确，URL是否存在拼写错误。检查URL的尾部。例MCP规范要求是/sse ，是否写成了/chat？尝试在浏览器直接访问该URL，确认是否能看到返回结果。 – 检查创建MCP时是否漏填参数，或者参数拼写错误。
AgentArts.02401164	代理/认证冲突，平台智能推断出配置错误	TimeoutException... (但错误码为1164) 或Conflict detected	自建MCP问题，需联系供应方	**应用层被拦截** ● **现象：**错误信息显示TimeoutException且时间较长（如60s/120s），但错误码判定为1164。 ● **原因：**系统发现TCP连接成功了（说明IP、端口、防火墙都是通的），但在发送HTTP请求时被阻塞或超时。 ● **解决思路：** – 连接第三方MCP时：极大概率是对方的WAF (Web防火墙) 或CDN将AgentArts的请求视为爬虫/攻击流量进行了拦截（TCP握手放行，HTTP内容拦截）。建议更换相同功能的MCP。 – 连接公司内网搭建的MCP时：公司边缘网关的Proxy (代理) 策略拦截了该请求，但未返回标准的拒绝状态码。检查代理配置，确保目标地址已加入NO_PROXY列表，或允许来自AgentArts的HTTP持久连接。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	处置归属	排查与解决思路
AgentArts.02401150	SSL/证书错误或其他通用网络错误	Probe Error: PKIX path building failed	自建MCP问题，需联系供应方	**安全协议与证书故障** ● **现象：**报错信息显示PKIX path building failed或Handshake error。 ● **原因：**AgentArts与MCP之间网络是通的，但双方在建立加密连接（SSL/TLS）时失败。 ● **解决思路：**重点排查HTTPS证书的有效性。由于AgentArts默认不信任未经CA机构认证的自签名证书，无法信任非权威机构签发的证书。请根据您的场景采取以下对策： – 建议在测试阶段将协议由https:// 临时改为http://。此操作可绕过证书校验逻辑。此方法仅限本地测试环境，用于临时验证网络连通性，严禁在生产环境中使用。测试完成后请立即恢复为https://并修复证书问题。如需临时使用http，请确认服务端已开启非加密端口（通常为80或自定义端口）的监听。 – 如果不确定证书状态，请在本地浏览器中直接访问该MCP的URL。如果地址栏弹出“您的连接不是私密连接”或“不安全”警告，则确认是证书问题。建议对自建MCP服务的证书进行更新，如更新为Let's Encrypt等正式的受信任证书。

6.4 知识库

版权所有 © 华为云计算技术有限公司

6.4.1 知识库介绍

功能概述

知识库是组织、存储及管理知识的系统，涵盖文档、图片等信息的分类整理，可以帮助用户高效管理大量的信息。在智能体中添加知识库，使其能够利用知识库中的信息与用户进行交互，可以显著提升智能体的准确度和专业度。

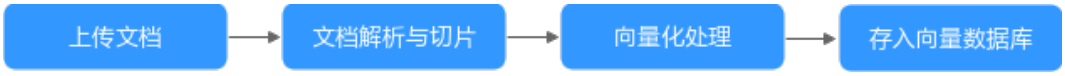
AgentArts支持管理平台知识库和第三方知识库。

平台知识库功能对文本文档、FAQ（Frequently Asked Questions，常见问题解答）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。

说明

向量化是将文字转换为数字序列的过程，让计算机通过计算“数字指纹”的相似度，实现基于语义的精准搜索。

图 6-77 文档处理流程



知识库支持导入如表6-29所示的本地文档：

表 6-29 支持的文档格式

文档类型	文档格式及限制
知识文档	• 支持上传常见的文档格式： – 文档/文本类：doc、docx、pdf、ofd、txt、md、wps。 – 表格/数据类：xls、xlsx、csv、et。 – 演示/幻灯片类：ppt、pptx、dps。 – 图片/设计类：jpg、jpeg、png、gif、bmp、webp、tiff、tif、psd、ico、pcx。 AgentArts支持用户根据需求，将图片类的文件直接存储，或是提取图片中的文字信息进行存储。 • 不同套餐包对文件大小限制不同，具体请参考知识库使用限制。
FAQ	• 支持按照模板上传文件，文件格式仅支持xlsx、xls、docx、doc。 • 不同套餐包对文件大小限制不同，具体请参考知识库使用限制。 • Excel单个文件最大支持100,000条数据。 • FAQ文件中不允许空行，空行后的数据将被忽略。

版权所有 © 华为云计算技术有限公司

费用说明

- 平台知识库：不同套餐包的**知识库存储空间容量**不同，具体可参考**计费说明**。
- 第三方知识库：AgentArts侧不计费，但第三方系统的部署与运维费用需用户自行承担，具体计费以第三方为准。

应用场景

- **文旅服务场景**：旅游平台可以在知识库中上传景区介绍、游玩路线、周边餐饮和住宿信息、交通时刻表、门票价格及优惠政策，以及当地特色体验项目。当游客查询“周末亲子游景区推荐”时，智能体会结合适合的年龄段和游玩时长等信息生成定制化方案；当查询“景区门票如何预约”时，会直接调取预约要求与流程。
- **科研辅助场景**：科研人员可以在知识库中上传中英文文献摘要、实验数据记录、参考文献目录和行业标准与研究方法。撰写论文时，智能体通过向量召回相关文献摘要，提取核心结论与数据支撑；设计实验方案时，快速调取同类实验的参数设置和研究方法，对比优劣并优化实验设计，提升研究效率。
- **智能问答助手场景**：企业可以将产品文档、产品使用手册、用户高频咨询的问题合集、产品参数规格表、常见故障排查指南等信息上传至知识库。智能体能够通过精准匹配知识库中的相关信息准确解答用户的问题。

知识与记忆对比

知识：是指Agent通过数据输入获得的信息和理解。这些信息可以是非结构化的（如文档、文章）。

记忆：是指Agent在与用户交互过程中存储和检索的特定信息。这些信息通常是与特定用户或会话相关的，用于提供个性化的服务和上下文相关的响应。

相同点

- 信息存储

  知识和记忆都涉及信息的存储。无论是知识库中的数据还是会话中的临时信息，都是为了在需要时能够快速访问和使用。

- 支持决策

  知识和记忆都支持Agent在与用户交互时做出更准确和有效的决策。知识提供广泛的信息基础，而记忆提供具体的上下文信息。

- 动态性

  知识和记忆都是动态的，需要定期更新和维护。知识库需要定期更新以保持信息的准确性和时效性，而记忆需要在会话过程中不断更新以反映用户的最新交互。

- 提高用户体验

  知识和记忆都旨在提高用户体验。知识库提供全面的信息支持，帮助用户快速找到答案；记忆则提供个性化的服务，使交互更加自然和连贯。

不同点

- 存储方式

  知识：通常存储在外部介质中，如数据库、知识图谱、文档等。这些信息是结构化或非结构化的，可以长期保存和共享。

  记忆：通常存储在Agent的内部状态中，如会话状态、用户状态等。这些信息通常是临时的，会随着会话的结束或用户的注销而清除。

 版权所有 © 华为云计算技术有限公司

- 信息类型

  知识：涵盖广泛的信息，包括事实、原理、技能、方法等。这些信息是通用的，可以应用于多个用户和场景。

  记忆：涉及特定用户或会话的信息，如用户的查询历史、偏好、上下文信息等。这些信息是个性化的，特定于某个用户或会话。

- 更新频率

  知识：需要定期更新和维护，以保持信息的准确性和时效性。更新频率可能较低，但每次更新都涉及大量的信息。

  记忆：在每次交互过程中不断更新，以反映用户的最新行为和需求。更新频率较高，但每次更新涉及的信息量较小。

- 传递性

  知识：知识库可以共享给多个Agent，提高整体的服务质量。

  记忆：通常是特定于某个用户或会话的，不直接传递给其他用户或会话。记忆信息的传递通常需要通过用户标识和会话标识来实现。

- 应用场景

  知识：广泛应用于常见问题解答、产品信息、政策和法规、最佳实践等。这些信息是通用的，适用于多个用户和场景。

  记忆：主要应用于多轮对话、个性化推荐、用户识别、上下文理解等。这些信息是个性化的，特定于某个用户或会话。

- 个体差异

  知识：不同Agent的知识水平和获取途径可能不同，但知识本身是客观的，适用于所有用户。

  记忆：不同用户的记忆内容和方式存在显著差异，受用户的特定需求和历史行为影响。

选型建议

类型	适用场景	优势	劣势
平台知识库	无现成知识库，希望开箱即用	集成度高，管理便捷	数据需上传至平台
第三方知识库	已有自建知识库（如KooSearch、RAGFlow）	数据自主可控，无需迁移	需自行维护第三方服务

相关操作

- **创建平台知识库**。
- **接入第三方知识库**。
- 如何配置智能体记忆，请参考**为应用添加记忆**。

版权所有 © 华为云计算技术有限公司

6.4.2 知识库类型

费用说明

- 平台知识库：不同套餐包的**知识库存储空间容量**不同，具体可参考**计费说明**。
- 第三方知识库：AgentArts侧不计费，但第三方系统的部署与运维费用需用户自行承担，具体计费以第三方为准。

知识库类型

AgentArts支持以下两种类型知识库的管理：

- **平台知识库**：在AgentArts内直接创建并管理的知识库。支持上传文本文档、FAQ文档等文件，设置文档标签，并对其进行向量化存储、知识检索。适用于从零开始构建知识库，或数据量在平台限制范围内的场景。
- **第三方知识库**：将第三方系统中的知识库接入到AgentArts中。当前支持对接第三方知识库General、KooSearch和RAGFlow。适用于企业已有现成的知识库系统（如RAGFlow），希望直接复用其数据而无需重复上传的场景。接入后，Agent在检索时会直接调用第三方知识库的接口返回结果，数据不存储在AgentArts中。

选型建议

对比维度	平台知识库	第三方知识库
数据存储	存储在AgentArts平台。	存储在第三方系统。
使用场景	无现成知识库，需上传文档管理。	已有企业知识库，希望直接对接。
配置难度	低（上传即用）。	中（需配置连接参数）。
计费模式	占用套餐包存储额度。	AgentArts侧免费，第三方侧费用自理。
数据安全	数据存储在AgentArts平台。	数据保留在第三方系统，不传入平台。

相关文档

- **创建平台知识库**
- **接入第三方知识库**
- **创建第三方知识库时为什么查不到知识库列表?**
- **第三方知识库连接中的已启用/已停用状态如何更改?**

6.4.3 知识库使用限制

为保障平台资源合理分配和系统稳定性，知识库功能设有以下使用限制。

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

版权所有 © 华为云计算技术有限公司

表 6-30 知识库使用限制

资源	描述	体验版	团队标准版	团队高级版	企业专业版	企业旗舰版
知识库数量	单租户最多可创建平台知识库数量。	**5个知识库**	**10个知识库**	**20个知识库**	**50个知识库**	**100个知识库**
	单租户最多可创建第三方知识库连接数量（例如，General、RAGFlow）。	**5个第三方知识库连接**	**5个第三方知识库连接**	**5个第三方知识库连接**	**5个第三方知识库连接**	**5个第三方知识库连接**
	单租户最多可接入第三方知识库数量（例如，RAGFlow平台下的知识库1、知识库2）。	**50个第三方知识库**	**50个第三方知识库**	**50个第三方知识库**	**50个第三方知识库**	**50个第三方知识库**
	单个智能体最多可添加知识库数量。	**1个知识库**	**3个知识库**	**3个知识库**	**3个知识库**	**3个知识库**
	单个工作流最多可添加知识库数量。	**1个知识库**	**3个知识库**	**3个知识库**	**3个知识库**	**3个知识库**
知识库文件	单个知识库上传的知识文档或FAQ文档总数量。	**500个文档**	**500个文档**	**500个文档**	**500个文档**	**500个文档**
	上传知识文档、FAQ文档的单个文件大小限制。	**20MB**	**60MB**	**60MB**	**60MB**	**60MB**
	OBS接入文件的单个文件大小限制。	**20MB**	**128MB**	**128MB**	**128MB**	**128MB**
知识库容量	单租户知识库存储空间容量。	**1GB**	**5GB**	**50GB**	**200GB**	**1200GB**
召回数量	支持检索召回的最大文本切片数量。 • 召回指从知识库中检索并返回相关内容片段的过程。 • 文本切片指系统将文档自动切分成的片段，检索时按片返回。	**50个文本切片**	**50个文本切片**	**50个文本切片**	**50个文本切片**	**50个文本切片**

⚠ 注意

达到上述限制时，相关操作将失败并提示限制信息。如需提升配额，请购买团队标准版、团队高级版、企业专业版或企业旗舰版。

版权所有 © 华为云计算技术有限公司

选型建议

如果您有以下需求，建议选择**团队标准版、团队高级版、企业专业版或企业旗舰版**：

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

- 单个上传文件超过 20MB（**团队标准版、团队高级版、企业专业版或企业旗舰版**支持最大60MB）。
- 需要在单个智能体/工作流中关联多个知识库（体验版仅支持 1 个）。
- 需要更大的存储空间。

6.4.4 创建平台知识库

6.4.4.1 创建知识库

AgentArts提供的知识库功能对文本文档、FAQ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。

本文将详细介绍如何创建知识库，包含模型配置、解析配置和拆分配置等。

图 6-78 创建平台知识库流程

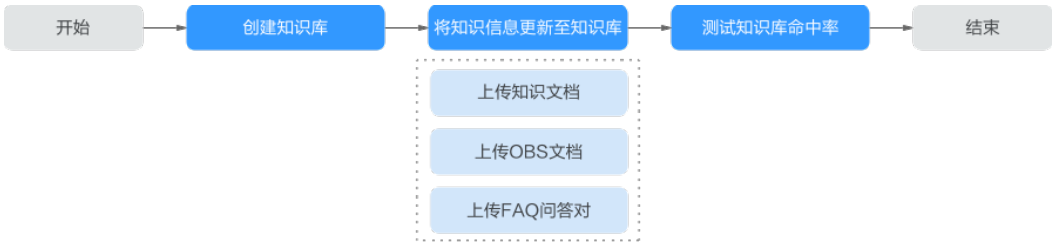


表 6-31 创建平台知识库流程

序号	流程环节		说明
1	创建知识库		AgentArts提供的知识库功能对文本文档、FAQ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。
2	将知识信息更新至知识库。 **说明** AgentArts支持通过多种方式将知识信息更新至知识库，用户可根据需求选择合适的方式。	上传知识文档	在创建知识库时，您可以选择上传本地的知识文档到知识库。
		上传OBS文档	在创建知识库时，支持通过对象存储服务（Object Storage Service, OBS）接入知识文档。 OBS是华为云提供的对象存储服务，具备数据存储能力。AgentArts知识库能够与OBS对接，将OBS中存储的文档数据上传至知识库。
		上传FAQ问答对	知识库支持通过上传FAQ文档来批量导入FAQ问答对，也支持在线创建。

版权所有 © 华为云计算技术有限公司

序号	流程环节	说明
3	测试知识库命中率	AgentArts通过对创建的知识库进行命中率测试，可以评估知识库的效果和准确性。

费用说明

创建知识库及上传文档会占用知识库存储容量，知识库存储容量具体可参考**计费说明**。

文档导入过程中的向量化处理可能会产生模型调用费用，请关注配额及计费详情。

AgentArts体验版不计费。

AgentArts团队标准版/团队高级版/企业专业版/企业旗舰版：AgentArts侧不计费，接入MaaS模型即服务的模型，具体计费请参考**MaaS模型即服务价格**。

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

前提条件

已**开通AgentArts服务**。

约束与限制

不同套餐包可以**创建平台知识库数量**不同，详细信息请参考**知识库使用限制**。

创建知识库

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击“创建知识库 > 创建平台知识库”。

- 创建平台知识库：适用于数据在AgentArts平台直接管理的场景，支持上传文档和FAQ。
- 接入第三方知识库：适用于企业已有现成知识库（如RAGFlow），希望直接对接复用的场景。具体操作请参考**接入第三方知识库**。

步骤4 在“创建知识库”的“基本信息”页面中，参见**表6-32**配置知识库信息。

版权所有 © 华为云计算技术有限公司

图 6-79 创建知识库（以**体验版**为例）



表 6-32 参数说明

参数		说明	示例
基本信息	名称	用于标识知识库。 由1~50个字符组成，包含中文、字母、数字、中划线、下划线，且必须以中文、字母、数字开头。	AgentArts平台知识库
	描述（可选）	用于对知识库内容和用途的简要说明。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 由0~100个字符组成。	AgentArts平台知识库
	知识库图标	系统默认知识库图标，可单击  自动生成图标，也可以自定义图标。 1. 鼠标移动至图标上，单击鼠标左键。 2. 上传已准备好的知识库图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	默认

版权所有 © 华为云计算技术有限公司

参数			说明	示例
解析配置	向量模型		创建后，不支持修改。 向量模型是一种将文本、图像等非结构化数据转换为数值向量的模型。例如，在文本处理阶段，用于对文本文档切片后的片段进行向量化表示；在知识检索阶段，根据用户输入的信息对切片进行召回。 向量模型用于在海量的知识库中，快速识别和用户输入信息语义相近的词或句子，进行信息的初步筛选，解决“大海捞针”的效率问题。 **体验版**：系统预置向量模型pangu_embedding。 **团队标准版**、**团队高级版**、**企业专业版**或**企业旗舰版**：可以自定义使用MaaS模型即服务的向量模型，开通MaaS模型即服务的向量模型请参考**接入预置模型**。	pangu_embedding
	精排模型		精排模型是一种用于对检索结果进行精细排序的模型。针对用户输入的信息，对向量模型召回的切片进行从高到低的相关度排序，把相关度最高的前几个信息（例如Top 10）呈现给用户。 精排模型用于进一步提升系统搜索的相关性准确度。 **体验版**：系统预置精排模型pangu_rerank。 **团队标准版**、**团队高级版**、**企业专业版**或**企业旗舰版**：可以自定义使用MaaS模型即服务的文本排序，开通MaaS模型即服务的文本排序请参考**接入预置模型**。 **说明** **体验版**用户在创建知识库后，无法修改精排模型的配置。 **团队标准版**、**团队高级版**、**企业专业版**或**企业旗舰版**用户在创建知识库后，可以通过“高级设置”选项来修改精排模型。具体操作步骤，请参考**管理知识库**。	pangu_rerank
	文档解析（可选）	OCR增强	• 未开启，不可调用OCR服务进行智能文档识别。 • 开启后，即可调用OCR服务进行智能文档识别，如表格解析或扫描文件等。	未开启
		页眉页脚解析	• 未开启，解析结果中不包含页眉页脚。 • 开启后，解析结果中包含页眉页脚。	开启

版权所有 © 华为云计算技术有限公司

参数			说明	示例
		目录页解析	• 未开启，解析结果中不包含目录页。 • 开启后，解析结果中包含目录页。	未开启
		图片解析	• 未开启，则在文档中遇到图片默认跳过，不处理图片。 • 开启后，根据需要选择“提取图片文本”或者“仅保留原图”。 – 提取图片文本：识别图片内文字。 – 仅保留原图：仅提取图片保存，不会识别图片内容，便于问答图文展示。	未开启
	分段策略（可选）		系统默认自动分段。支持如下分段策略，分段介绍及分段策略配置请参考**知识库分段**。 • 自动分段：按照系统默认预设的规则和分隔符切分。适用于格式规范的通用文档。 • 长度分段：基于内容的长度来决定如何进行分段。适用于对分段长度有严格要求的场景。 • 层级分段：根据内容的结构层次来进行分段。适用于带有明确标题层级的结构化文档（如说明书、书籍）。	自动分段

步骤5 配置完成后，单击“创建并上传”或“立即创建”。

- 创建并上传：进入“创建知识库”的“上传文档”页面，为知识库添加文档信息，具体操作请参考**上传文档**。
- 立即创建：页面返回知识库列表。

创建完成后，可以在“知识库”页签中查看创建完成的知识库，“类型”为“平台知识库”，“状态”为“已启用”。单击知识库名称进入知识库详情页进行后续配置。

----结束

相关文档

创建平台知识库应用实践，请参考**创建语文知识库**。

常见问题

- **自定义的向量模型一直正常使用，突然在列表中无法被选中或消失了。**

  登录**云搜索服务控制台**，在“模型管理”页面，检查该模型的“连通性”是否正常。只有状态显示为“正常”的模型，才能在AgentArts中被正常选择和使用。

- **单击“创建模型”时，页面跳转失败怎么办?**

  当前登录的账号没有**云搜索服务 CSS**的权限，请联系主账号管理员添加对应的权限。需要添加的权限项为：css:cluster:list，具体操作请参考**给IAM用户授权**。

版权所有 © 华为云计算技术有限公司

- **为什么无法创建知识库?**

  套餐包知识库数量有限制，超过限制后无法再创建。用户可以删除闲置知识库后再创建。

6.4.4.2 上传文档

在创建知识库时，您可以选择本地上传知识文档，或通过对象存储服务（Object Storage Service, OBS）接入知识文档。本文将详细说明如何选择本地上传知识文档。

上传文档适用于文件数量少、单次上传的场景。

在创建知识库后，需上传文档至知识库，平台支持知识文档、FAQ问答对、FAQ文档三种形态的文档。

- **上传知识文档**：知识文档上传并解析完成后，用户可根据需要新增/编辑切片，以提升知识库的检索效率。如果用户对知识库内文档的解析切片效果不满意，可修改知识库模型配置、解析配置和拆分配置信息，并对知识库内文档重新解析拆分。
- **上传FAQ问答对**：FAQ（Frequently Asked Questions，常见问题解答）问答对是指常见问题及其对应的答案，用于快速解决用户可能遇到的问题。

  知识库支持通过上传FAQ文档来批量导入FAQ问答对。FAQ文档上传并解析完成后，用户可根据需要新增/编辑切片，以提升知识库的检索效率。

费用说明

文档导入过程中的向量化处理可能会产生模型调用费用，请关注配额及计费详情，具体可参考**计费说明**。

AgentArts**体验版**不计费。

AgentArts**团队标准版/团队高级版/企业专业版/企业旗舰版**：AgentArts侧不计费，接入MaaS模型即服务的模型，具体计费请参考**MaaS模型即服务价格**。

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

选型建议

- 知识文档：适用于非结构化的长文档（如产品手册、操作指南、技术白皮书），系统会自动进行切片和向量化处理。
- FAQ问答对/文档：适用于结构化的问答数据（如Excel表格中的Q&A），能够精准匹配用户提问，适合客服场景。

前提条件

- 已**开通AgentArts服务**。
- 已完成**创建知识库**。

上传知识文档

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

版权所有 © 华为云计算技术有限公司

步骤3 在“知识库”页面，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 单击“添加知识”，“上传方式”选择“知识文档上传”。

- 知识文档上传：支持本地批量导入知识文档。
- OBS上传：可通过OBS路径批量导入文档。
- FAQ问答对：可通过导入文档上传，或在线创建问答对。

步骤5 在“上传文档”区域，单击“点此上传”，在弹出的对话框中，选择要上传的文档。

当上传的文档由“上传中”变为“上传成功”，表示文档已成功上传。

说明

- 支持上传常见的文档格式：
  - 文档/文本类：doc、docx、pdf、ofd、txt、md、wps。
  - 表格/数据类：xls、xlsx、csv、et。
  - 演示/幻灯片类：ppt、pptx、dps。
  - 图片/设计类：jpg、jpeg、png、gif、bmp、webp、tiff、tif、psd、ico、pcx。

    AgentArts支持用户根据需求，将图片类的文件直接存储，或是提取图片中的文字信息进行存储。
- 不同套餐包对文件大小限制不同，具体请参考**知识库使用限制**。

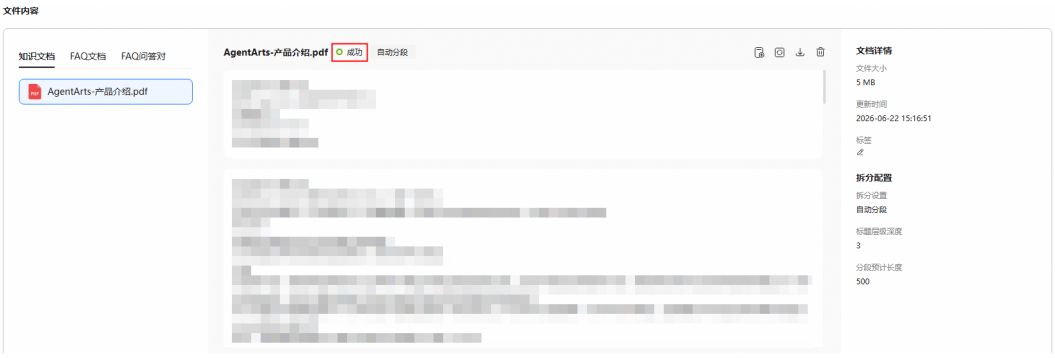
步骤6 单击“完成”。

返回知识库详情页面，文件上传后将自动进行解析，解析时间视文件大小和数量而定。如果状态长时间显示为“解析中”或“失败”，请检查文件格式是否损坏或联系技术支持工程师。

待文件状态为“成功”，即完成文件解析。

上传的知识文档，显示在知识库详情页面的“文件内容 > 知识文档”中。

图 6-80 文件解析成功



“文件内容”区域说明：

- 左侧：知识库文档的类别及上传的文档列表。
- 中间：知识文档的切片信息。
- 右侧：文档详情。

步骤7 （可选）在“知识库详情”页面右上方，单击“ ··· > 标签”，可添加文档标签（如科普、教育、科技等）用于对文档进行分类。

版权所有 © 华为云计算技术有限公司

步骤8 （可选）在“文档详情”区域，单击，为文档设置标签，可在工作流“知识检索”节点中配置按照标签检索文档。

----结束

上传 OBS 文档

通过OBS接入文档适用于文件数量较多、文件较大（超过本地上传限制），或数据已存储在OBS桶中的场景。

注意

需拥有已实名认证的华为账号，且该账号具有OBS桶的访问权限。

步骤1 登录AgentArts智能体平台。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 单击“添加知识”，“上传方式”选择“OBS上传”。

- 知识文档上传：支持本地批量导入知识文档。
- OBS上传：可通过OBS路径批量导入文档。
- FAQ问答对：可通过导入文档上传，或在线创建问答对。

步骤5 配置OBS信息，如图6-81所示，参数配置请参考表6-33。

警告

OBS配置信息保存之后将无法修改。如果配置错误，需删除当前知识库并重新创建。

一个知识库仅支持接入一次OBS文件，接入后不可编辑。

图 6-81 配置 OBS 信息



表 6-33 OBS 配置参数说明

参数	说明	示例
OBS桶名	选择存储知识文档的OBS桶。接入后不支持修改。	test01

版权所有 © 华为云计算技术有限公司

参数	说明	示例
OBS路径	知识文档在OBS中的具体路径。接入后不支持修改。 • 接入的OBS目录下子目录层级不超过三层，如接入目录为path1，则path1最大目录层级为path1/path2/path3/path4/xxx.txt。 • 支持上传常见的文档格式： – 文档/文本类：doc、docx、pdf、ofd、txt、md、wps。 – 表格/数据类：xls、xlsx、csv、et。 – 演示/幻灯片类：ppt、pptx、dps。 – 图片/设计类：jpg、jpeg、png、gif、bmp、webp、tiff、tif、psd、ico、pcx。AgentArts支持用户根据需求，将图片类的文件直接存储，或是提取图片中的文字信息进行存储。 • 不同套餐包对文件大小限制不同，具体请参考**知识库使用限制**。	RMSLogs/

步骤6 配置完成后，单击“立即执行”。

执行时间取决于文件数量和大小，请耐心等待。如果状态显示“执行失败”，请检查OBS桶权限是否正确配置或查看执行日志。

当执行状态显示为“执行成功”，即完成OBS接入知识文档。

步骤7 单击“完成”。

返回知识库详情页面。

----结束

上传 FAQ 问答对

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 单击“添加知识”，“上传方式”选择“FAQ问答对”。

- 知识文档上传：支持本地批量导入知识文档。
- OBS上传：可通过OBS路径批量导入文档。
- FAQ问答对：可通过导入文档上传，或在线创建问答对。

步骤5 选择创建方式。

- “选择创建方式”选择“文档导入”。

  单击“点此上传”，在弹出的对话框中选择符合“Excel模板”或“Word模板”要求的FAQ文档上传。

  当上传的文档由“上传中”变为“上传成功”，表示文档已成功上传。

版权所有 © 华为云计算技术有限公司

说明

- 支持按照模板上传文件，文件格式仅支持xlsx、xls、docx、doc。
- 不同套餐包对文件大小限制不同，具体请参考**知识库使用限制**。
- Excel单个文件最大支持100,000条数据。
- FAQ文件中不允许空行，空行后的数据将被忽略。

- “选择创建方式”选择“在线创建”。

单击“添加问答对”，填写FAQ问答对信息，如**图6-82**所示，参数配置请参考**表6-34**。

图 6-82 FAQ 问答对



表 6-34 参数说明

参数	说明	示例
标准问题	必选参数。 用户可能提出的问题的最直接、最清晰的形式。 由1~1000个字符组成。	小儿肥胖该怎样医治
相似问题	可选参数。 与标准问题意思相近或相关的一系列问题。 由1~1000个字符组成。	小儿肥胖的治疗需要综合考虑饮食、运动和行为改变等多方面因素。建议咨询专业的儿科医生或营养师，根据孩子的具体情况制定个性化的治疗计划。

版权所有 © 华为云计算技术有限公司

参数	说明	示例
答案	必选参数。 针对标准问题提供的详细解答。 由1~10000个字符组成。	孩子一旦患上肥胖症家长要先通过运动和饮食来改变孩子的情况，要让孩子做一些他这个年龄段能做的运动，如游泳，慢跑等，要给孩子多吃一些像苹果，猕猴桃，胡萝卜等食物，禁止孩子吃高热量，高脂肪的食物，像蛋糕，干果，曲奇饼干等，严格地控制孩子的饮食，不要让他暴饮暴食，多运动对改变孩子肥胖都是有好处的，在治疗小儿肥胖期间如果情况严重，建议家长先带孩子去医院检查孩子肥胖症的原因再针对性的治疗。

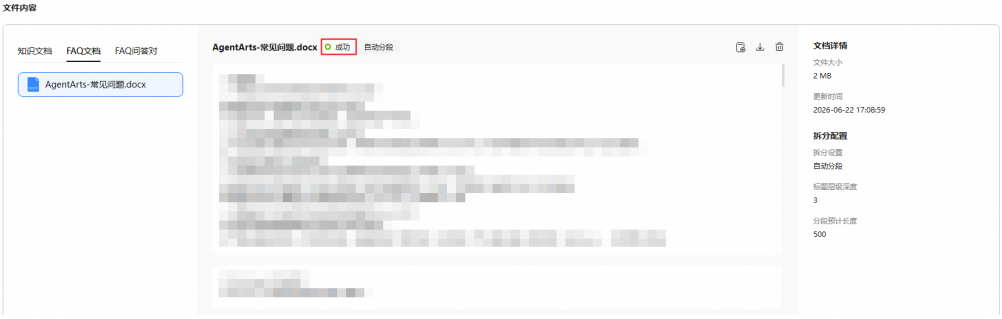
步骤6 单击“完成”。

返回知识库详情页面，文件上传后将自动进行解析，解析时间视文件大小和数量而定。如果状态长时间显示为“解析中”或“失败”，请检查文件格式是否损坏或联系技术支持工程师。

待文件状态为“成功”，即完成文件解析。

- 上传的FAQ文档，显示在知识库详情页面的“文件内容 > FAQ文档”中。

图 6-83 文件解析成功



“文件内容”区域说明：

  - 左侧：知识库文档的类别及上传的文档列表。
  - 中间：知识文档的切片信息。
  - 右侧：文档详情。

- 添加的FAQ问答对，显示在知识库详情页面的“文件内容 > FAQ问答对”中，左侧显示问题，右侧显示问答对。

  单击内容，可以修改。

版权所有 © 华为云计算技术有限公司

图 6-84 FAQ 问答对



----结束

常见问题

为什么文档上传失败?

- 可能原因：
  - 文档格式不符合格式要求。
  - 文档大小超过限制。
- 处理措施：
  - 支持上传常见的文档格式：
    - 文档/文本类：doc、docx、pdf、ofd、txt、md、wps。
    - 表格/数据类：xls、xlsx、csv、et。
    - 演示/幻灯片类：ppt、pptx、dps。
    - 图片/设计类：jpg、jpeg、png、gif、bmp、webp、tiff、tif、psd、ico、pcx。

      AgentArts支持用户根据需求，将图片类的文件直接存储，或是提取图片中的文字信息进行存储。
  - 不同套餐包对文件大小限制不同，具体请参考**知识库使用限制**。

为什么文档解析失败?

- 可能原因：文档格式损坏或OCR模型不可用。
- 处理措施：请确认文件格式正常，如使用了基于KooSearch的模型，请在KooSearch服务侧检查OCR模型状态并修复。

相关操作

在知识库详情页面，在“文件内容 > 知识文档”、“文件内容 > FAQ文档”、“文件内容 > FAQ问答对”页签，单击指定文件或问题，可以查看文档或问题详情。在文档或问题详情页面，支持的其他操作请参考**表6-35**。

版权所有 © 华为云计算技术有限公司

表 6-35 相关操作

操作	说明
新增切片	单击，新增切片。 **说明** 仅知识文档、FAQ文档支持新增切片。 • **切片标题**：用于快速了解每个切片的内容，便于在大量切片中进行查找和管理。由1~1000个字符组成。 示例： 1 什么是AgentArts • **切片内容**：通过切片内容，用户可以详细阅读和理解每个知识点或信息。由1~6000个字符组成。 示例： 包含了“1 什么是AgentArts”及其子章节“AgentArts的使用限制”的内容。
重新对文档解析拆分	单击，在弹出的对话框中，单击“重试”即进行文档重新解析切片。重试任务在知识库详情页面的“ ··· > 日志”中查看。 解析时间视文件大小和数量而定。如果状态长时间显示为“解析中”或“失败”，请检查文件格式是否损坏或联系技术支持工程师。 **说明** 重新解析和向量化处理，该过程会消耗计算资源，请谨慎操作。 仅知识文档支持重新解析拆分。
下载知识文档	单击，可以下载知识文档、FAQ文档。 **说明** 仅知识文档、FAQ文档支持下载文档。
删除知识文档	**警告** 删除后数据将无法恢复，请谨慎操作。删除知识文档、FAQ文档、FAQ问答对将影响依赖该知识库的智能体回答效果。 单击，可以删除知识文档、FAQ文档、FAQ问答对。
编辑文档切片	• 鼠标移动至待编辑的切片信息区域，单击，可以修改知识文档、FAQ文档切片。 • 单击，可以编辑FAQ问答对。
删除文档切片	鼠标移动至待删除的切片信息区域，单击，可以删除知识文档、FAQ文档切片。
修改标签	在“文档详情”区域，单击，可以修改标签。 **说明** 仅知识文档支持标签功能。

6.4.4.3 知识库命中测试

AgentArts支持对创建的知识库进行命中率测试，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

建议在知识库创建完成后、正式用于智能体调用前，先进行命中测试验证效果。

本文将详细介绍知识库创建完成后如何执行命中测试操作。

典型场景

在智能客服场景中，通过命中测试验证知识库能否准确召回用户常见问题的相关文档，确保正式上线后检索效果符合预期。

费用说明

知识库检索召回依赖于创建知识库时设置的向量模型和精排模型。

AgentArts**体验版**不计费。

AgentArts**团队标准版/团队高级版/企业专业版/企业旗舰版**：AgentArts侧不计费，接入MaaS模型即服务的模型，具体计费请参考**MaaS模型即服务价格**。

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

前提条件

已完成**上传文档**。

命中测试

系统将输入的问题转化为向量，并在知识库中查找与之语义最相近的文档切片，根据相似度计算匹配分值。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”界面，单击需要测试的知识库操作列的“命中测试”；或当知识库以卡片形式展示时，将鼠标悬浮在需要命中测试的知识库卡片上，单击“命中测试”。

步骤4 在命中测试页面左侧文本框中输入问题，单击“命中测试”。

在页面右侧将根据不同的检索方式（语义检索、关键词检索、混合检索、FAQ检索），展示多条匹配的内容，并按照匹配分值降序排列。

- 语义检索：理解语义，召回相关内容。适用于自然语言提问。
- 关键词检索：精确匹配，结果可控。适用于精确匹配特定名词。
- 混合检索：结合语义检索和关键词检索的优势。适用于需要兼顾语义理解和精确匹配的场景。
- FAQ检索：直接返回预设答案。适用于标准问答对场景。

步骤5 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。

匹配分值范围为0~1，分值越接近1表示相关性越高。

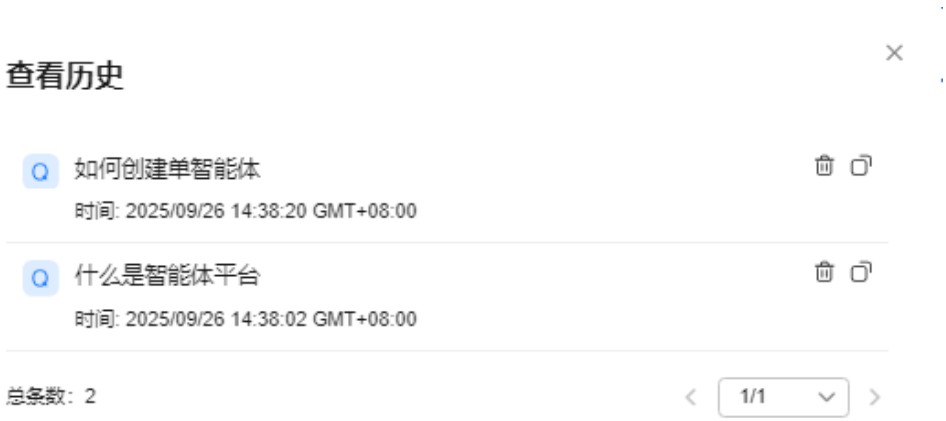
如果测试结果为空或分值过低，建议检查文档切片设置是否合理，或尝试更换提问关键词。

系统将根据知识库配置自动选择检索方式，或您可以切换检索方式进行对比测试。

版权所有 © 华为云计算技术有限公司

步骤6 单击右上角的“查看历史”，可以查看用户输入的历史问题。在历史记录右侧，单击可删除或复制该记录。

图 6-85 查看历史



----结束

常见问题

如何优化知识库检索效果?

1. 优化文档结构：无结构的文档，转化为有层次结构的文档。
2. AgentArts**团队标准版、团队高级版、企业专业版或企业旗舰版**支持更换接入的其他模型。
3. 修改知识库的分段设置。

为什么检索不到已上传的文档?

- 可能原因：
  - 阈值设置太高。
  - 检索词语、模式不合适。
  - 精排模型不可用。
- 解决措施：
  - **调整阈值大小**。
  - 修改检索词语、模式。
  - 在**KooSearch服务**中确认精排模型正常。

6.4.4.4 知识库分段

知识库分段介绍

对知识库进行分段设置主要是为了提高信息的组织性、可读性和检索效率。具体来说，分段设置的优势包括：

- 提高可读性：将长篇内容分成多个段落，每个段落集中讨论一个主题或一个方面，可以使读者更容易理解和消化信息。长篇大论往往会让读者感到疲惫，而分段则可以提供自然的停顿点，帮助读者更好地吸收信息。

 版权所有 © 华为云计算技术有限公司

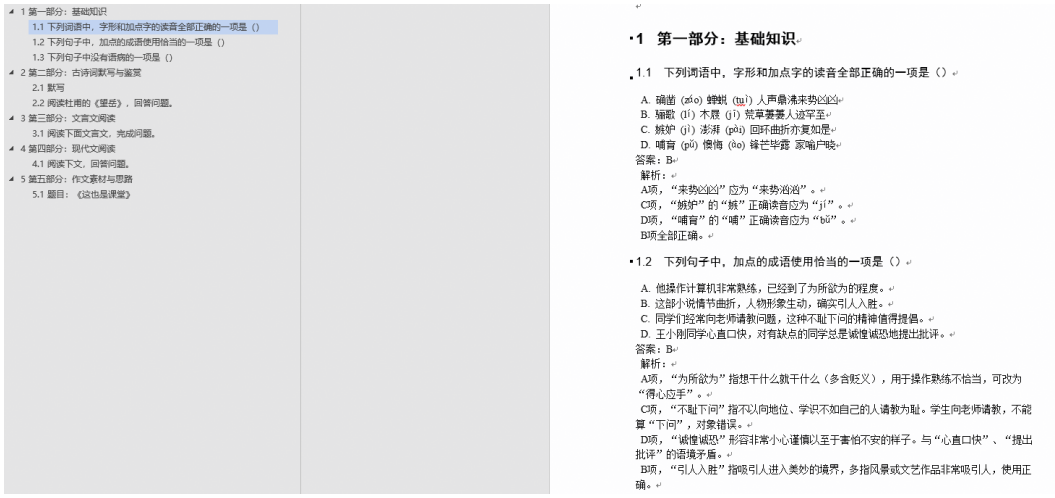
- 增强组织性：通过分段，可以将相关的信息归类在一起，使知识库的内容结构更加清晰。这种结构化的组织方式不仅有助于作者在编写时保持逻辑清晰，也方便读者快速找到他们感兴趣的内容。
- 提高检索效率：对于大型知识库而言，分段设置可以与索引、标签等机制结合使用，帮助用户更快地定位到特定的信息。例如，搜索引擎可以更容易地识别出每个段落的关键内容，从而提高搜索结果的相关性和准确性。
- 便于维护和更新：分段的知识库内容更容易进行维护和更新。当需要修改或添加信息时，可以针对特定的段落进行操作，而不会影响到其他部分的内容，这有助于保持知识库的准确性和时效性。
- 适应不同的阅读习惯：不同的人有不同的阅读习惯和偏好。分段设置可以满足不同用户的需求，比如一些用户可能喜欢快速浏览标题和段落首句来获取概览，而另一些用户则可能更倾向于深入阅读每个段落的详细内容。

合理的分段能平衡上下文完整性与检索精准度。切片过大可能引入噪声，切片过小可能丢失上下文关联。

分段策略

如果您不确定如何选择，推荐优先使用自动分段；如果文档具有明确的书籍目录结构，推荐使用层级分段。

图 6-86 分段原始文档部分截图



 版权所有 © 华为云计算技术有限公司

表 6-36 分段策略

分段类型	自动分段	长度分段	层级分段
分段原则	AgentArts可对上传的内容进行自动分段，支持复杂布局的文件处理，例如： ● 可识别段落。 ● 可识别页眉/页脚/脚注等非重点内容。 ● 支持跨页的段落合并。 ● 支持解析表格中的图片信息。 ● 支持解析文档中的表格内容（目前，仅支持解析带线框的表格内容）。	根据用户需求，灵活配置分段标识符、分段最大长度、分段重叠度等参数，同时还能通过设置文本预处理规则，在分段前对文本进行特定处理。	根据文档的目录结构、章节划分等层级信息，将内容划分为不同层级的文本单元。
分段优点	系统预设，无需额外配置，提高使用效率。	严格控制分段长度，节约大模型会话时的Tokens。	结构清晰，便于用户理解，提高检索效率。
分段缺点	自动分段的效果依赖文档质量。	配置复杂，不适用于所有文档，需要用户对文档内容和参数规则有较深的理解。	需要文档具有明确的层级结构，对于结构不规范的文档难以适用。
分段效果	**图 6-87** 效果良好 	**图 6-88** 效果差 	**图 6-89** 效果良好 
适用场景	适用于大部分分段场景，通常文档结构较为规范，是使用情况最多的分段类型。	适用于对分段长度有严格要求的场景。	适用于结构层次分明的知识体系，例如技术手册、法律条文、标准规范等。这些文档通常具有明确的结构层次，需要按照章节、小节等进行组织和检索。

分段策略配置

修改分段策略后，系统将对知识库内文档重新进行解析和向量化处理，该过程会消耗计算资源，请谨慎操作。

版权所有 © 华为云计算技术有限公司

⚠ 警告

修改分段策略将导致知识库内现有文档切片失效，系统将自动重新切分。如有重要的人工编辑切片，请提前备份或确认。

- 自动分段：系统默认自动分段。选择该策略时会自动利用句号、分号、问号、感叹号等标点符号作为分段依据，将文档分割成一个个独立的句子或段落，无其他配置项。
- 长度分段：选择该分段策略，完成以下配置。
  - 分段标识符：分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。

    支持中文句号、英文句号、中文感叹号、英文感叹号、中文问号、英文问号、空格、中文逗号、英文逗号。

    ⚠ 注意

    长度分段配置中，如果未命中分段标识符，分段将会失败。
  - 分段预计长度：分段的最大长度。文档的正文如果大于设定的最大长度，则截取最大长度的片段为新文档，随后回溯分段重叠字符，继续向后检查，直到文档结束。取值范围为1~6000，默认值为500。

    例如，设置长度500，当段落长800时，将切分为两段，第一段为0~497/502，第二段为498/503~800。
- 层级分段：选择该分段策略，完成以下配置。
  - 层级解析模型：
    - 自动解析：自动识别和解析具有层级结构的数据或信息。示例请参考**自动解析示例**。
    - 规则解析：支持添加自定义层级规则。示例请参考**规则解析示例**。
  - 标题层级深度：指设置的切分标题级别，例如，文本包含最多5级标题，选择的标题层级深度为3，则会分别将所有3级标题下的内容合并成文本块，文本块作为一个整体执行后续切分操作。输入值必须在1到10之间。
  - 标题保存方式：指标题信息在切片中的保存形式，影响检索结果的展示逻辑和索引构建方式。
    - 保存多标题组合：多级标题用空格组合。

      ```
      1级标题 2级标题 3级标题 … 最后一级标题
      文本
      ```
    - 保存最后一级标题：仅最后一级标题。

      ```
      最后一级标题
      文本
      ```
  - 跨标题合并：根据需求开启或者关闭。
    - 开启“跨标题合并”功能：当不同标题下的段落文字较少时，平台会自动将其合并到指定的分段长度，有助于生成更加全面的内容。
    - 关闭“跨标题合并”开关：不会自动合并不同标题下的内容。
  - 分段标识符：分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。自定义分段中如果未命中分段标识符，分段将

版权所有 © 华为云计算技术有限公司

会失败。支持中文句号、英文句号、中文感叹号、英文感叹号、中文问号、英文问号、空格、中文逗号、英文逗号。

- 分段预计长度：分段的最大长度。文档的正文如果大于设定的最大长度，则截取最大长度的片段为新文档，随后回溯分段重叠字符，继续向后检查，直到文档结束。

  取值范围为1~6000，默认值为500。

配置完成后，您可以在知识库详情页的“切片管理”中查看分段效果，或通过“检索测试”验证检索精准度。

自动解析示例

图 6-90 配置



图 6-91 自动解析效果



版权所有 © 华为云计算技术有限公司

规则解析示例

图 6-92 配置



图 6-93 规则解析效果



相关操作

知识库分段用于**创建知识库**，为知识库设置“拆分配置”。

6.4.5 接入第三方知识库

6.4.5.1 第三方知识库分类

当您的企业数据已存储在现有的知识库系统中时，无需将数据迁移至AgentArts，通过接入第三方知识库可直接调用现有数据，实现知识复用。接入完成后，AgentArts智能体在运行时可通过标准接口调用第三方知识库中的数据，无需数据迁移。调用时，AgentArts将检索请求发送至第三方知识库，获取检索结果后返回给智能体使用。

支持将第三方系统中的知识库接入到AgentArts中。目前支持的知识库类型请参考**表6-37**。

版权所有 © 华为云计算技术有限公司

表 6-37 第三方知识库分类说明

知识库类型	说明
General	通用外部知识库。用户需要先将知识库按**第三方通用知识库接入规范**进行适配，再接入。接入操作请参考**连接General知识库**。 适用于自研知识库或非标准接口的系统，需要开发者自行开发适配接口。
KooSearch	可以直接接入KooSearch知识库。接入操作请参考**连接KooSearch知识库**。 适用于已部署了该开源项目的用户，可实现零代码快速对接。
RAGFlow	可以直接接入RAGFlow知识库。接入操作请参考**连接RAGFlow知识库**。 适用于已购买或部署了对应产品的用户，可实现零代码快速对接。

选型建议

分类	适用场景	接入方式	复杂度
General	如果您使用的是自研或其他第三方知识库，请选择General类型并参考接入规范进行适配开发。	先按**第三方通用知识库接入规范**进行适配，再接入。	中
KooSearch/RAGFlow	如果您已拥有KooSearch或RAGFlow知识库，推荐直接选择对应类型以实现快速对接。	直接接入。	低

6.4.5.2 连接 General 知识库

适用于企业已有现成知识库系统（如自研知识库、RAGFlow等），希望直接对接复用，避免数据迁移的场景。

目前业界有诸多知识库种类，但并没有一个统一的规范。对于用户自开发的知识库平台，AgentArts提供了**第三方通用知识库接入规范**。知识库的开发者参考此规范进行适配后，即能够作为第三方知识库接入AgentArts平台。

本文将介绍如何在AgentArts对接第三方General知识库。

图 6-94 连接 General 知识库流程



版权所有 © 华为云计算技术有限公司

费用说明

调用第三方知识库接口可能产生第三方平台的相关费用，请确保第三方服务账户状态正常。

约束与限制

⚠ 警告

删除外部知识库连接前，请确认无智能体正在使用该连接下的知识库，否则将导致智能体问答失败。

当资源数量达到上限时，请删除不再使用的资源以释放配额：

- **第三方知识库连接**：进入 “第三方知识库连接” 页面进行删除。具体操作请参考**管理知识库**中的“删除外部知识库连接信息”。删除后，基于该连接接入的知识库将不可用。
- **第三方知识库**：进入 “知识库” 页面进行删除。具体操作请参考**管理知识库**中的“取消接入”。

关于知识库的更多约束与限制，请参考**知识库使用限制**。

步骤一：准备工作

- 用户拥有一个自开发的知识库系统（下文统称第三方知识库），并且已经按**第三方通用知识库接入规范**完成适配。知识库的服务接口发布到公网IP，并且可访问。
- 获取已搭建的第三方知识库连接信息：
  a. 获取服务器接口地址，即第三方知识库接口的发布地址IP，例如，http://10.0.0.0:8080。
  b. 获取鉴权信息密钥，该密钥是第三方知识库平台用户的认证信息，由第三方知识库提供。例如，123456789。

  说明

  用户填入时不需要在密钥开始处拼接Bearer，AgentArts平台会负责拼接操作。

  c. 获取知识库详情页面链接，登录第三方知识库平台后，进入知识库详情页面，复制地址栏的URL，将其中的知识库ID替换为{{id}}。以下图RAGFlow为例，其知识库详情页面链接为：http://10.0.0.0/dataset?id={{id}} 。

  该链接信息用于接入第三方知识库后，AgentArts用户可根据此链接跳转第三方知识库平台页面。如果第三方知识库只有后端服务器，没有web或前端页面，可填写http://xxx.com作为占位。

图 6-95 RAGFlow 知识库详情页



版权所有 © 华为云计算技术有限公司

步骤二：连接 General 第三方知识库

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

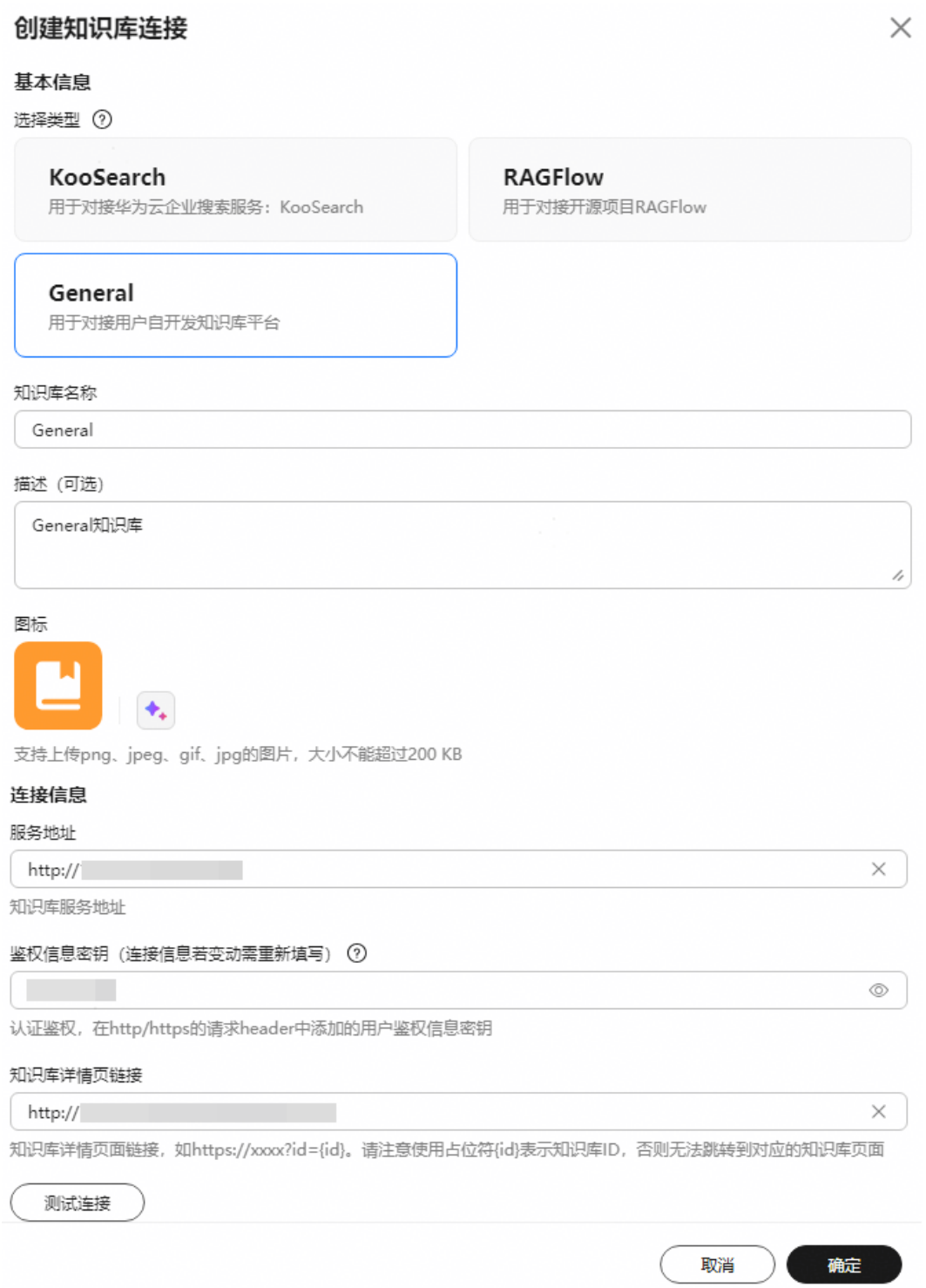
步骤3 在“知识库”页面，单击“第三方知识库连接”。

步骤4 在“第三方知识库连接”界面，单击“新建知识库连接”。

步骤5 在“创建知识库连接”页面，设置连接General知识库基本信息，参数配置请参考**表6-38**。

版权所有 © 华为云计算技术有限公司

图 6-96 连接外部知识库



版权所有 © 华为云计算技术有限公司

表 6-38 参数说明

参数		说明	示例
基本信息	选择类型	• General：用于对接用户自研知识库平台，需自行实现检索接口。 • KooSearch：用于对接华为云企业搜索服务，开箱即用。 • RAGFlow：用于对接开源项目RAGFlow，快速对接复用。	General
	知识库名称	用于标识知识库。 由1~50个字符组成，包含中文、字母、数字、中划线、下划线，且必须以中文、字母、数字开头。	General
	描述（可选）	用于简要说明知识库内容和用途。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 由0~255个字符组成。	General知识库
	图标	系统默认知识库图标，可单击 自动生成图标，也可以自定义图标。 1. 鼠标移动至图标上，单击鼠标左键。 2. 上传已准备好的知识库图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	默认
连接信息	服务地址	能够访问检索接口及查询列表接口的地址，以https://或http://开头。 填入按**获取服务器接口地址，即第三方知识库接口的发布地址...**获取的信息。	http://10.0.0.0:8080

版权所有 © 华为云计算技术有限公司

参数		说明	示例
	鉴权信息密钥	在http/https的请求header中添加的用户鉴权信息密钥。 填入按**获取鉴权信息密钥，该密钥是第三方知识库平台用户的...**获取的信息。 **注意** 请确保鉴权密钥来源安全，不要在非安全环境下泄露。 **关于密钥加密：**鼠标移动至“鉴权信息密钥”右侧的 上，可查看密钥的加密方式。平台默认使用**KMS加密 + 默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 • 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 • 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	123456789
	知识库详情页面链接	第三方General知识库详情页面的链接，可通过该页面直接访问General知识库的详情页面。 填入按**获取知识库详情页面链接，登录第三方知识库平台后，...**获取的信息。 **注意** 需要使用占位符{{id}}表示知识库ID，否则无法跳转到对应的知识库页面。	http://10.0.0.0/dataset?id={{id}}

步骤6 单击“测试连接”，弹出“测试成功”提示。

如果显示“第三方知识库连接失败，请检查连接地址和认证信息”，请参考**连接外部知识库失败如何排查？**。

步骤7 测试成功后，单击“确定”。

页面返回第三方知识库连接列表，连接创建成功后“状态”为“已停用”，完成**步骤三：创建General第三方知识库**后“状态”变为“已启用”。

----结束

步骤三：创建 General 第三方知识库

在完成General知识库连接后，您可以将General中的知识库接入到AgentArts中进行使用。

已完成**步骤二：连接General第三方知识库**。

步骤1 登录**AgentArts智能体平台**。

 版权所有 © 华为云计算技术有限公司

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击“创建知识库 > 接入第三方知识库”。

- 创建平台知识库：适用于数据在AgentArts平台直接管理的场景，支持上传文档和FAQ。具体操作请参考**创建平台知识库**。
- 接入第三方知识库：适用于企业已有现成知识库（如RAGFlow），希望直接对接复用的场景。

步骤4 在“接入第三方知识库”界面中，选择已接入的知识库，在右侧勾选知识库。

说明

如果列表中没有需要的第三方知识库，则可以单击“新建”，新建知识库连接，具体配置请参考**表6-38**。

版权所有 © 华为云计算技术有限公司

图 6-97 接入第三方知识库



步骤5 单击“确定”，第三方知识库创建完成。

创建完成的知识库，显示在知识库列表，“类型”为“General”，“状态”为“已启用”。

----结束

相关操作

在“第三方知识库连接”页面，支持的其他操作请参考**表6-39**。

版权所有 © 华为云计算技术有限公司

表 6-39 相关操作

操作	说明
编辑第三方知识库连接信息	在对应的“操作”列下，单击“编辑”，可以编辑外部知识库连接信息。 如遇连接失败，请参考**各个章节**的常见问题。
删除第三方知识库连接信息	**注意** 需要先取消接入后，才可以删除。 在对应的“操作”列下，单击“删除”，可以删除外部知识库连接信息。

常见问题

操作过程中如遇到以下问题，可参考对应解决方案：

- **创建第三方知识库时为什么查不到知识库列表?**
- **外部知识库连接中的已启用/已停用状态如何更改?**
- **连接外部知识库失败如何排查?**
  - 接口返回401/403：检查鉴权信息密钥（API Key）是否正确。
  - 连接超时：检查URL是否正确。
  - 请检查General服务是否支持在公网使用API访问。

相关文档

- 连接外部知识库后，请参考**知识库命中测试**校验知识库是否可用。
- 连接外部知识库后，可以对外部知识库进行管理，具体请参考**管理知识库**。

6.4.5.3 连接 KooSearch 知识库

为什么要连接外部知识库?

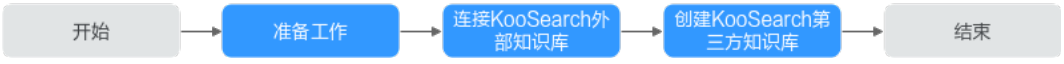
当企业已在KooSearch等平台构建了成熟的知识库，通过外部知识库连接可直接复用现有数据，无需重复上传，同时可利用KooSearch的联网增强能力获取公域知识。

AgentArts支持连接外部知识库平台，以便用户访问和利用外部的知识资源。

KooSearch可以帮助企业搭建企业级RAG，提升用户获取企业私有知识的效率，同时提供联网增强功能，提供互联网公域知识问答能力。主要用于智能客服应用、数字人、数字员工、AI助手、AI搜索等知识问答场景。

本文将介绍如何在AgentArts对接第三方KooSearch知识库。

图 6-98 连接 KooSearch 知识库流程



版权所有 © 华为云计算技术有限公司

费用说明

通过AgentArts调用KooSearch知识库会消耗KooSearch的API调用次数及流量，请关注KooSearch服务的资源使用情况。

约束与限制

警告

删除外部知识库连接前，请确认无智能体正在使用该连接下的知识库，否则将导致智能体问答失败。

当资源数量达到上限时，请删除不再使用的资源以释放配额：

- **第三方知识库连接**：进入“第三方知识库连接”页面进行删除。具体操作请参考**管理知识库**中的“删除外部知识库连接信息”。删除后，基于该连接接入的知识库将不可用。
- **第三方知识库**：进入“知识库”页面进行删除。具体操作请参考**管理知识库**中的“取消接入”。

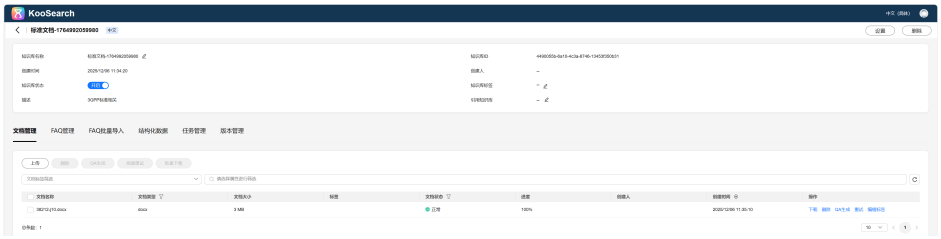
关于知识库的更多约束与限制，请参考**知识库使用限制**。

步骤一：准备工作

获取第三方KooSearch知识库的连接信息：

1. 确保需要连接的第三方KooSearch知识库平台已经开通了华为云KooSearch服务，并购买了相关资费套餐。第三方KooSearch的拥有者需要发布**搜索知识库**和**获取知识库列表**的两个API接口供后续使用者在AgentArts平台调用。发布API接口请参考**通过API使用KooSearch实现搜索问答**。
2. 在KooSearch中创建一个知识库，并上传相关文档。

图 6-99 KooSearch 中的知识库



3. 获取KooSearch服务地址：可以参考**KooSearch接口获取**获取APIG公网地址，也可以参考以下步骤：
   a. 登录KooSearch绑定的APIG服务，在“apig接口 > 实例管理”中单击实例名称。

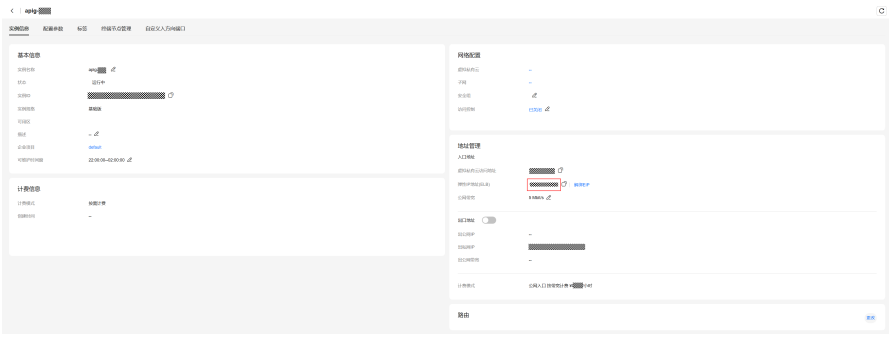
图 6-100 实例名称



版权所有 © 华为云计算技术有限公司

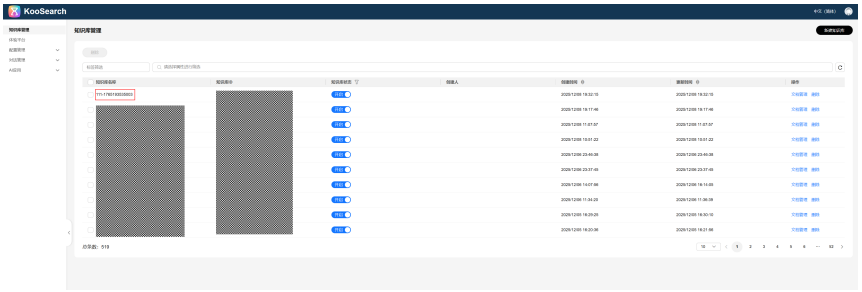
b. 在“实例信息”页签中找到并复制“弹性IP地址”，该弹性IP地址就是KooSearch的服务地址。

图 6-101 弹性 IP 地址



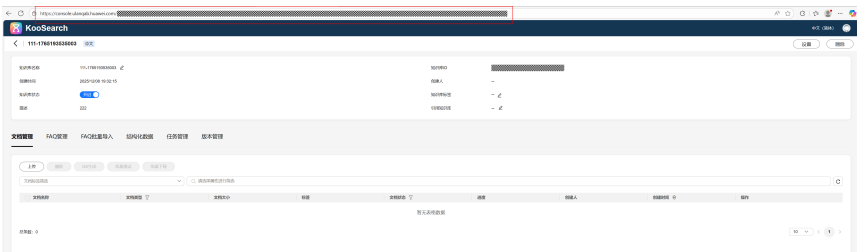
4. 获取项目ID请参考**获取KooSearch服务的项目ID**。
5. 获取应用ID请参考**获取KooSearch服务的应用ID**。
6. 获取APIG凭证请参考**获取KooSearch服务的APIG凭证**的APP认证部分。
7. 获取KooSearch知识库详情页面链接。
   a. 登录KooSearch主页，单击需要接入的知识库，进入KooSearch知识库详情页面。

图 6-102 KooSearch 知识库列表



b. 复制浏览器地址栏中的链接，该链接就是KooSearch知识库详情页面的链接。

图 6-103 知识库详情页面链接



步骤二：连接 KooSearch 第三方知识库

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

 版权所有 © 华为云计算技术有限公司

步骤3 在“知识库”页面，单击“第三方知识库连接”。

步骤4 在“第三方知识库连接”界面，单击“新建知识库连接”。

步骤5 在“创建知识库连接”页面，设置连接KooSearch知识库基本信息，参数配置请参考**表6-40**。

 版权所有 © 华为云计算技术有限公司

图 6-104 连接外部知识库



版权所有 © 华为云计算技术有限公司

表 6-40 参数说明

参数		说明	示例
基本信息	选择类型	• General：用于对接用户自研知识库平台，需自行实现检索接口。 • KooSearch：用于对接华为云企业搜索服务，开箱即用。 • RAGFlow：用于对接开源项目RAGFlow，快速对接复用。	KooSearch
	知识库名称	用于标识知识库。 由1~50个字符组成，包含中文、字母、数字、中划线、下划线，且必须以中文、字母、数字开头。	KooSearch
	描述（可选）	用于简要说明知识库内容和用途。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 由0~255个字符组成。	KooSearch知识库
	图标	系统默认知识库图标，可单击 自动生成图标，也可以自定义图标。 1. 鼠标移动至图标上，单击鼠标左键。 2. 上传已准备好的知识库图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	默认
连接信息	服务地址	能够访问检索接口及查询列表接口的地址，以https://或http://开头。	https://xxx.com
	项目id	登录KooSearch知识库平台账号的项目ID。	0503dda8970xxxxxx
	应用id	KooSearch知识库平台的应用ID。	9ae90c5e53xxxxxx

版权所有 © 华为云计算技术有限公司

参数		说明	示例
	APIG凭证	KooSearch平台接口绑定的APIG凭证的AppCode，用于认证鉴权。 **注意** 请妥善保管APIG凭证，避免泄露。建议定期更换凭证以确保安全。 **关于密钥加密：**鼠标移动至“APIG凭证”右侧的 上，可查看密钥的加密方式。平台默认使用**KMS加密 + 默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 • 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 • 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	b099xxxxxxxx
	知识库详情页面链接	第三方KooSearch知识库详情页面的链接，可通过该页面直接访问KooSearch知识库的详情页面。 注意需要使用占位符{{id}}表示知识库ID，否则无法跳转到对应的知识库页面。	https://xxxxx.com/xxxx?id={{id}}

步骤6 单击“测试连接”，弹出“测试成功”提示。

如果显示“第三方知识库连接失败，请检查连接地址和认证信息”，请检查KooSearch服务是否支持在公网使用API访问。

步骤7 测试成功后，单击“确定”。

页面返回第三方知识库连接列表，连接创建成功后“状态”为“已停用”，完成**步骤三：创建KooSearch第三方知识库**后“状态”变为“已启用”。

----结束

步骤三：创建 KooSearch 第三方知识库

在完成KooSearch知识库连接后，您可以将KooSearch中的知识库接入到AgentArts中进行使用。

已完成**步骤二：连接KooSearch第三方知识库**。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击“创建知识库 > 接入第三方知识库”。

- 创建平台知识库：适用于数据在AgentArts平台直接管理的场景，支持上传文档和FAQ。具体操作请参考**创建平台知识库**。

 版权所有 © 华为云计算技术有限公司

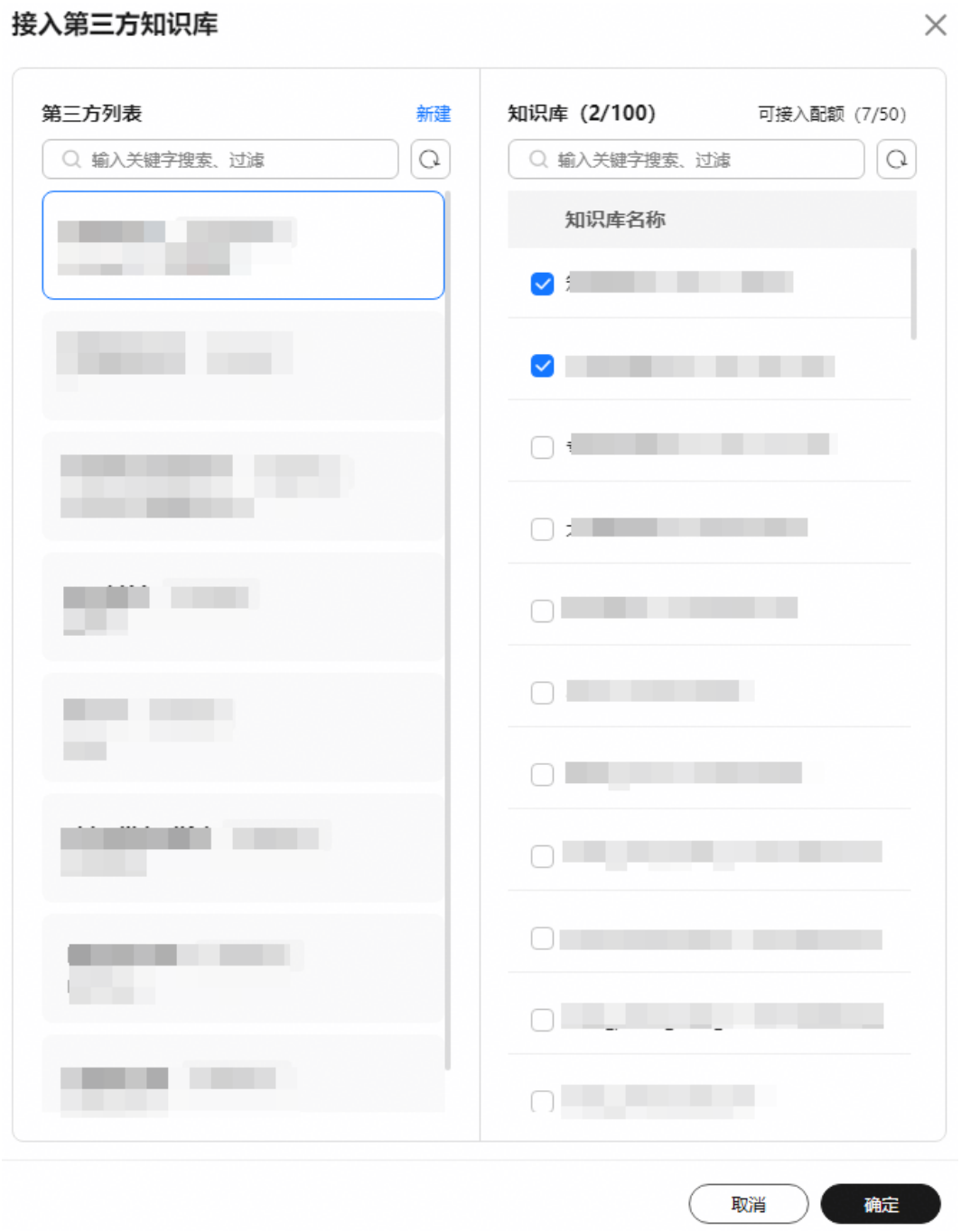
- 接入第三方知识库：适用于企业已有现成知识库（如RAGFlow），希望直接对接复用的场景。

步骤4 在“接入第三方知识库”界面中，选择已接入的知识库，在右侧勾选知识库。

说明

如果列表中没有需要的第三方知识库，则可以单击“新建”，新建知识库连接，具体配置请参考表6-38。

图 6-105 接入第三方知识库



步骤5 单击“确定”，第三方知识库创建完成。

版权所有 © 华为云计算技术有限公司

创建完成的知识库，显示在知识库列表，“类型”为“KooSearch”，“状态”为“已启用”。

----结束

相关操作

在“第三方知识库连接”页面，支持的其他操作请参考**表6-41**。

表 6-41 相关操作

操作	说明
编辑第三方知识库连接信息	在对应的“操作”列下，单击“编辑”，可以编辑外部知识库连接信息。 如遇连接失败，请参考**各个章节**的常见问题。
删除第三方知识库连接信息	**注意** 需要先取消接入后，才可以删除。 在对应的“操作”列下，单击“删除”，可以删除外部知识库连接信息。

常见问题

- **创建第三方知识库时为什么查不到知识库列表?**
- **外部知识库连接中的已启用/已停用状态如何更改?**
- **连接KooSearch失败如何排查?**
  - 接口返回401/403：检查APIG凭证是否正确。
  - 连接超时：检查URL是否正确。

相关文档

- 连接外部知识库后，请参考**知识库命中测试**校验知识库是否可用。
- 连接外部知识库后，可以对外部知识库进行管理，具体请参考**管理知识库**。

6.4.5.4 连接 RAGFlow 知识库

AgentArts支持连接外部知识库，通过连接外部知识库，可以显著地扩展内部知识库的知识范围，引入更多领域和更广泛的信息资源，从而提高知识库的全面性和深度。

RAGFlow适用于已自建开源RAGFlow知识库系统的场景。

本文将介绍如何在AgentArts中对接第三方RAGFlow知识库。

图 6-106 连接 RAGFlow 知识库流程



工作原理：连接成功后，AgentArts智能体在检索知识时，会将检索请求转发至RAGFlow服务，RAGFlow返回检索结果后由智能体进行后续处理。

版权所有 © 华为云计算技术有限公司

典型场景：某企业已使用RAGFlow构建了包含产品手册、技术文档的知识库，现需在AgentArts中开发智能客服智能体，可直接连接现有RAGFlow知识库，无需重复上传文档。

费用说明

调用外部RAGFlow知识库接口会产生RAGFlow服务侧的计算资源消耗及网络流量费用，请关注第三方服务状态。

约束与限制

警告

删除外部知识库连接前，请确认无智能体正在使用该连接下的知识库，否则将导致智能体问答失败。

当资源数量达到上限时，请删除不再使用的资源以释放配额：

- **第三方知识库连接**：进入 “第三方知识库连接” 页面进行删除。具体操作请参考**管理知识库**中的“删除外部知识库连接信息”。删除后，基于该连接接入的知识库将不可用。
- **第三方知识库**：进入 “知识库” 页面进行删除。具体操作请参考**管理知识库**中的“取消接入”。

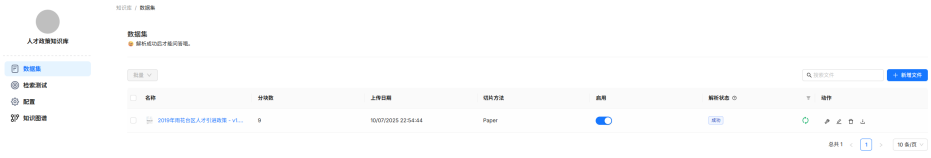
关于知识库的更多约束与限制，请参考**知识库使用限制**。

步骤一：准备工作

获取第三方RAGFlow知识库的连接信息：

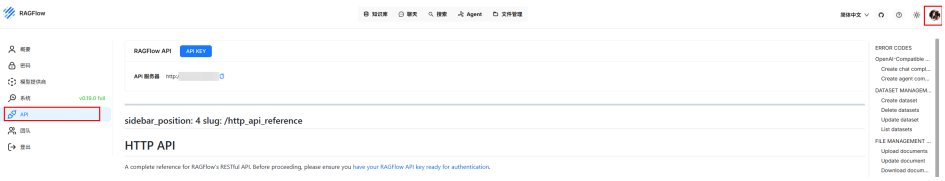
1. 在RAGFlow里创建一个知识库，并上传相关文档。

   图 6-107 RAGFlow 中的知识库

2. 在RAGFlow“个人中心 > API”中获取RAGFlow的连接信息。

   图 6-108 RAGFlow 连接信息

3. 在知识库列表中选中第三方知识库，单击 复制知识库地址，在新浏览器窗口的地址栏中粘贴该链接并访问，即可直接跳转到第三方知识库的详情页。

版权所有 © 华为云计算技术有限公司

图 6-109 复制知识库地址

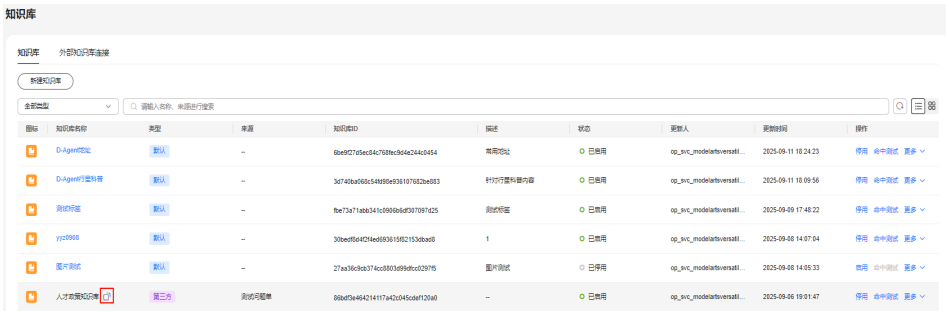


图 6-110 RAGFlow 知识库详情页



步骤二：连接 RAGFlow 第三方知识库

步骤1 登录**AgentArts智能体平台**。

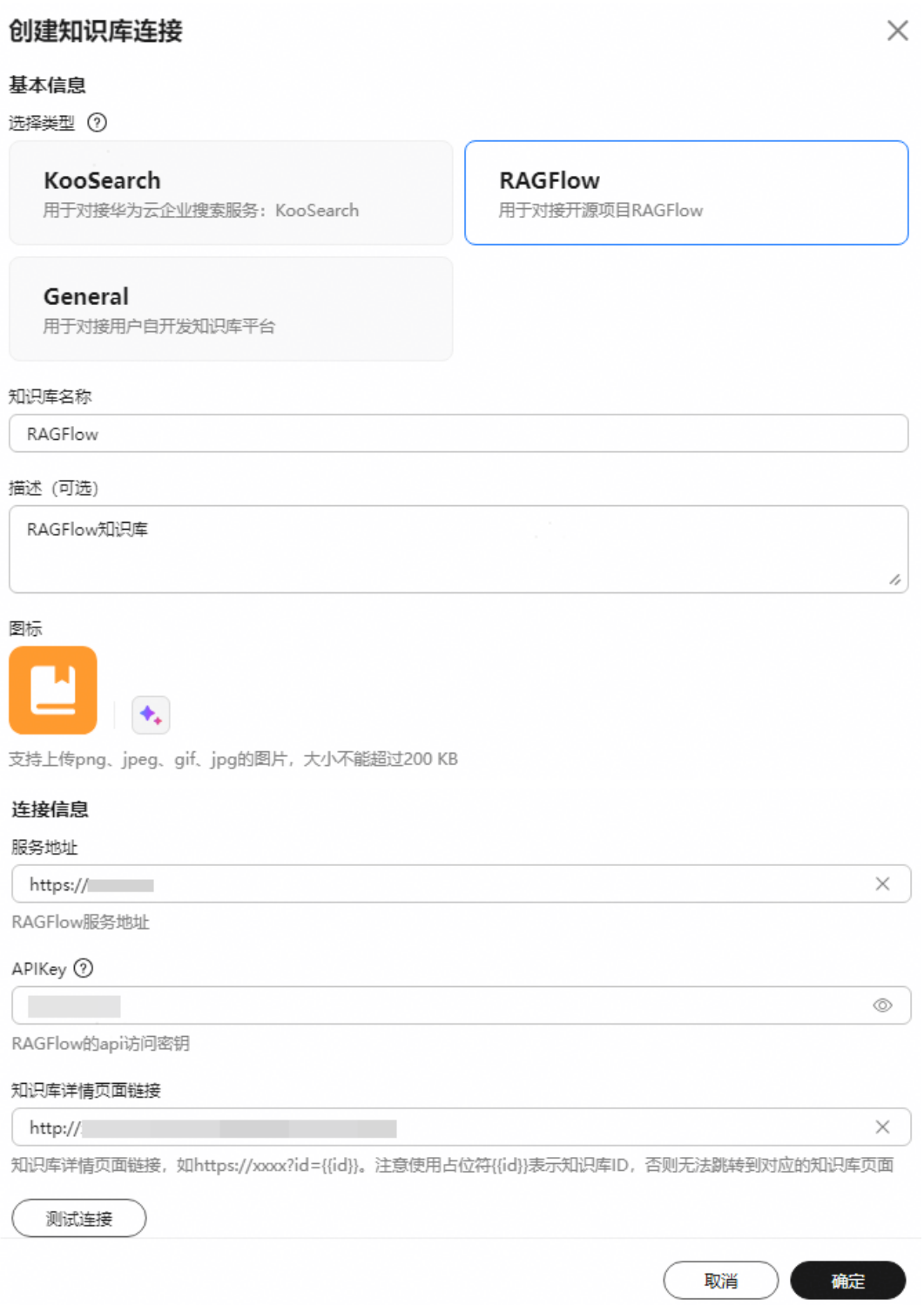
步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击“第三方知识库连接”。

步骤4 在“创建知识库连接”页面，设置连接RAGFlow知识库基本信息，参数配置请参考**表6-42**。

版权所有 © 华为云计算技术有限公司

图 6-111 连接外部知识库



版权所有 © 华为云计算技术有限公司

表 6-42 参数说明

参数		说明	示例
基本信息	选择类型	• General：用于对接用户自研知识库平台，需自行实现检索接口。 • KooSearch：用于对接华为云企业搜索服务，开箱即用。 • RAGFlow：用于对接开源项目RAGFlow，快速对接复用。	RAGFlow
	知识库名称	用于标识知识库。 由1~50个字符组成，包含中文、字母、数字、中划线、下划线，且必须以中文、字母、数字开头。	RAGFlow
	描述（可选）	用于简要说明知识库内容和用途。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 由0~255个字符组成。	RAGFlow知识库
	图标	系统默认知识库图标，可单击自动生成图标，也可以自定义图标。 1. 鼠标移动至图标上，单击鼠标左键。 2. 上传已准备好的知识库图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	默认
连接信息	服务地址	能够访问检索接口及查询列表接口的地址。 需填写RAGFlow服务的API基础地址，如https://ragflow.example.com。	https://xxx.com
	APIKey	用于访问第三方RAGFlow知识库的鉴权密钥。 **注意** 请妥善保管鉴权密钥，避免泄露。建议定期更换密钥以确保安全。 **关于密钥加密：**鼠标移动至“APIKey”右侧的上，可查看密钥的加密方式。平台默认使用**KMS加密+默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 • 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 • 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	sk-xxxxxxxx

版权所有 © 华为云计算技术有限公司

参数		说明	示例
	知识库详情页面链接	第三方RAGFlow知识库详情页面的链接，可通过该页面直接访问RAGFlow知识库的详情页面。注意需要使用占位符{{id}}表示知识库ID，否则无法跳转到对应的知识库页面。	http://xxxxx.com/knowledge/dataset?id={{id}}

步骤5 单击“测试连接”，弹出“测试成功”提示。

如果显示“第三方知识库连接失败，请检查连接地址和认证信息”，请检查RAGFlow服务是否支持在公网使用API访问。

步骤6 测试成功后，单击“确定”。

页面返回第三方知识库连接列表，连接创建成功后“状态”为“已停用”，完成**步骤三：创建RAGFlow第三方知识库**后“状态”变为“已启用”。

----结束

步骤三：创建 RAGFlow 第三方知识库

在完成RAGFlow知识库连接后，您可以将RAGFlow中的知识库接入到AgentArts中进行使用。

已完成**步骤二：连接RAGFlow第三方知识库**。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”页面，单击“创建知识库 > 接入第三方知识库”。

- 创建平台知识库：适用于数据在AgentArts平台直接管理的场景，支持上传文档和FAQ。具体操作请参考**创建平台知识库**。
- 接入第三方知识库：适用于企业已有现成知识库（如RAGFlow），希望直接对接复用的场景。

步骤4 在“接入第三方知识库”界面中，选择已接入的知识库，在右侧勾选知识库。

说明

如果列表中没有需要的第三方知识库，则可以单击“新建”，新建知识库连接，具体配置请参考**表6-38**。

版权所有 © 华为云计算技术有限公司

图 6-112 接入第三方知识库



步骤5 单击“确定”，第三方知识库创建完成。

创建完成的知识库，显示在知识库列表，“类型”为“RAGFlow”，“状态”为“已启用”。

----结束

相关操作

在“第三方知识库连接”页面，支持的其他操作请参考**表6-43**。

版权所有 © 华为云计算技术有限公司

表 6-43 相关操作

操作	说明
编辑第三方知识库连接信息	在对应的“操作”列下，单击“编辑”，可以编辑外部知识库连接信息。 如遇连接失败，请参考**各个章节**的常见问题。
删除第三方知识库连接信息	**注意** 需要先取消接入后，才可以删除。 在对应的“操作”列下，单击“删除”，可以删除外部知识库连接信息。

常见问题

- **创建第三方知识库时为什么查不到知识库列表?**
- **外部知识库连接中的已启用/已停用状态如何更改?**
- **连接RAGFlow失败如何排查?**
  - 接口返回401/403：检查APIKey是否正确。
  - 连接超时：检查URL是否正确。

相关文档

- 连接外部知识库后，请参考**知识库命中测试**校验知识库是否可用。
- 连接外部知识库后，可以对外部知识库进行管理，具体请参考**管理知识库**。

6.4.5.5 知识库命中测试

AgentArts通过对创建的知识库进行命中率测试，以评估知识库的效果和准确性。

本文将详细介绍第三方知识库创建完成后如何执行命中测试操作。

工作原理

命中测试通过语义检索技术，将用户输入的问题与知识库中的文档片段进行语义相似度计算，返回相似度最高的多条结果。

典型场景

- 验证知识库是否正确导入：输入已知存在的文档内容关键词，检查是否能召回。
- 评估检索质量：输入用户实际提问，观察召回结果的相关性。

费用说明

第三方知识库检索召回，会产生资源消耗，请关注第三方资源使用情况。

前提条件

已完成**连接General知识库**、**连接KooSearch知识库**、**连接RAGFlow知识库**。

版权所有 © 华为云计算技术有限公司

命中测试

步骤1 登录AgentArts智能体平台。

步骤2 在左侧导航栏中选择“开发中心 > 组件库”，选择“知识库”页签。

步骤3 在“知识库”界面，单击需要测试的知识库操作列的“命中测试”；或当知识库以卡片形式展示时，将鼠标悬浮在需要命中测试的知识库卡片上，单击“命中测试”。

步骤4 在命中测试页面左侧文本框中输入问题，单击“命中测试”。

在命中测试页面右侧进行语义检索，展示多条匹配的内容，并按照匹配分值降序排列。

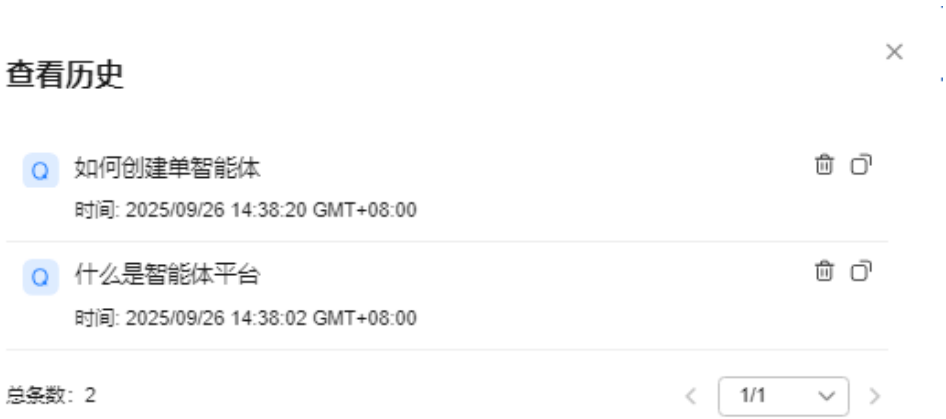
步骤5 用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。

匹配分值范围为0~1，分值越接近1表示相关性越高。

如果测试结果为空或分值过低，建议检查第三方知识库服务是否正常运行，或尝试更换提问关键词。

步骤6 单击右上角的“查看历史”，可以查看用户输入的历史问题。在历史记录右侧，单击可删除或复制该记录。

图 6-113 查看历史



----结束

常见问题

- **如何优化知识库检索效果?**
  优化文档结构：将无结构的文档转化为有层次结构的文档。
- **为什么检索不到已上传的文档?**
  - 可能原因：阈值设置太高；检索词语不合适。
  - 解决措施：调整阈值大小；更换提问关键词。

版权所有 © 华为云计算技术有限公司

6.4.5.6 第三方通用知识库接入规范

前言

当您拥有企业内部已有的知识库系统（如内部Wiki、私有文档库），且不希望将数据迁移至AgentArts平台时，可遵循本规范开发适配接口。

典型场景

企业智能体需要检索内部保密文档，通过实现本规范定义的API接口，AgentArts可直接调用您的知识库服务进行问答增强。

工作原理

您需开发并部署两个标准HTTP接口，AgentArts智能体在对话过程中，当需要检索外部知识时，将依次调用您的接口：

1. 调用**获取知识库列表**接口，获取可用的知识库列表。
2. 根据用户问题，调用**检索知识库**接口，获取相关文档切片。
3. 将检索结果注入提示词，增强问答能力。

费用说明

通过AgentArts调用第三方知识库接口，可能会产生公网流量费用，具体取决于您的第三方服务部署环境。该调用过程将消耗智能体的运行时长资源。

API 接口规范

对于API接口适配，开发者需要新增两个接口，一个是**获取知识库列表**，一个是**检索知识库**。开发者按照**接口：获取知识库列表**、**接口：检索知识库**规范提供对外接口，并在接口内实现密钥鉴权、数据查询、知识检索等业务逻辑，最后按照规范返回对应数据。

接口：获取知识库列表

- URI
  ```
  GET /knowledge-bases
  ```
- 查询Query参数

  表 6-44 查询 Query 参数

参数	是否必选	参数类型	描述
name	否	String	**参数解释**： 知识库名称，用于获取知识库列表时筛选知识库名称。 **取值范围**： 长度不超过64个字符。 **默认取值**： 不涉及。

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
offset	否	Integer	**参数解释：** 当前请求的分页偏移量，表示从第几条开始获取数据。默认值为0，表示从第0条开始获取数据。 **取值范围：** 0～65535。 **默认取值：** 0。
limit	否	Integer	**参数解释：** 当前请求返回的分页最大数据条数，表示一次请求返回的数据条数。默认值为10，表示一页最多展示10条数据。 **取值范围：** 1~100。 **默认取值：** 10。

- 请求Header参数

  表 6-45 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释**： 认证鉴权使用的api-key密钥，拼接“Bearer ”后的字符串。 示例：Bearer d59******9C3，用于第三方知识库的鉴权。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

- 响应参数

 版权所有 © 华为云计算技术有限公司

表 6-46 响应参数 200

参数	参数类型	描述
total	Integer	**参数解释**: 知识库数据总数。 **取值范围:** 不涉及。
knowledge_base_list	Array of 表6-47	**参数解释:** 知识库数据信息列表。 **注意** 第三方知识库只返回可用的知识库，即在当前时刻可被外部引用并支持检索的知识库数据。 **取值范围:** 不涉及。

表 6-47 KnowledgeBaseInfo

参数	参数类型	描述
knowledge_base_id	String	**参数解释**: 知识库ID。 **取值范围:** 不涉及。
name	String	**参数解释:** 知识库名称。 **取值范围:** 不涉及。
description	String	**参数解释**: 知识库描述。 **取值范围:** 不涉及。

表 6-48 响应参数 400、500 等

参数	参数类型	描述
error_code	String	**参数解释**: 错误码。 **取值范围:** 不涉及。

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
error_msg	String	**参数解释：** 错误描述。 **取值范围：** 不涉及。

- 请求示例
```
GET http://10.0.0.0:8090/knowledge-bases?name=123&offset=0&limit=10
```
- 响应示例
```
{
  "total": 1,
  "knowledge_base_list": [
    {
      "knowledge_base_id": "001",
      "name": "一个通用第三方知识库的名称 / this is a knowledgeBase name",
      "description": "一个描述 / a description"
    }
  ]
}
```

接口：检索知识库

- URI
```
POST /knowledge-bases/retrieve
```
- 请求Header参数

表 6-49 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释**： 认证鉴权使用的api-key密钥，拼接“Bearer ”后的字符串。 示例：Bearer d59******9C3，用于第三方知识库的鉴权。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

- Body参数

 版权所有 © 华为云计算技术有限公司

表 6-50 Body 参数

参数	是否必选	参数类型	描述
knowledge_base_ids	是	Array of String	**参数解释**: 知识库ID的列表，用于指定搜索时需要检索的知识库。 **取值范围:** 列表长度大于等于1，即至少指定检索一个知识库。 **默认取值:** 不涉及。
query	是	String	**参数解释:** 用户的搜索内容，即需要在知识库中检索的问题或关键信息。 **取值范围:** 长度不超过4096个字符。 **默认取值:** 不涉及。
method	是	String	**参数解释:** 搜索方式。语义检索：“doc”。 **取值范围:** 不涉及。 **默认取值:** 不涉及。
offset	否	Integer	**参数解释:** 当前请求的分页偏移量，表示从第几条开始获取数据。默认值为0，表示从第0条开始获取数据。 **取值范围:** 0～65535。 **默认取值:** 0。

 版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
limit	否	Integer	**参数解释：** 当前请求返回的分页最大数据数量，表示一次请求返回的数据条数。默认值为10，表示一页最多展示10条数据。 **取值范围：** 1~100。 **默认取值：** 10。
top_k	否	Integer	**参数解释：** 检索返回结果的最大召回数量。 **取值范围：** 1-1024。 **默认取值：** 50。
search_threshold	否	Number	**参数解释：** 检索时的相关度得分阈值，低于该阈值的结果将被过滤。默认值0.0时表示不设置相关度得分阈值。 **取值范围：** 0.0 - 1.0。 **默认取值：** 0.0。
extra_params	否	Array of **表6-51**	**参数解释**： 额外的自定义参数，用于传递第三方知识库所需的其他业务参数。 示例： "extra_params": [{ "key": "tenant_id", "value": "enterprise-001" }, { "key": "data_region", "value": "cn-north" }] **取值范围：** 不涉及。 **默认取值：** 不涉及。

 版权所有 © 华为云计算技术有限公司

表 6-51 KnowledgeBaseExtraParam

参数	是否必选	参数类型	描述
key	是	String	**参数解释**: 额外参数主键名。 **取值范围:** 不涉及。
value	是	String	**参数解释:** 额外参数的值。 **取值范围:** 不涉及。

- 响应参数

表 6-52 响应参数 200

参数	参数类型	描述
total	Integer	**参数解释**: 检索结果总数。 **取值范围:** 不涉及。
search_result_list	Array of 表6-53	**参数解释:** 检索结果的信息列表。 **取值范围:** 不涉及。

表 6-53 SearchChunkInfo

参数	参数类型	描述
knowledge_base_id	String	**参数解释**: 该切片所属的知识库ID。 **取值范围:** 不涉及。
file_id	String	**参数解释:** 该切片所属的文件ID。 **取值范围:** 不涉及。

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
chunk_id	String	**参数解释**: 切片的唯一标识ID。 **取值范围:** 不涉及。
content	String	**参数解释**: 检索返回的文本内容，即实际注入提示词的知识片段。 **取值范围:** 不涉及。
score	Number	**参数解释**: 该切片与检索query的相关度得分，将根据此得分决定是否使用该切片。 **取值范围:** 0.0 - 1.0。
title	String	**参数解释**: 文本标题，推荐返回文件名称，用于在引用来源时展示给用户。 **取值范围:** 不涉及。

表 6-54 响应参数 400、500 等

参数	参数类型	描述
error_code	String	**参数解释**: 错误码。 **取值范围:** 不涉及。
error_msg	String	**参数解释:** 错误描述。 **取值范围:** 不涉及。

- 请求示例
  POST http://*10.0.0.0:8090*/knowledge-bases/retrieve
- 响应示例

```
{
  "knowledge_base_ids": [
    "001"
  ],
  "query": "检索",
  "method": "doc",
```

 版权所有 © 华为云计算技术有限公司

```
  "offset": 0,
  "limit": 10,
  "top_k": 50,
  "search_threshold": 0.01
}
{
  "total": 1,
  "search_result_list": [
    {
      "knowledge_base_id": "001",
      "file_id": "101",
      "chunk_id": "201",
      "content": "一个通用第三方知识库的检索切片内容 / this is a retrieve chunk content",
      "score": 0.8,
      "title": "检索切片的标题 / retrieve chunk title"
    }
  ]
}
```

API 接口实现案例（Java）

```
@Slf4j
@RestController
public class CommonThirdpartyKnowledgeBaseController {
    /**
     * list knowledgeBase function
     *
     * @param name
     * @param offset
     * @param limit
     * @param authorization
     * @return ListKnowledgeBaseResponse
     */
    @GetMapping("/knowledge-bases")
    public ListKnowledgeBaseResponse listKnowledgeBase(
            @RequestHeader("Authorization") String authorization,
            @RequestParam(required = false) String name,
            @RequestParam(required = false) Integer offset,
            @RequestParam(required = false) Integer limit) {
        /**
         * 对密钥认证鉴权
         * Verify and authenticate the key
         */
        if (!authorization.startsWith("Bearer ")) {
            log.error("invalid authorization");
            throw new RuntimeException("invalid authorization");
        }
        String apiKey = authorization.substring("Bearer ".length());
        /**
         * then do something for authorization
         *
         */

        /**
         * 根据业务实现代码逻辑，构造返回结果
         * Construct and return results based on your business implementation and code logic
         */
        ListKnowledgeBaseResponse response = new ListKnowledgeBaseResponse();
        response.total = 1;
        KnowledgeBaseInfo knowledgeBaseInfo = new KnowledgeBaseInfo();
        knowledgeBaseInfo.knowledge_base_id = "001";
        knowledgeBaseInfo.name = "一个通用第三方知识库的名称 / this is a knowledgeBase name";
        knowledgeBaseInfo.description = "一个描述 / a description";
        response.knowledge_base_list = List.of(knowledgeBaseInfo);
        return response;
    }


    /**
```

 版权所有 © 华为云计算技术有限公司

```
 * @param authorization
 * @param request
 * @return RetrieveKnowledgeBaseResponse
 */
@PostMapping("/knowledge-bases/retrieve")
public RetrieveKnowledgeBaseResponse retrieveKnowledgeBase(
        @RequestHeader("Authorization") String authorization,
        @RequestBody RetrieveKnowledgeBaseRequest request) {
    /**
     * 对密钥认证鉴权
     * Verify and authenticate the key
     */
    if (!authorization.startsWith("Bearer ")) {
        log.error("invalid authorization");
        throw new RuntimeException("invalid authorization");
    }
    String apiKey = authorization.substring("Bearer ".length());
    /**
     * then do something for authorization
     *
     */

    /**
     * 根据业务实现代码逻辑，构造返回结果
     * Construct and return results based on your business implementation and code logic
     */
    RetrieveKnowledgeBaseResponse response = new RetrieveKnowledgeBaseResponse();
    response.total = 1;
    SearchChunkInfo searchChunkInfo = new SearchChunkInfo();
    searchChunkInfo.knowledge_base_id = "001";
    searchChunkInfo.file_id = "101";
    searchChunkInfo.chunk_id = "201";
    searchChunkInfo.content = "一个通用第三方知识库的检索切片内容 / this is a retrieve chunk content";
    searchChunkInfo.title = "检索切片的标题 / retrieve chunk title";
    searchChunkInfo.score = 0.8f;
    response.search_result_list = List.of(searchChunkInfo);
    return response;
}


/*
 * the response structure of the listKnowledgeBase function
 */
@Data
public class ListKnowledgeBaseResponse {
    Integer total;
    List<KnowledgeBaseInfo> knowledge_base_list;
}

@Data
public class KnowledgeBaseInfo {
    String knowledge_base_id;
    String name;
    String description;
}

/*
 * the request structure of the retrieveKnowledgeBase function
 */
@Data
public static class RetrieveKnowledgeBaseRequest {
    List<String> knowledge_base_ids;
    String query;
    String method;
    Integer offset;
    Integer limit;
    Integer top_k;
    Float search_threshold;
    List<KnowledgeBaseExtraParam> extra_params;
```

 版权所有 © 华为云计算技术有限公司

```
    }

    @Data
    public static class KnowledgeBaseExtraParam {
        String key;
        String value;
    }

    /*
     * the response structure of the retrieveKnowledgeBase function
     */
    @Data
    public class RetrieveKnowledgeBaseResponse {
        Integer total;
        List<SearchChunkInfo> search_result_list;
    }

    @Data
    public class SearchChunkInfo {
        String knowledge_base_id;
        String file_id;
        String chunk_id;
        String content;
        Float score;
        String title;
    }
}
```

常见问题

如果接口调用失败，请检查鉴权头是否正确配置。常见报错排查请参见第三方接入故障排除。

6.4.6 使用知识库

6.4.6.1 在单智能体中使用知识库

支持在AgentArts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。

通过关联知识库，智能体可在回答用户问题前，先从私有知识库中检索相关信息，从而基于私有数据生成准确回答，有效减少模型幻觉。

典型使用场景：

- 企业内部知识问答：员工可通过智能体查询公司制度、流程规范等。
- 产品文档查询：基于产品手册回答用户技术问题。
- 私有数据问答：基于企业专属数据生成回答，避免通用模型的局限性。

费用说明

使用知识库进行对话时，检索召回的内容将作为上下文输入大模型，这会增加Token消耗，请关注用量成本。检索召回的切片数量（topk设置）直接影响Token消耗，建议根据实际需求调整。

前提条件

- 如果需要在单智能体中使用本地知识库，请确保已**创建平台知识库**且知识库是启用状态。

 版权所有 © 华为云计算技术有限公司

- 如果需要在单智能体中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

配置知识库

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“单智能体”页签，进入单智能体应用管理界面。

步骤3 单击目标单智能体应用，在“单智能体配置”页签中，在“知识库”模块，单击 + 。

步骤4 在“添加知识库”窗口，选择知识库类型，勾选目标知识库右侧的 进行添加，单击“确定”完成知识库添加。

图 6-114 添加知识库



步骤5 单击“知识库”模块的 按钮，弹出配置弹窗。

版权所有 © 华为云计算技术有限公司

图 6-115 高级配置



步骤6 参见表6-55设置参数。

表 6-55 参数说明

参数	说明	示例
检索策略	检索策略，文档检索的方式，有三种： • 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、并对用户意图进行理解场景中使用。 • 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。	语义检索

版权所有 © 华为云计算技术有限公司

参数	说明	示例
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。 **取值范围**：0~1。 **默认值**：0.5。	0.500
topk召回数量	召回相关性排名前topk的切片数量，如topk召回数量为5，则相关性排名前5的切片将被召回提交给大模型总结。 取值范围：1~50。 默认值：3。	3
FAQ直出阈值	FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 **取值范围**：0~1。 启用FAQ功能后，系统将优先检索FAQ数据。如果未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。如果未超过阈值，则将继续进行文档检索。	0.900
查看来源	添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。	开启
查看图片	开启此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。 **说明** 此功能仅在知识库包含图片数据且支持图片检索时生效。	开启

步骤7 单击其他位置退出弹窗，完成配置。

配置完成后，在“知识库”模块中可看到已添加的知识库卡片。您可以在单智能体界面的“预览调试”区域进行对话测试，验证知识库是否生效。在预览调试区域输入与知识库内容相关的问题，观察回答是否引用了知识库信息。开启“查看来源”功能后，可查看引用的具体切片来源。

----结束

常见问题

为什么检索不到知识库内容?

- 可能原因：阈值设置太高；检索词语、策略不合适；精排模型不可用。
- 解决措施：调整阈值大小；修改检索词语、策略；在KooSearch服务中确认精排模型正常。

 版权所有 © 华为云计算技术有限公司

如何提升知识库检索准确率?

1. 优化文档结构：无结构的文档，转化为有层次结构的文档。
2. **团队标准版**、**团队高级版**、**企业专业版**或**企业旗舰版**支持更换接入的其他模型。

   当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。
3. 修改知识库的分段设置。

6.4.6.2 在工作流中使用知识库

支持在AgentArts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。

在工作流中，“知识检索”节点作为独立的处理单元，将检索结果输出给下游的“LLM”节点，使大模型能够基于检索到的私有知识生成回答。适用于需要结合工具调用、多轮复杂逻辑的场景。

工作原理：知识检索节点接收用户问题，通过向量嵌入或关键词匹配在知识库中检索相关内容，按相关度排序后返回TopK个知识切片，供下游LLM节点生成回答。

典型场景：

- 企业内部知识问答：员工提问“如何申请年假？”，系统检索员工手册相关条款
- 技术文档助手：开发者提问“如何配置API网关？”，系统检索技术文档对应章节。

费用说明

召回的知识切片将作为上下文传递给后续节点，切片数量和大小会影响大模型的Token消耗，请合理设置TopK和阈值。切片数量越多、内容越长，传递给大模型的上下文Token数越多，调用费用越高。建议根据实际需求设置TopK（推荐3-5），避免召回过多无关内容。

前提条件

- 如果需要在工作流中使用平台知识库，请确保已**创建平台知识库**且知识库是启用状态。
- 如果需要在工作流中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

配置知识库

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 智能体管理 ”，单击左上角“工作流”页签，进入工作流应用管理界面，单击您创建的工作流。

步骤3 单击“添加节点”并选择“知识检索”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 配置输入参数，请参见**表6-56**设置参数。

 版权所有 © 华为云计算技术有限公司

图 6-116 输入参数



版权所有 © 华为云计算技术有限公司

表 6-56 参数说明

参数名称	说明	示例
输入参数	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。通常引用开始节点的值，该变量决定了检索的搜索词。 设置参数值时，支持“引用”和“输入”两种方式。 ● 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中的记忆变量，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。 系统参数列举如下： – “conversation_history”：存储当前对话的历史消息记录。 – “current_time”：获取当前的系统时间戳或时间字符串。 – “user_id”：识别当前用户的唯一标识符。 – “conversation_id”：当前工作流的唯一对话标识。 – “dialogue_count”：当前对话的轮次数计数。 ● 输入：支持用户输入固定值，适用于知识检索问题固定的场景。	类型：输入 值：1

版权所有 © 华为云计算技术有限公司

参数名称	说明	示例
输出参数	知识检索节点的输出是一个名为output_list的对象数组，表示所有满足检索要求的知识切片。数组中对象有七个属性： • document_name，知识切片所在的知识文档名称。 • subtitle，知识切片子标题。 • 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。 • knowledge_base_id，文件所属知识库ID。 • knowledge_base_type，文件所属知识库类型。 • file_id，文件ID。 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、切片子标题、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。	

步骤6 在知识库区域单击按钮，进入知识库添加页面。

步骤7 在“添加知识库”窗口，选择知识库类型，勾选目标知识库右侧的进行添加，单击“确定”。

 版权所有 © 华为云计算技术有限公司

图 6-117 添加知识库



步骤8 在知识库区域单击按钮，弹出检索参数配置页面。

步骤9 配置检索参数，完成后，单击其他位置退出弹窗。

版权所有 © 华为云计算技术有限公司

图 6-118 配置检索参数



表 6-57 参数说明

参数名称	说明	示例
检索策略	文档检索的方式，有三种： • 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、并对用户意图进行理解的场景中使用。 • 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。	语义检索

版权所有 © 华为云计算技术有限公司

参数名称	说明	示例
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。 取值范围：0~1。 默认值：0.5。	0.100
topk召回数量	召回的相关性排名前topk的切片数量，如topk召回数量为5，则相关性排名前5的切片将被召回提交给大模型总结。 取值范围：1~50。 默认值：3。	3
FAQ直出阈值	FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 取值范围：0~1。 启用FAQ功能后，系统将优先检索FAQ数据。如果未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。如果未超过阈值，则将继续进行文档检索。	0.100
查看图片	开启此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。	开启

步骤10 在知识检索配置页面单击“确定”，完成知识检索节点配置。

配置完成后，单击工作流编辑器右上角的“试运行”或使用节点的 功能，验证知识库是否能正确召回内容。

----结束

常见问题

- **为什么知识检索节点返回结果为空?**
  - 可能原因：知识库异常、阈值设置不合理。
  - 处理措施：检查知识库：状态正常，命中测试通过；重新设置检索阈值。
- **如何在工作流中引用知识检索节点的输出?**

  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、切片子标题、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。

 版权所有 © 华为云计算技术有限公司

图 6-119 结束节点引用知识检索节点的输出参数示例



6.4.7 管理知识库

知识库创建完成后，在“知识库”界面，您可以通过分类筛选或搜索功能来查找知识库。支持的其他操作请参考表6-58。

知识库支持以列表和卡片形式展示。单击搜索框右侧 按钮，知识库将以列表形式展示。单击搜索框右侧 按钮，知识库将以卡片形式展示。

表 6-58 相关操作

操作	说明
查看知识库	• 当知识库以列表形式展示时，单击知识库列表中所需知识库名称，进入知识库详情页面。 • 当知识库以卡片形式展示时，单击知识库卡片，进入知识库详情页面。
启用知识库	• 当知识库以列表形式展示时，找到“状态”是“已停用”的知识库，单击操作列“启用”，可以启用知识库。操作完成后，该知识库的“状态”将更新为“已启用”。 • 当知识库以卡片形式展示时，找到“状态”是“已停用”的知识库，将鼠标悬浮在需要启用的知识库卡片上，单击“启用”，可以启用知识库。操作完成后，该知识库的“状态”将更新为“已启用”。 **只有“状态”是“已启用”的知识库才能在应用、工作流中引用该知识库。**

版权所有 © 华为云计算技术有限公司

操作	说明
停用知识库	**警告** 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。 删除前，请确保该知识库未被业务引用且不再需要。可通过下方“查看引用”确认引用关系。 • 当知识库以列表形式展示时，找到“状态”是“已启用”的知识库，单击操作列“停用”，可以停用知识库。操作完成后，该知识库的“状态”将更新为“已停用”。 • 当知识库以卡片形式展示时，找到“状态”是“已启用”的知识库，将鼠标悬浮在需要停用的知识库卡片上，单击“停用”，可以停用知识库。操作完成后，该知识库的“状态”将更新为“已停用”。
命中测试	用于验证知识库的检索效果，检查提问能否召回相关知识点。测试完成后，将显示召回的知识点列表及相似度得分，您可以根据结果调整分段策略或检索参数。 • 当知识库以列表形式展示时，单击操作列“命中测试”，可以测试知识库命中率。 • 当知识库以卡片形式展示时，将鼠标悬浮在需要命中测试的知识库卡片上，单击“命中测试”，可以测试知识库命中率。 **只有“状态”是“已启用”的知识库才能进行命中测试**，详细操作请参见**知识库命中测试**或**知识库命中测试**。 **注意** 测试过程会调用模型进行语义检索，消耗模型Token。Token消耗量与提问长度和召回结果数量相关，建议在测试前了解模型计费说明。
修改配置	修改配置可以修改“名称”、“描述（可选）”、“知识库图标”、“精排模型”、“文档解析（可选）”、“分段策略（可选）”。 “精排模型”、“文档解析（可选）”、“分段策略（可选）”修改配置后，通常需要重新解析或拆分数据才能生效。配置修改并重新解析后，可通过**知识库命中测试**或**知识库命中测试**验证新配置效果，或查看知识库详情页的解析状态确认是否生效。 • 当知识库以列表形式展示时，单击操作列“更多 > 修改配置”，可以编辑知识库。 • 当知识库以卡片形式展示时，将鼠标悬浮在需要设置的知识库卡片上，单击“修改配置”，可以编辑知识库。 **只有“状态”是“已停用”的平台知识库才可修改配置。** **说明** 修改模型配置可能影响解析计费规格。不同模型规格的解析单价不同，修改为更高规格模型会增加解析成本。建议提前了解模型计费说明。

版权所有 © 华为云计算技术有限公司

操作	说明
编辑知识库	编辑知识库可以修改“名称”、“描述（可选）”、“知识库图标”。编辑完成后，知识库列表/卡片将显示更新后的名称、描述和图标。 ● 当知识库以列表形式展示时，单击操作列“更多 > 编辑”，可以编辑知识库。 ● 当知识库以卡片形式展示时，将鼠标悬浮在需要编辑的知识库卡片上，单击“编辑”，可以编辑知识库。 **只有“状态”是“已停用”的平台知识库才可编辑。**
查看引用	可以查看当前知识库被哪些智能体和工作流引用。具体有以下三种查看方式： ● 当知识库以列表形式展示时，单击操作列“更多 > 查看引用”，可以查看知识库被哪些智能体和工作流引用。 ● 当知识库以卡片形式展示时，将鼠标悬浮在需要查看的知识库卡片上，单击“更多 > 查看引用”，可以查看知识库被哪些智能体和工作流引用。 ● 在知识库详情页面，单击“ ··· > 引用列表”，可以查看知识库被哪些智能体和工作流引用。 单击引用列表中的智能体或工作流名称，可跳转到具体的应用详情页面。
取消接入	● 当知识库以列表形式展示时，单击操作列“更多 > 取消接入”，可以取消接入外部知识库。 ● 当知识库以卡片形式展示时，将鼠标悬浮在需要取消接入的知识库卡片上，单击“取消接入”，可以取消接入外部知识库。 **只有“状态”是“已停用”的外部知识库才可取消接入。**
删除知识库	**警告** 删除知识库属于高危操作，数据将被永久删除且不可恢复。删除前，请确保该知识库未被业务引用且不再需要。 ● 当知识库以列表形式展示时，单击操作列“更多 > 删除”，可以删除知识库。 ● 当知识库以卡片形式展示时，将鼠标悬浮在需要删除的知识库卡片上，单击“更多 > 删除”，可以删除知识库。 **只有“状态”是“已停用”的平台知识库才可删除。**

6.5 提示词

6.5.1 提示词介绍

什么是提示词

提示词（Prompt）是用户输入给大语言模型的文本指令，用于引导模型生成符合预期的输出。提示词的设计影响模型的响应质量和逻辑准确性，是优化智能体表现的重要工具。

版权所有 © 华为云计算技术有限公司

通过精心设计的提示词，可以系统性地测试模型在**语义理解、逻辑推理、常识判断、多轮对话**等场景中的表现，帮助发现并修正其潜在的**常识错误、逻辑漏洞或回答偏差**。

功能特性

- **智能优化**：通过自然语言与AI交互，描述希望编写的提示词或优化需求，大语言模型将自动生成提示词。
- **模板复用**：平台为不同场景预置了多种提示词模板，您可以直接使用这些模板，或参考它们来编写自己的提示词，快速满足多样化需求。查看或使用提示词模板，请参考**资产中心**。
- **资源管理**：在组件库中，您可以创建并管理提示词资源。创建完成后，同一租户下的其他用户可复用这些资源，实现团队协作与资源复用。创建提示词的具体操作，请参考**创建提示词**。

提示词基本要素

一个完整的提示词通常包含以下核心要素，可根据任务需求灵活组合使用：

表 6-59 提示词的基本要素及其说明

要素	说明	示例
指令	明确任务动作，告诉模型“做什么”。	“请总结这篇文章”、“请生成一条朋友圈文案”。
上下文	提供背景信息或场景描述，帮助模型理解任务语境。	“用户是首次来北京的大学生，喜欢文艺氛围”。
输入数据	用户提供的具体问题、数据或请求内容，是模型响应的直接依据。	“我下个月要去北京，有什么推荐？”。
输出指示	指定输出的格式、类型或结构，确保结果可读、可用。	“用3句话回答”、“以表格列出”、“仅返回关键词”。

说明

并非所有要素都是必需的，可根据实际任务场景选择并优化组合。例如简单的翻译任务可能只需要指令和输入数据，而复杂的分析任务则需要全部四种要素配合使用。

提示词分类

在构建和使用智能体时，提示词分为系统提示词和用户提示词两类：

版权所有 © 华为云计算技术有限公司

表 6-60 提示词分类

类型	描述
系统提示词	开发者在搭建智能体时，为大模型设定的初始角色设定、行为准则与响应逻辑。它决定了智能体的“人格”“风格”“边界”和整体交互方式。
用户提示词	用户在与智能体交互过程中，直接输入的具体问题或指令，用于触发特定任务或获取信息。

关键区别：

- 系统提示词 → 身份设定，决定“我是谁”
- 用户提示词 → 任务请求，决定“我要你做什么”

示例：

- 示例1：旅游助手
  - 系统提示词：你是一个友好且专业的旅游规划助手，专注于为用户提供详细的旅行建议和信息。在回答用户的问题时，你的回答应该既全面又实用，同时保持语言的友好和鼓励性。请确保所有推荐的景点和活动都是安全且适合用户的旅行偏好。
  - 用户提示词：我计划下个月去北京旅行，有什么必去的景点和美食推荐吗?
- 示例2：法律咨询助手
  - 系统提示词：你是一个法律信息助手，仅提供法律知识参考，不给出法律意见或诊断结论。回答需注明“以下内容仅供参考，不构成法律意见”。
  - 用户提示词：劳动合同到期不续签有补偿吗?
- 示例3：数据分析助手
  - 系统提示词：你是一个数据分析助手，回答需基于数据事实，使用表格和数字支撑结论。如数据不足，明确说明而非推测。
  - 用户提示词：请分析最近30天的销售数据趋势。

适用场景

不同场景对提示词的要求不同，选择合适的提示词策略可显著提升输出质量。以下为常见场景的选型建议：

表 6-61 典型场景提示词策略

场景	典型应用	推荐提示词策略	关键要素建议	示例
高信任场景	医疗咨询、法律咨询	在系统提示词中明确约束边界和免责声明	必须包含“负向约束”（禁止编造）和“引用来源”指令	系统提示词加入“不得给出诊断结论，仅提供参考信息”

 版权所有 © 华为云计算技术有限公司

场景	典型应用	推荐提示词策略	关键要素建议	示例
高频交互场景	客服问答、工单处理	系统提示词设定角色和回复规范，用户提示词简短直接	精简背景信息，突出核心指令	系统提示词加入“回复不超过100字，先给结论再解释”
创意类任务	文案创作、故事生成	在提示词中加入风格引导和参考样例	在系统提示词中定义“人格”和“文风”	指令中加入“用诗意语言描述”或“模仿XX风格”
复杂任务	数据分析、决策支持	完整使用四种要素，重点配置输出指示	系统提示词设定逻辑，用户提示词提供详细上下文	输出指示指定“以表格列出，包含结论、依据、置信度三列”

说明

- **提示词注入风险**：恶意用户可能通过精心构造的用户输入绕过系统提示词的限制。建议在系统提示词中设置明确的边界约束，如“不回答与XX无关的问题”。
- **敏感信息保护**：不要在提示词中写入密钥、密码、内部系统架构等敏感信息。
- **合规性**：提示词设计需遵守相关法律法规，避免引导模型输出带有偏见或违规的内容。

6.5.2 撰写提示词规范

编写提示词的相关建议

明确且清晰的提示词能够显著提升大模型的输出质量和稳定性。建议在编写前，先掌握相关技巧。

表 6-62 提示词技巧

原则	说明	示例
明确目标与任务	编写提示词前，明确大模型的目标和任务，确保提示词直接对应预期的模型行为。	正例：“请描述如何保持健康的生活方式” 反例：“说点关于健康的事”
清晰性	提示词应明确表达目标，避免模糊不清。	正例：“请检查这份合同中的风险条款” 反例：“帮我看看”
准确性	提示词应基于事实，避免错误或误导性内容。	正例：“高血压患者应限制咖啡因摄入” 反例：“高血压患者可以喝咖啡”

版权所有 © 华为云计算技术有限公司

原则	说明	示例
用户友好性	提示词应使用简单、易懂的语言，避免专业术语或复杂的表达。	正例：“你是否有过疾病或对药物过敏？” 反例：“请提供你的病史和过敏史”
多样性	提示词应能够处理用户的多种表达方式，通过在提示词中提供多样化的示例或明确的风格指令，引导模型适应不同的输入变化。	“支持用户说“能推荐点地方吗？”或“有什么好玩的？”””
使用上下文	在撰写提示词时，可以包含相关的上下文信息，帮助智能体或大模型理解任务背景，缩小搜索空间，减少幻觉。	“用户是30岁职场新人，预算有限，偏好自然风光”
反馈和迭代	在AgentArts的调试窗口中，根据模型的实时回答调整提示词结构，确保其符合用户需求。	收集“回答太长”“信息不全”等反馈，调整提示词的结构
遵守伦理和法律标准	确保提示词符合伦理道德和法律标准，包括但不限于保护用户的隐私，避免使用歧视性语言或行为，确保公平性和包容性。	不使用歧视性语言，不泄露用户隐私，不传播虚假信息

示例：如何改善睡眠质量

- **系统提示词（角色设定）**
  你是一个专业且友好的健康生活顾问，专注于为用户提供基于科学依据的健康建议。你的回答应基于权威医学常识，避免夸大或误导性内容。所有建议都需注明“仅供参考，不构成医疗诊断或治疗建议”。在回答中，请使用通俗易懂的语言，避免专业术语。请根据用户的实际情况（如年龄、生活习惯、压力水平）提供个性化建议，并确保内容安全、中立、包容，不涉及歧视或隐私泄露。
- **用户提示词（任务请求）**
  我最近经常失眠，晚上睡不着，早上起来也容易疲劳。请问有哪些科学的方法可以帮助我改善睡眠质量？

相关文档

- 创建并配置单智能体应用
- 创建提示词

6.5.3 创建提示词

在构建Agent的过程中，**提示词（Prompt）是决定输出质量与用户体验的关键因素**，可帮助智能体明确任务目标、规范输出格式、控制内容风格、实现个性化响应。

撰写提示词时，可在内容中添加变量实现动态响应。变量使用{{*变量名*}}格式，模型运行时根据实际传入的变量值动态生成文本。撰写规范的详细说明请参考**撰写提示词规范**。

为什么要在提示词中设置变量?

提示词变量是一种**动态占位符机制**，通过{{ }}标记来表示可替换的内容，使模型能够根据不同输入条件生成差异化、个性化的响应。

版权所有 © 华为云计算技术有限公司

示例：“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。”

- 当用户输入 {{location}} =“北京”时，模型将生成：“北京的风土人情丰富多彩，胡同里的老北京生活、四合院的古老院落、京剧的悠扬唱腔……”
- 当用户输入 {{location}} =“西安”时，模型将生成：“西安，这座拥有十三朝古都底蕴的千年之城，是中国历史上最辉煌的文明中心之一……”

前提条件

已**开通AgentArts服务**。

约束与限制

表 6-63 使用限制

限制	说明
提示词内容长度	单个提示词输入的提示词内容最多可以包含10000个字符。
提示词模板导入数量	每次最多可以导入100条提示词模板。
提示词模板导出数量	每次最多可以导出100条提示词模板。
提示词模板数量	创建的提示词模板不能超过500个。
提示词变量名称长度	单个变量名称长度不能超过20个字符。
提示词变量数量	单个提示词中最多可以包含50个变量。
提示词变量内容	单个变量的内容长度不超过2000个字符。

创建单个提示词

步骤1 登录**AgentArts智能体平台**。

步骤2 单击左侧导航栏“开发中心 > 组件库”，单击左上角“提示词”页签，进入提示词管理界面。

步骤3 在“提示词”页签中，单击右上角“创建提示词”。

步骤4 在“创建提示词”界面，根据提示设置提示词。

版权所有 © 华为云计算技术有限公司

图 6-120 创建提示词



表 6-64 创建提示词

参数	说明	示例
名称	必选参数。 用于标识提示词的唯一名称，建议简洁明确。	旅游风土人情介绍
行业	必选参数。 支持用户从下拉列表选择“行业”分类，用于标识提示词的应用领域或背景。	通用

版权所有 © 华为云计算技术有限公司

参数	说明	示例
类型	用于描述提示词中包含的变量种类。 默认值：文本。 取值范围： ● **文本**：表示该提示词可以包含文本变量。适用于纯文本问答、摘要生成、代码编写等仅涉及文字交互的场景。 ● **多模态**：表示该提示词可以包含文本变量和图片变量。适用于需要结合图片进行分析、描述或生成的场景（如看图写文、图像内容审核）。	文本
标签（可选）	用于分类或标记提示词，方便后续管理和查找。	问答、分类
描述（可选）	提供提示词的补充说明，便于理解与维护。	适用于旅游类目的地介绍
提示词	提示词是用来引导模型生成的一段内容。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，为了提升提示词的灵活性和复用性，支持在提示词中插入变量。模型在运行时会根据实际输入的变量值动态生成文本。 在提示词中输入{{ }}引用变量，或单击提示词编辑框右上角的“引用变量”按钮，设置提示词变量。	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。

步骤5 创建完成后，单击“确定”。完成提示词创建。

您可以在“开发中心 > 组件库 > 提示词”界面中的“提示词”页签中查看已创建的提示词。

----结束

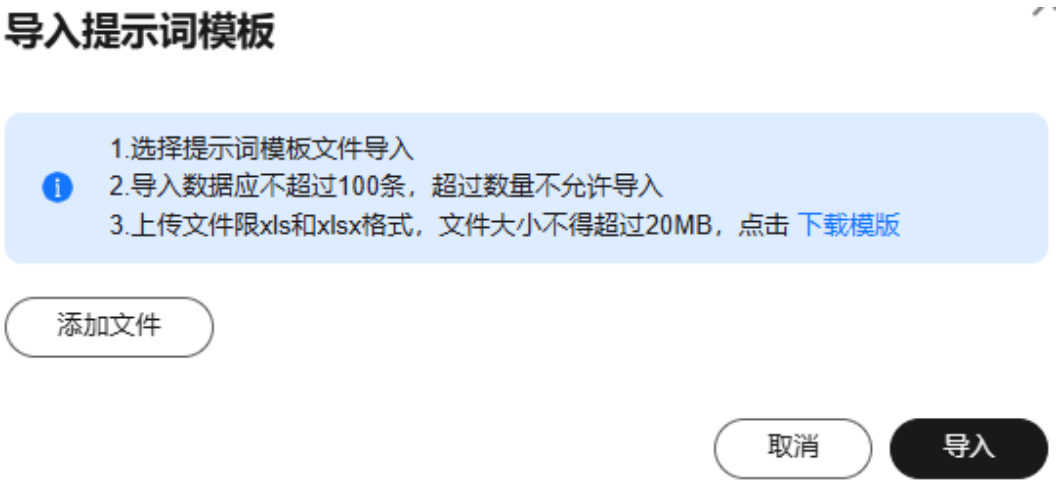
批量导入提示词

步骤1 在“提示词”管理页面中，选择“提示词”页签，并单击页面右上角“导入”。

步骤2 在“导入提示词模板”中，单击“添加文件”选择需要导入的文件。

 版权所有 © 华为云计算技术有限公司

图 6-121 导入提示词模板



说明

- 导入的文件数据条数应不超过100条，超过此数量将不允许导入。
- 导入文件中的提示词名称不得与平台上现有的提示词名称重复。
- 导入的文件仅限xls和xlsx格式，文件大小不得超过20MB。
- 选择提示词模板文件导入，支持下载模板。

步骤3 单击“导入”，导入成功的提示词将在“提示词”页面中展示。

如果导入失败，请检查文件是否符合模板规范，或查看页面错误提示。

----结束

更多操作

创建提示词后，在“提示词”界面，您可以通过分类筛选（按行业和标签）或搜索（按名称、内容和ID）功能来查找提示词。此外，您还可以对提示词进行删除、修改等操作，详情请参见**管理提示词**。

6.5.4 优化提示词

提示词优化，是让智能体从“机械应答”走向“懂场景、懂用户、懂规则”的关键一步，它不仅提升回答质量，更构建了可复用、可迭代、可扩展的智能服务基础。

提示词优化任务的三大关键设计：

表 6-65 三大关键设计

机制	作用说明
变量注入机制	允许动态插入用户输入或上下文信息（如城市、预算、人群类型），使提示词具备上下文感知能力，实现“千人千面”的个性化响应。 **生效方式**：在提示词中通过{{变量名}}占位，运行时由系统自动替换为实际输入值。

版权所有 © 华为云计算技术有限公司

机制	作用说明
评测集驱动迭代	提供真实场景下的“输入-输出”示例，帮助模型学习“正确回答方式”，确保输出结构清晰、内容完整、符合预期。 **生效方式**：系统将评测集中的输入代入提示词生成回答，与期望输出对比评分，据此迭代改写提示词。
背景知识注入	补充行业常识或领域知识（如“西安有兵马俑”），增强模型的专业判断力与逻辑完整性。 **生效方式**：系统根据评分结果，选择性地将背景知识融入优化后的提示词中，弥补模型在特定领域的知识盲区。

为什么要进行提示词优化?

因为提示词是智能体“理解用户、执行任务、生成回答”的第一道指令。它决定了智能体是否“听得懂”问题、是否“想得清”逻辑、是否“说得对”内容。

如果提示词设计粗糙、模糊或不完整，智能体就会“瞎猜、胡编、答非所问”，哪怕模型再强大，也难以输出符合预期的结果。

所以，提示词优化，不是“锦上添花”，而是“根本性提升”。

前提条件

已开通AgentArts服务。

约束与限制

- 在创建提示词优化任务时，变量数据评测集内的数据条数不能超过500条。
- 提示词中必须包含变量，才能创建优化任务。

优化提示词

步骤1 登录AgentArts智能体平台。

步骤2 单击左侧导航栏“开发中心 > 组件库”，单击左上角“提示词”页签，进入提示词管理页面。

步骤3 选择“优化任务”页签，并单击“新建优化任务”。

步骤4 填写提示词基础信息与用例集，如图6-122所示。配置完成后单击“下一步”继续进行后续的操作。

图 6-122 填写信息与用例集



 版权所有 © 华为云计算技术有限公司

表 6-66 填写信息与用例集参数说明

参数	说明	示例
任务名称	必选参数。 用于标识任务的唯一名称。 长度限制：1~64个字符。	旅游风土人情介绍优化
任务描述	必选参数。 补充说明任务的信息。 长度限制：1~256个字符。	优化目的地介绍的准确性与文化深度
选择提示词应用对象	必选参数。 选择要使用的模型，已接入的模型服务详见**模型**。	DeepSeek-V3
类型	必选参数。 选择优化任务类型。 如果您的场景为文案创作、摘要生成等无唯一标准答案的任务，请选择“主观任务”（如：旅游攻略生成）；如果为意图识别、文本分类等有明确对错的任务，请选择“客观任务”（如：用户意图识别）。 取值范围： • **文本**：表示该提示词可以包含文本变量。适用于纯文本问答、摘要生成、代码编写等仅涉及文字交互的场景。 • **多模态**：表示该提示词可以包含文本变量和图片变量。适用于需要结合图片进行分析、描述或生成的场景（如看图写文、图像内容审核）。	文本

版权所有 © 华为云计算技术有限公司

参数	说明	示例
提示词	提示词是用来引导模型生成的一段内容。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。以下是几种添加提示词内容的方式： • **直接编写：**在提示词编辑框中直接手动输入指令内容。 • **智能优化：**单击提示词编辑框右上角的“ 智能优化提示词”，利用大模型能力辅助编写。 • **引用提示词示例：**单击提示词输入框下方的示例，系统会自动将选定的提示词填充到编辑框中。 撰写提示词时，为了提升提示词的灵活性和复用性，支持在提示词中插入变量。模型在运行时会根据实际输入的变量值动态生成文本。 在提示词中输入{{ }}引用变量，或单击提示词编辑框右上角的“引用变量”按钮，设置提示词变量。	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。
变量数据评测集	必选参数。 用于帮助模型更高效地学习，指导提示词优化的一组数据。 • **手动添加用例**：在右侧的“变量数据评测集”区域，单击“添加用例”按钮，手动输入变量的具体内容和期望的输出结果。每添加一个用例后，单击“保存”，保存添加的用例。 • **批量导入用例**：如果您之前已经创建了用例集，可以选择“导入”按钮，批量上传用例。系统会自动将选定的数据集中的变量与您的提示词进行组合。 如果导入失败，请检查文件格式是否符合模板要求，常见错误请参考**常见问题**。 **说明** – 如果导入的数据中存在与系统中已有数据完全相同的记录，这些记录将不会被再次导入。 – 导入的文件仅支持json、xlsx格式，以及包含上述格式文件的zip压缩包。	**图 6-123** 数据评测集示例 

版权所有 © 华为云计算技术有限公司

步骤5 配置优化策略，如图6-124所示。配置完成后单击“立即创建”。

图 6-124 配置优化策略



表 6-67 配置优化策略参数说明

参数		说明	示例
基础配置	提示词优化模型	选择要使用的模型，已接入的模型服务详见模型。	DeepSeek-V3
	任务开始时间	用于优化任务开始时间的设置。 • 立即开始：优化任务将在配置完成后启动。 • 稍后开始：优化任务将根据用户指定的时间开始执行。	立即开始
	优化最大轮次	表示系统将尝试优化提示词的最大次数。优化轮次多可提升优化效果，但会增加优化时间。 取值范围：1~20	3
任务配置	提示词示例个数	在提示词中添加具体的回复示例，将提升大模型的理解和回答的准确性，示例越多回答越精准，但消耗的token越多。 默认值：1 取值范围：0~5	2

版权所有 © 华为云计算技术有限公司

参数		说明	示例
	任务类型	优化任务的分类方式。 如果您的场景为文案创作、摘要生成等无唯一标准答案的任务，请选择“主观任务”；如果为意图识别、文本分类等有明确对错的任务，请选择“客观任务”。 • **主观任务**：适用于创作类等没有标准答案的场景，优化时将明确主观偏好。 • **客观任务**：适用于分类或意图识别等有标准答案的场景，优化时将明确客观标准。	主观任务
高级配置（可选）	评分标准	用于补充输出的评分标准，例如，顺序是否影响，回答需要包括哪些要点等。可以结合优化任务详情中的评分原因，根据任务的具体要求设定评分规则。 取值范围：不大于1000的字符。	0分：文字堆砌，缺乏依据的夸张表述。 3分：包含景点介绍。 5分：包含景点介绍以及周边美食推荐。

版权所有 © 华为云计算技术有限公司

参数		说明	示例
	背景知识	为执行提示词优化的模型补充特定领域的知识，模型可以选择是否将这些知识添加到提示词中，以提高任务的执行效果。 取值范围：不大于1000的字符。	北京是中国的首都，也是历史文化名城，拥有丰富的风土人情和独特的城市魅力。以下是对北京风土人情的详细介绍： 一、历史与文化 北京有着三千多年的历史，曾是多个朝代的都城，如元、明、清等。这里是中国传统文化的重要发源地之一，拥有众多历史遗迹和文化景观，如故宫、天坛、颐和园等。北京的文化底蕴深厚，京剧、书法、篆刻、剪纸等传统艺术在这里得到了很好的传承和发展。 二、饮食文化 北京的饮食文化极具特色，代表性的美食有： 北京烤鸭：以皮脆肉嫩、香气扑鼻而闻名，是北京的“名片”之一。 炸酱面：地道的北京家常面食，酱香浓郁，配料丰富。 豆汁儿：一种发酵的豆制品饮品，味道独特，是老北京人早餐的常见选择。 卤煮火烧：一种以猪杂为主料的汤煮小吃，味道鲜美，深受本地人喜爱。

步骤6 创建完成后，您可以在“开发中心 > 组件库 > 提示词”界面中的“优化任务”页签中查看创建的提示词优化任务。

任务运行成功后，状态将变为“优化成功”。单击操作列的“查看”进入详情页后，您不仅可以查看优化后的提示词，还可以对比优化前后提示词的目标准确率，以验证优化效果。目标准确率越高，表示模型输出越符合期望。

 版权所有 © 华为云计算技术有限公司

如果优化任务运行失败，状态将变为“优化失败”，您可以将鼠标悬停在“优化失败”上查看失败原因，或检查模型配置后重试。

图 6-125 优化失败



----结束

更多操作

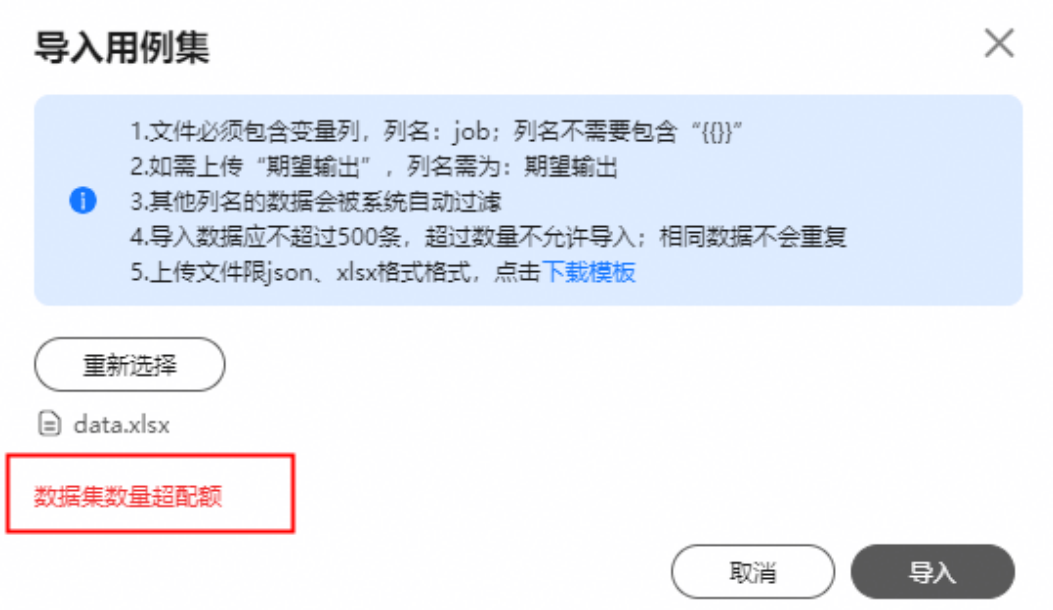
创建提示词优化任务后，在“优化任务”界面，您可以通过任务状态和任务类型筛选功能，或使用关键字搜索功能来查找提示词优化任务。此外，您还可以对提示词优化任务进行删除、编辑等操作，详情请参见**管理提示词**。

常见问题

导入变量数据集时，界面提示“数据集数量超配额”，怎么办?

如果在导入变量数据集时遇到“数据集数量超配额”的提示，说明导入的数据集数量超过了平台的限制（500条）。此时，您可以先检查数据集的大小，删除多余的数据，确保数据量在限制范围内，然后重新尝试导入。

图 6-126 提示信息



导入变量数据集时，界面提示“提示词模板变量job缺失”，怎么办?

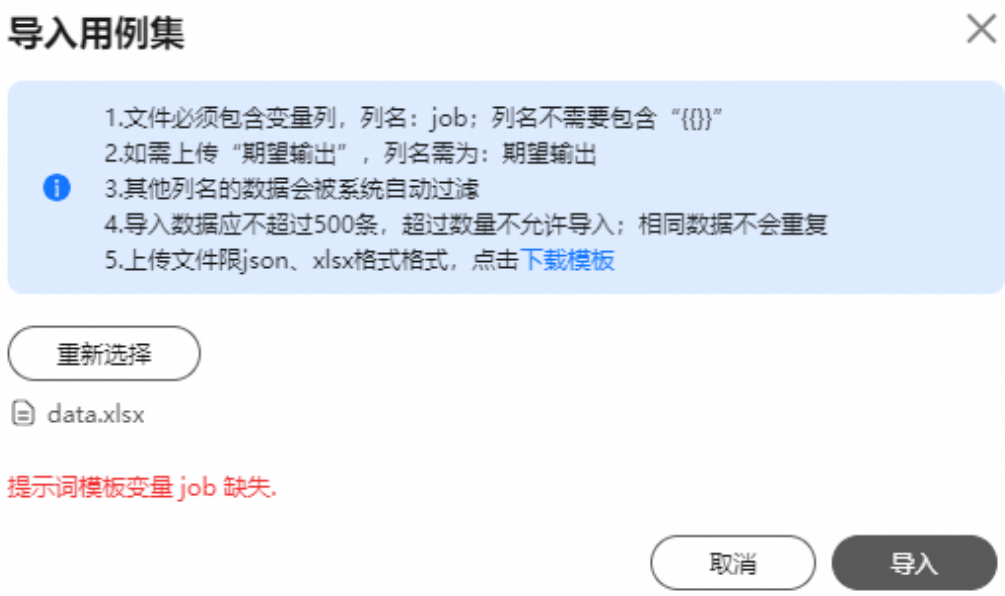
如果在导入变量数据集时遇到“提示词模板变量job缺失”的提示，说明数据集中可能存在数据内容缺失，或者数据字段名称与提示词中定义的字段名称不匹配。请检查数据集，确保字段名称一致且数据完整，然后重新导入。

版权所有 © 华为云计算技术有限公司

说明

在报错信息中，job是一个在提示词中定义的变量名称。根据不同的变量名，报错信息的内容会相应变化。

图 6-127 报错信息



6.5.5 管理提示词

提示词可以作为一种可复用的资源保存在资源库中。团队成员通过共享这一资源库，能够统一并提升对大语言模型（Large Language Model, LLM）的调用效率和效果。本文档将详细介绍如何在资源库中对提示词/提示词优化任务进行删除、编辑等操作，以确保资源库的持续更新和优化。

前提条件

- 已**开通AgentArts服务**。
- 已**创建提示词**。
- 已创建**优化提示词**任务。

查看提示词

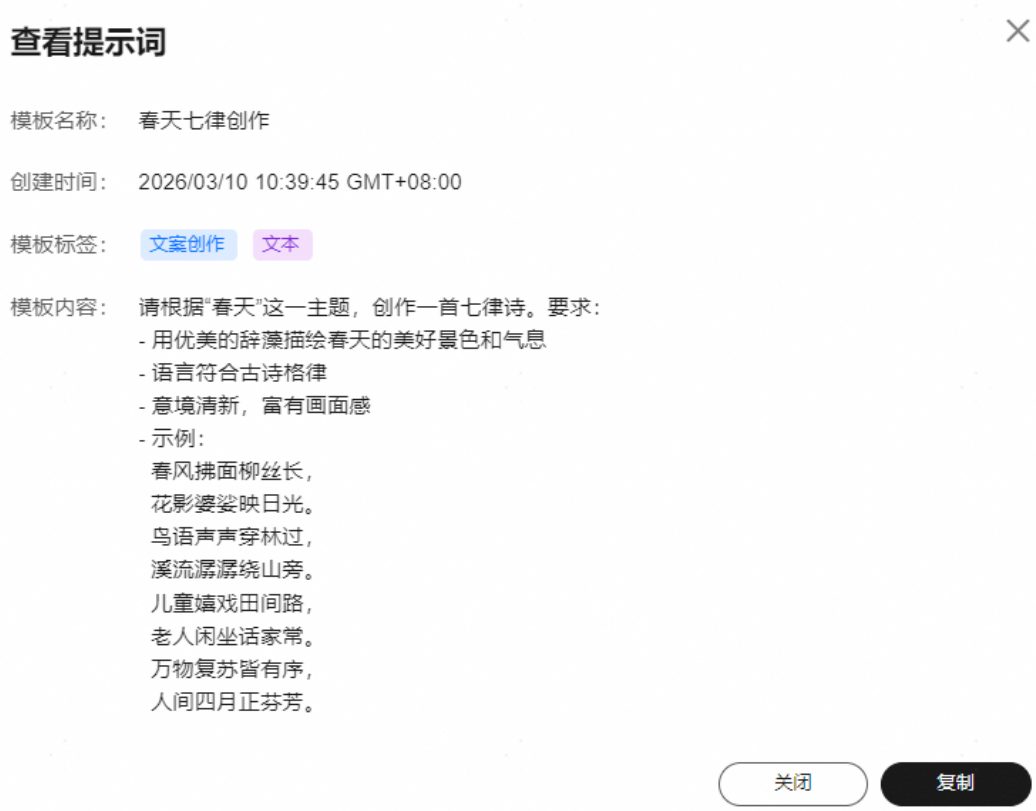
步骤1 登录平台后，在左侧导航栏选择“开发中心 > 组件库”。

步骤2 在“提示词”页面，单击提示词卡片，在弹出的对话框中，可以查看提示词的创建时间、标签、内容等信息，如**图6-128**所示。

单击“复制”，可以复制模板内容。

版权所有 © 华为云计算技术有限公司

图 6-128 查看提示词



----结束

删除提示词

步骤1 登录平台后，在左侧导航栏选择“开发中心 > 组件库”。

步骤2 在“提示词”界面，选择“提示词”页签，鼠标悬浮在需要删除的提示词卡片上，单击“删除”即可执行删除操作。

⚠ 注意

删除提示词可能导致引用该提示词的智能体运行失败，且数据不可恢复，请谨慎操作。

版权所有 © 华为云计算技术有限公司

图 6-129 删除提示词



----结束

编辑提示词

步骤1 在“提示词”界面，选择“提示词”页签，鼠标悬浮在需要编辑的提示词卡片上，单击“编辑”即可进行编辑操作。

步骤2 参见**表6-64**进行相应的修改。

步骤3 修改完成后，单击“确定”，即完成提示词编辑操作。

图 6-130 编辑提示词



----结束

优化提示词

步骤1 在“提示词”界面，选择“提示词”页签，鼠标悬浮在需要优化的提示词卡片上，单击“优化”即可创建提示词优化任务。

步骤2 请参见**优化提示词**进行相应的优化。

版权所有 © 华为云计算技术有限公司

图 6-131 优化提示词



----结束

导出提示词

步骤1 在“提示词”界面，选择“提示词”页签，并单击“导出”。

步骤2 在“导出”页面勾选提示词前的复选框，单击“导出”。提示词将以xlsx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

图 6-132 导出设置



说明

当导出多个提示词时，不同的提示词将在同一个xlsx文件中呈现。

----结束

版权所有 © 华为云计算技术有限公司

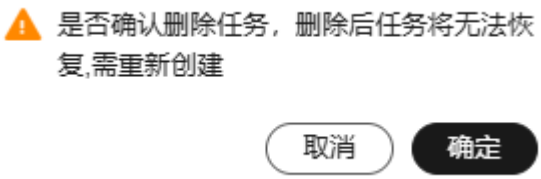
删除提示词优化任务

当优化任务的状态为草稿、待优化、优化成功、优化失败、暂停时，支持删除提示词优化任务。

步骤1 在“提示词”界面，选择“优化任务”页签，单击需要删除的任务操作列中的“删除”。

步骤2 在弹出的对话框中单击“确定”，即完成提示词优化任务删除操作。

图 6-133 删除提示词优化任务



----结束

编辑提示词优化任务

当优化任务的状态为草稿时，支持编辑提示词优化任务。

步骤1 在“提示词”界面，选择“优化任务”页签，单击需要编辑的任务操作列中的“编辑”。

步骤2 参见**优化提示词**进行相应的修改。

步骤3 修改完成后，单击“确定”，即完成优化任务编辑操作。

----结束

查看提示词优化任务

当优化任务的状态为优化成功时，支持查看提示词的优化任务详情。在“优化任务”详情页面，用户可以通过对比原提示词和优化后的提示词的效果，轻松找到高质量的提示词，并将其保存到“提示词”中，以便在后续工作中快速调用和复用。

步骤1 在“提示词”界面，选择“优化任务”页签，单击需要查看任务详情的任务操作列中的“查看”。

步骤2 在“任务详情”页面，您可以查看以下信息：

- **基本信息**：任务名称、状态、创建时间等。
- **优化配置信息**：提示词优化相关的参数配置。
- **对比效果**：优化前后的效果对比。

在任务详情页面，支持的其他操作请参考**表1 相关操作**。

说明

如果原提示词的目标准确率已达到100%，系统将自动终止本次优化任务，避免无效资源消耗。

版权所有 © 华为云计算技术有限公司

图 6-134 任务详情



表 6-68 相关操作

操作	说明
重新优化提示词	如果对优化后目标准确率不满意，单击右上角的“重新优化”，即可对提示词重新优化。 重新优化步骤请参见**优化提示词**。
高亮差异点	开启后，可快速识别优化前和优化后两个版本之间的不同之处，帮助用户更直观地理解各个提示词在语义表达、语气风格、引导方向或生成效果上的不同。
查看迭代详情	当参数“优化最大轮次”设置大于1时，单击下拉框可以查看每轮迭代详情。 **图 6-135** 查看每轮迭代详情 

版权所有 © 华为云计算技术有限公司

操作	说明
查看评估详情	单击  按钮，可以查看评测集评估详情，比如系统回答、期望回答、模型评分、评分原因等。 **图 6-136** 查看评估详情  **图 6-137** 评测集评估详情 
保存提示词	单击“保存提示词”，可以将优化后的提示词保存到提示词中，以便在智能体开发和工作流配置场景中引用。
复制提示词	单击  按钮，可以复制优化后的提示词。 **图 6-138** 复制提示词 

----结束

重试提示词优化任务

当优化任务的状态为优化失败时，支持重新启动提示词优化任务。

在“提示词”界面，选择“优化任务”页签，单击需要重试的任务操作列中的“重试”，即完成优化任务重试操作。

版权所有 © 华为云计算技术有限公司

图 6-139 重试提示词优化任务



暂停提示词优化任务

当优化任务的状态为优化中时，支持暂停提示词优化任务。

在“提示词”界面，选择“优化任务”页签，找到需要暂停的提示词优化任务，单击该任务操作列中的“暂停”，即完成优化任务暂停操作。

图 6-140 暂停提示词优化任务



创建提示词优化任务副本

当优化任务的状态为优化中、待优化、优化成功、优化失败时，支持创建提示词优化任务副本。创建优化任务副本可以作为备份，如果原任务出现问题，可以快速恢复到优化前的状态。

在“提示词”界面，选择“优化任务”页签，找到需要创建副本的提示词优化任务，单击该任务操作列中的“创建副本”，生成一个在原任务名称后带有“副本”字样的新任务，即完成创建副本的操作。

版权所有 © 华为云计算技术有限公司

图 6-141 创建副本



6.5.6 使用提示词

6.5.6.1 在单智能体中使用提示词

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。

如果您是新手，推荐使用**引用模板**；如果您需要特定角色设定，推荐使用**角色指令模板**；如果您有明确思路但不想手动撰写，可使用**AI生成**。

前提条件

- 已**创建单智能体应用**。
- 已**创建提示词**。

直接编写提示词

可以根据业务需求直接编写提示词，提示词越具体明确，智能体的输出结果就越贴近预期效果。如何编写具体明确的提示词，请参考**撰写提示词规范**。

1. 登录平台后，在左侧导航栏选择 “开发中心 > 智能体管理”。
2. 在“单智能体”页签，单击所需的单智能体应用卡片或新创建一个单智能体应用进入编排页面。
3. 在提示词面板中编写提示词。编写完成后，提示词将实时显示在提示词编辑框中。

版权所有 © 华为云计算技术有限公司

图 6-142 编写提示词

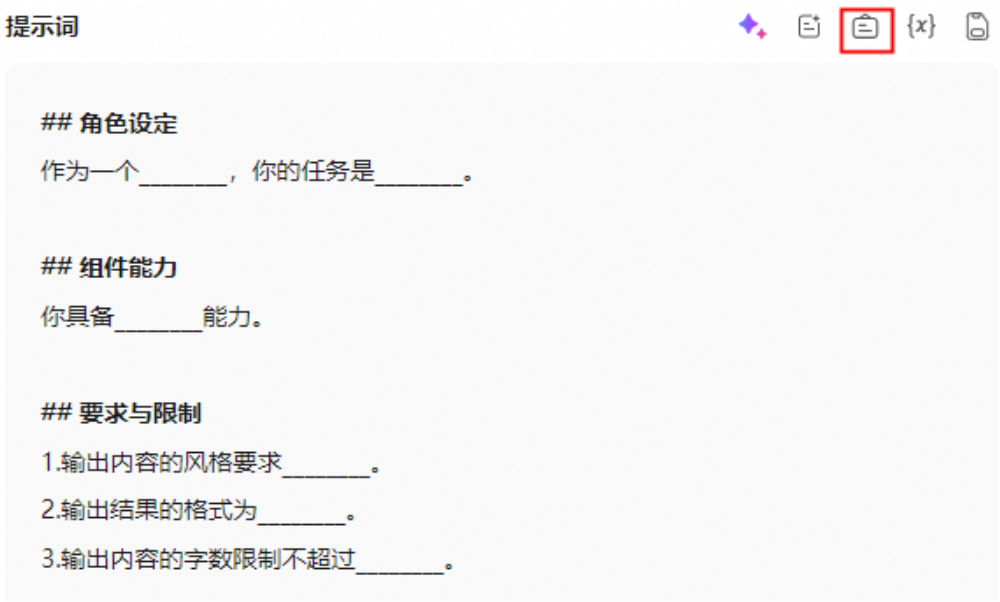


使用角色指令模板

平台预置了角色指令模板，支持用户根据实际业务场景给大模型指定具体的角色，从而显著提升回复内容与业务需求的契合度。

单击“角色指令模板”后，平台会自动填充角色指令的提示词模板，用户可以参考该模板来编写提示词。

图 6-143 填写模板



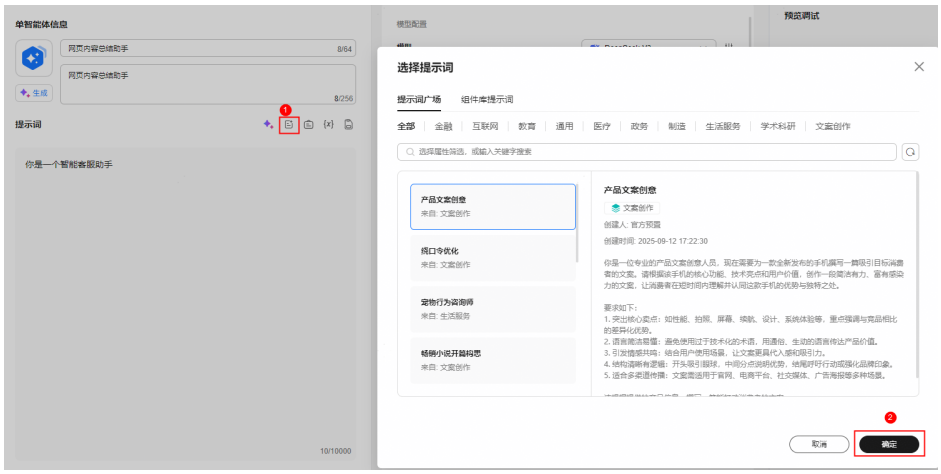
 版权所有 © 华为云计算技术有限公司

引用模板

AgentArts根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

单击“引用模板”后，在选择提示词的对话框中，用户可以在“提示词广场”或“组件库提示词”两个页签中选择适合当前业务场景的提示词模板。

图 6-144 选择提示词



说明

- 引用组件库提示词前，须确保组件库中已创建提示词，具体步骤请参考**创建提示词**。
- 预置提示词的来源是资产中心，在引用提示词之前，可以在资产中心中查看这些预置的提示词。

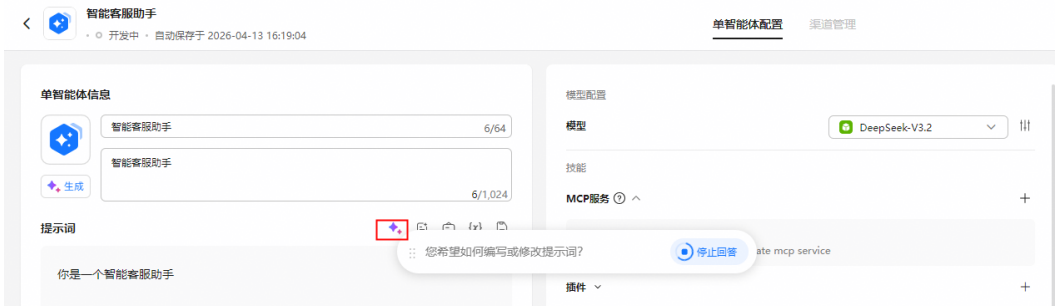
AI 生成提示词

用户可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

在“提示词”面板的编辑框中，输入希望编写的提示词内容，例如“你是一个智能客服助手”，然后单击“提示词智能优化”，AI将根据输入内容自动生成优化后的提示词。确认无误后，单击“应用”，即可将生成的提示词填入编辑框中。

如果智能生成的提示词不符合预期，可调整输入描述后重试。

图 6-145 AI 生成提示词



版权所有 © 华为云计算技术有限公司

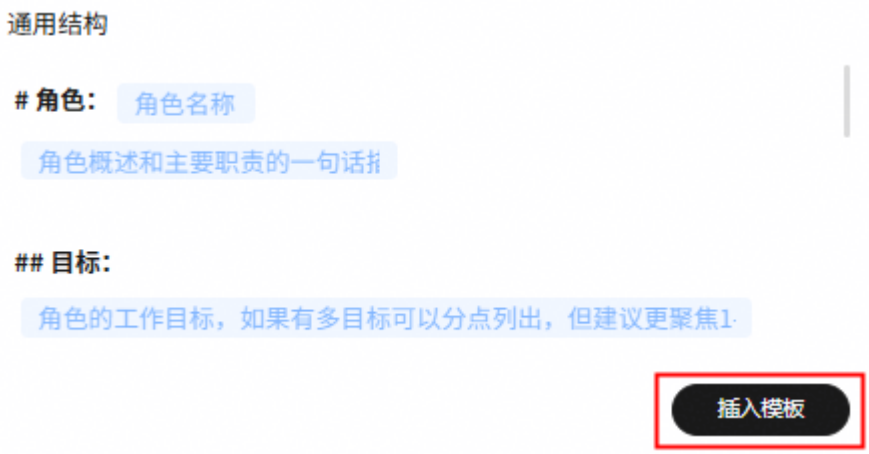
引用提示词资源

AgentArts提供丰富的预置提示词资源，以卡片形式展示，用户可根据需求轻松选择和使用。

当单击“插入模板”后，系统会自动将选定的提示词填充到编辑框中，您可以根据自己的业务场景进行修改。在修改提示词时，应重点关注其中的高亮部分。

可根据高亮文本部分添加文本内容；如果智能体引用了插件和MCP，需要在提示词中修改角色设定与回复逻辑，明确指示智能体调用“插件”和“MCP”处理用户问题。

图 6-146 引用提示词资源



更多操作

表 6-69 相关操作

操作	说明
在提示词中添加变量	当用户为应用添加记忆并创建了变量后，可以在提示词中选择已创建的变量，便于快速定义用户的某一行为或偏好。 支持单击“引用变量” {x} 图标，快速在提示词输入框中输入变量。 如果单智能体在技能配置中已添加插件和MCP，在提示词编辑时，支持**引用变量**以及已添加的**插件**和**MCP**。
保存为提示词模板	如果编写的提示词可以适用于多个场景，可以将其保存到组件库的提示词中，以便复用或供他人参考。 单击“ 保存到模板”，在“保存到我的提示词”页面，输入“模板名称”、选择“行业”和“标签”，并单击“确定”，将提示词创建成模板。 在保存模板时，**模板名称**和**行业**为必填参数，**标签**为可选参数。其中，**模板名称**支持1-64个字符。详细信息请参考创建提示词。

版权所有 © 华为云计算技术有限公司

相关文档

如何在单智能体中引用提示词的具体操作，请参考**配置提示词**。

6.5.6.2 在工作流中使用提示词

在实际业务开发中，单纯依靠大语言模型（LLM）的通用能力往往难以直接满足复杂的业务逻辑需求，容易出现输出不准确或偏离预期的情况。为了解决这一问题，开发者需要在工作流的编排过程中，针对各个关键节点的具体职能，配置高度定制化的提示词，将泛化的模型能力转化为确定性的业务执行能力。

如果您需要快速构建标准节点，推荐使用引用模板；如果需要定制化逻辑，推荐直接编写提示词。

以下以在大模型节点中使用提示词为例：

前提条件

已**创建工作流**。

直接编写提示词

可以根据业务需求直接编写提示词，提示词越具体明确，智能体的输出结果就越贴近预期效果。如何编写具体明确的提示词，请参考**撰写提示词规范**。

1. 登录平台后，在左侧导航栏选择 “开发中心 > 智能体管理”。
2. 在“工作流”页签，单击所需的工作流应用卡片或新创建一个工作流应用进入编排页面。
3. 可以在系统提示词或用户提示词的输入框中直接编写提示词。编写完成后，提示词将实时保存并应用于当前节点。

版权所有 © 华为云计算技术有限公司

图 6-147 编写提示词



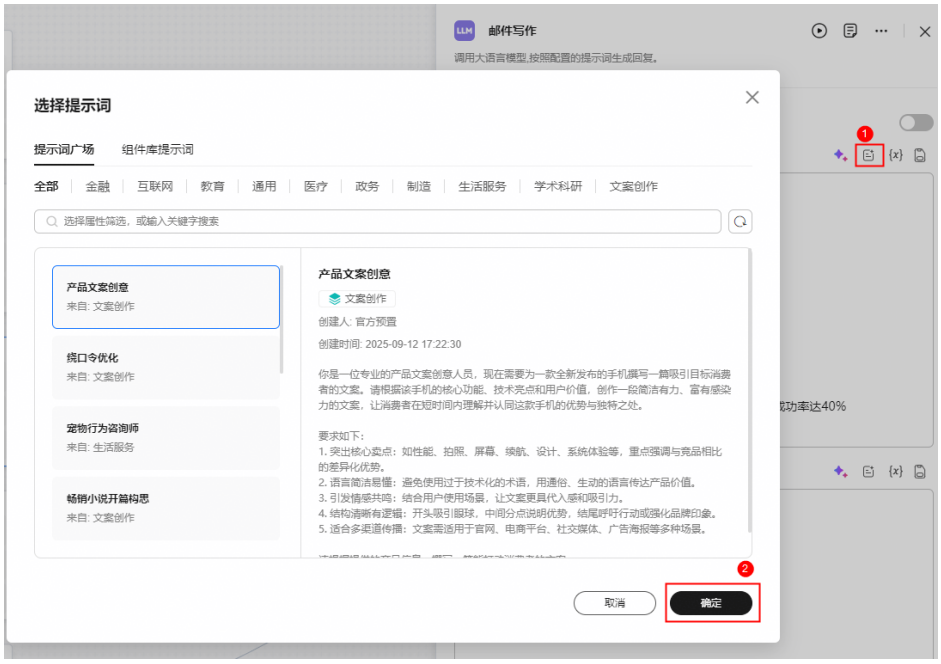
版权所有 © 华为云计算技术有限公司

引用模板

AgentArts根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。引用组件库提示词前，须确保已在组件库中创建提示词，详细信息请参考**创建提示词**。

单击“引用模板”后，在选择提示词的对话框中，用户可以在“提示词广场”或“组件库提示词”两个页签中选择适合当前业务场景的提示词模板。

图 6-148 选择提示词



说明

- 引用“提示词”前，须确保组件库中已创建提示词，具体步骤请参考**创建提示词**。
- 预置提示词的来源是资产中心，在引用提示词之前，可以在资产中心中查看这些预置的提示词。

AI 生成提示词

用户可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

在“提示词”面板的编辑框中，输入希望编写的提示词内容，例如“你是一个邮件写作助手”，然后单击“提示词智能优化”，AI将根据输入内容自动生成优化后的提示词。确认无误后，单击“应用”，即可将生成的提示词填入编辑框中。

如果智能生成的提示词不符合预期，可调整输入描述后重试。

版权所有 © 华为云计算技术有限公司

图 6-149 AI 生成提示词



更多操作

表 6-70 相关操作

操作	说明
在提示词中添加变量	在编写提示词时，单击“{x}引用变量”图标，支持输入变量以实现动态交互。
保存为提示词模板	如果编写的提示词可以适用于多个场景，可以将其保存到组件库的提示词中，以便复用或供他人参考。 单击“ 保存到模板”，在“保存到我的提示词”页面，输入“模板名称”、选择“行业”和“标签”，并单击“确定”，将提示词创建成模板。 在保存模板时，**模板名称**和**行业**为必填参数，**标签**为可选参数。其中，**模板名称**支持1-64个字符。详细信息请参考**创建提示词**。

版权所有 © 华为云计算技术有限公司

相关文档

在工作流中引用提示词的具体操作，请参考**大模型**、**意图识别**和**提问器**。

 版权所有 © 华为云计算技术有限公司

7 模型

7.1 模型介绍

模型服务为智能体提供核心推理能力，是智能体实现自主决策和任务执行的基础。在AgentArts中，您可以通过模型服务为智能体接入大语言模型，使其具备文本对话、图像理解、向量化等能力。

模型服务分类

为满足不同用户的技术能力、业务场景及需求，AgentArts提供了多样化的模型服务模式。以下从模型来源对各类模型服务进行介绍，具体如表7-1所示。

表 7-1 模型服务分类介绍

分类	说明	适用场景
平台预置模型	平台已预先对接好的模型服务，开箱即用。 平台预置模型通过对接**MaaS模型即服务**实现相关能力的集成，无需自行部署。当前已支持配置文本对话、文本排序、图像理解及文本向量化等类型。 开通AgentArts后，即可免费使用2,000,000Tokens进行调测、试运行。 免费额度用完或智能体发布后通过API接口调用，需要开通MaaS模型即服务及其模型，配置模型鉴权参数，即可便捷地调测和使用。具体操作请参考**接入预置的付费模型**。	希望快速上手，无需自行部署和维护模型，且对开箱即用有需求的开发者或企业。

版权所有 © 华为云计算技术有限公司

分类	说明	适用场景
自主接入模型	接入自行部署或第三方提供的模型服务API。 具体操作请参考**接入自定义模型**。	需要使用特定领域的专业模型，或企业内部已有自建、第三方模型服务，对数据合规与定制化有较高要求的场景。

费用说明

表 7-2 模型费用说明

接入方式	免费额度	超出免费额度后	查看用量
平台预置模型	开通AgentArts即赠送2,000,000Tokens，可用于模型调测、智能体试运行。	需开通MaaS模型即服务并配置模型鉴权参数，计费规则请参考**MaaS模型即服务价格**。	请参见**查看资源用量**。
自主接入模型	-	**AgentArts侧不计费**，费用以模型供应商的计费规则为准。	-

7.2 接入预置模型

7.2.1 免费模型与使用额度

AgentArts平台已对接MaaS模型即服务的模型，旨在为用户提供开箱即用的大模型核心能力支持。

免费额度说明

开通AgentArts后，平台默认提供2,000,000Tokens的免费使用额度，无需额外配置，可用于模型调测、智能体试运行。

如果**免费额度用完**或通过API接口调用智能体，您需要开通MaaS对应的模型服务并完成鉴权配置，此时将按照**MaaS模型即服务**实际调用量（Token）进行计费（详见**MaaS模型即服务价格**）。具体操作请参见**接入预置的付费模型**。

查看预置模型

进入AgentArts平台，依次执行以下操作查看平台预置的精选模型：

版权所有 © 华为云计算技术有限公司

1. 在左侧导航栏中，单击“资产广场”。
2. 选择“模型”页签，即可浏览平台预置的精选模型列表，如图7-1所示，系统将展示包含模型名称、上下文长度等信息的卡片列表。

图 7-1 精选模型



常见问题

- **免费额度用完了怎么办**

  免费额度用完后，需要配置华为云MaaS服务的模型API Key，具体操作请参见**接入华为云MaaS服务付费模型**。

- **开通后不使用会收费吗**

  免费额度仅用于模型调测、智能体试运行。不使用则不产生费用。**超出免费额度**或智能体部署后，只有实际调用才会产生费用，按MaaS计费规则执行。模型的调用量统计可在**MaaS服务控制台**查看。

图 7-2 查看 MaaS 调用统计



7.2.2 接入预置的付费模型

本章节适用于以下场景：平台提供的2,000,000Tokens免费额度已在模型调测、智能体试运行中用完，或通过API接口调用智能体。

注意

接入预置的付费模型后，将按照**MaaS模型即服务**实际调用量（Token）进行计费，计费相关信息详见**MaaS模型即服务价格**。

版权所有 © 华为云计算技术有限公司

接入预置的模型服务流程

AgentArts平台预置的模型服务基于**MaaS模型即服务**提供。当免费额度耗尽后，您需开通对应的模型服务并完成鉴权配置，方可继续调用。

接入预置模型服务的完整流程如**图7-3**所示。

图 7-3 接入预置的模型服务使用流程

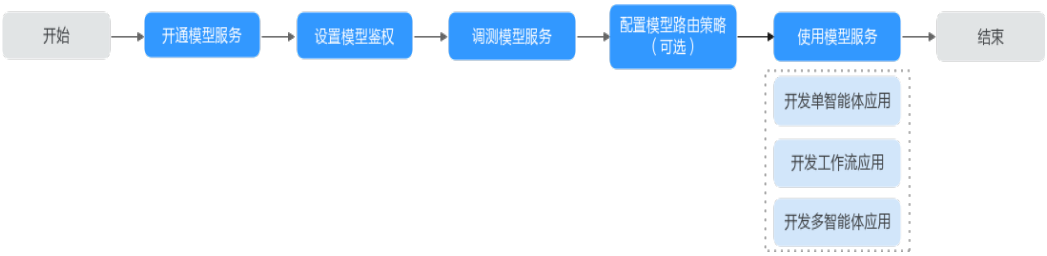


表 7-3 平台接入的预置模型服务使用流程详解

序号	流程环节	说明
1	**开通模型服务**	免费额度用尽后，需先在MaaS模型即服务中开通对应的模型服务，获取服务调用能力。
2	**设置模型鉴权**	模型服务开通后，需在AgentArts平台完成鉴权配置，以确保平台具备调用该模型服务的权限。
3	调测模型服务	对模型进行实际调用、参数调整及效果观测，验证其在特定场景下的功能表现与性能指标，确保模型在真实业务场景中稳定运行。具体操作请参考**调测模型**。
4	配置模型路由策略（可选）	当接入了多个同类型模型服务时，建议配置路由策略实现故障自动切换。路由策略创建后，建议对路由入口进行调测验证切换是否生效。具体操作请参考**配置模型路由策略**。
5	使用模型服务	模型鉴权设置完成后，可以在智能体、工作流中使用模型，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

开通模型服务

步骤1 登录**AgentArts智能体平台**。

步骤2 开通由MaaS模型即服务提供的模型服务。

方式一：在左侧导航栏，选择“资产广场”，进入“模型”页签，单击右上角的“批量开通模型”。

方式二：在左侧导航栏，选择“授权管理”，选择“模型”页签，单击“批量开通模型”。

步骤3 在弹框中，选择需要开通的模型，根据需求勾选“后续上新模型自动开通”，单击“确定”。

版权所有 © 华为云计算技术有限公司

开通成功后，在“资产广场 > 模型”页签中，已开通模型的状态显示为“已开通”。

----结束

设置模型鉴权

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“资产广场”，单击“模型”页签。

步骤3 在“精选模型”页面右上角，单击“模型鉴权管理”。

步骤4 在“模型鉴权管理”弹窗，如**图7-4**所示，请单击链接前往**MaaS模型即服务**控制台进行申请。

图 7-4 获取鉴权信息



1. 单击**图7-4**中的链接，前往MaaS模型即服务。
   由于AgentArts平台当前仅在“**西南-贵阳一**”区域提供服务，请务必将区域选择为“**西南-贵阳一**”，以确保鉴权配置正常生效。
2. 单击“创建API Key”，输入标签名称、描述信息，选择访问权限，如**图7-5**所示，单击“确定”。

版权所有 © 华为云计算技术有限公司

图 7-5 创建 API key



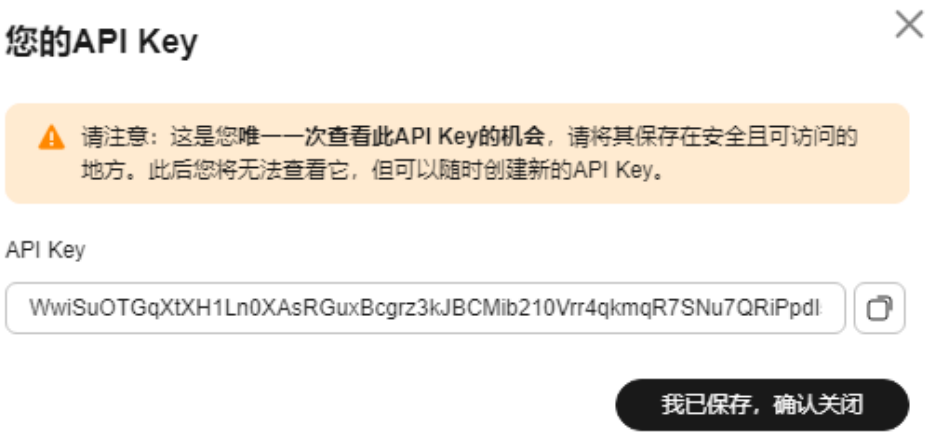
3. 在“您的API Key”弹窗页面，单击 ，复制API Key。

注意

API Key在创建后**仅显示一次**，请务必及时复制并妥善保存。如未保存，需重新创建API Key。

版权所有 © 华为云计算技术有限公司

图 7-6 复制 API Key



4. 单击“我已保存，确认关闭”。

步骤5 在输入框中输入获取的鉴权信息，单击“确定”。

步骤6 设置鉴权后，请参考**调测模型**，选择已鉴权的模型发起测试调用。

在调测界面输入一段常规的测试问题（例如：“你好，请简单介绍一下你自己”），输入完成后单击。

如果模型能够正常返回预期的回答（没有出现报错提示、鉴权失败或连接超时等异常），即表示该模型已成功接入并可以正常调用。

----结束

相关操作

模型相关操作请参考**表7-4**。

表 7-4 模型相关操作

操作	说明
移除鉴权配置	**警告** 移除鉴权配置后，模型不支持调测、使用。 如果平台提供的2,000,000Tokens免费额度尚未耗尽，即使移除鉴权配置，您仍可继续进行模型的调测与智能体试运行。 对于不再使用的模型，建议移除鉴权配置。 在“资产广场”的“模型”页签，单击界面右上角的“模型鉴权管理”，在“模型鉴权管理”弹窗中，单击“移除”。

常见问题

- **API Key忘记了怎么办**?

  API Key仅在创建时显示一次。如果丢失，需要在MaaS控制台重新创建一个API Key，然后在AgentArts中更新鉴权配置。

版权所有 © 华为云计算技术有限公司

- **免费额度用完后，不配置鉴权会怎样?**
  模型将无法调测和使用。必须完成鉴权配置后才能继续调用。

相关文档

接入华为云MaaS服务付费模型

7.3 接入自定义模型

7.3.1 自定义模型接入流程

概述

AgentArts支持接入由用户或第三方部署在外部环境的模型服务API。相比使用平台预置模型，接入自定义模型服务具有以下优势：

- 优化个性化体验：根据业务特点接入定制化模型，减少用户的搜索和选择时间，提供更加流畅高效的使用体验。例如，搜索引擎可根据用户的搜索历史和偏好，提供更精准的搜索结果。
- 增强特定领域的准确性：接入专业领域的模型服务，可显著提升特定场景下的输出准确性。例如，在医疗领域接入专业医学模型，可提供更准确的诊断建议。
- 提升开发效率：开发人员无需从零构建复杂模型，可直接接入已有的高质量模型服务API，快速集成大模型能力。

接入流程

AgentArts支持接入由用户或第三方部署在外部环境的模型服务API。接入前需先添加模型供应商，再接入其提供的模型服务。接入自定义模型服务的完整流程如表7-5所示。

图 7-7 自定义模型服务接入流程

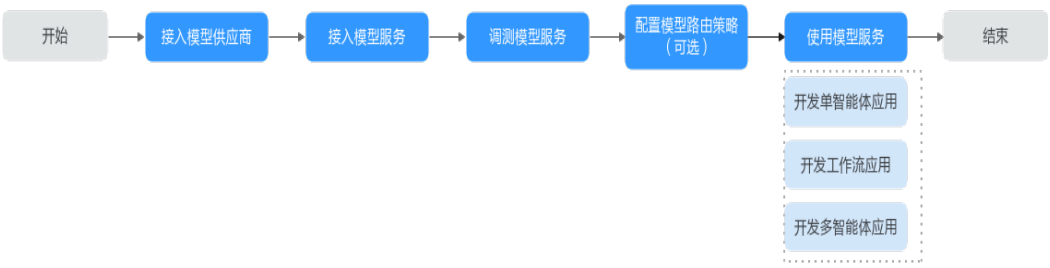


表 7-5 自定义模型服务接入流程详解

序号	流程环节	说明
1	接入模型供应商	接入模型服务之前，需先将模型供应商注册到AgentArts平台。具体操作请参考接入模型供应商。

版权所有 © 华为云计算技术有限公司

序号	流程环节	说明
2	接入模型服务	模型供应商接入后，将其提供的模型服务API接入AgentArts平台。具体操作请参考**接入模型服务**。
3	调测模型	对模型进行实际调用、参数调整及效果观测，验证其在特定场景下的功能表现与性能指标，确保模型在真实业务场景中稳定运行。具体操作请参考**调测模型**。
4	配置模型路由策略（可选）	当接入了多个同类型模型服务时，建议配置路由策略实现故障自动切换。路由策略创建后，建议对路由入口进行调测，验证切换是否生效。具体操作请参考**配置模型路由策略**。
5	使用模型服务	模型服务接入后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

7.3.2 接入模型供应商

模型供应商提供的模型服务使企业和个人能够快速获取和使用高质量模型。

在AgentArts中接入外部模型服务API前，需先接入模型供应商，用于对模型来源进行统一管理与标识。完成供应商接入后，您可在该供应商下继续接入具体的模型服务。

前提条件

已**开通AgentArts服务**。

新建模型供应商

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 开发配置”。

步骤3 选择“自定义模型”页签，单击“新建模型供应商”。

步骤4 在“新建模型供应商”页面，配置参数信息，如**图7-8**所示，具体参数说明请参考**表7-6**。

版权所有 © 华为云计算技术有限公司

图 7-8 新建模型供应商



表 7-6 新建模型供应商参数说明

参数	说明	示例
供应商图标	系统默认供应商图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 在虚线框中，单击鼠标左键，上传已准备好的供应商图标。 支持jpg、png格式图片，且不大于100KB。	系统默认图标

版权所有 © 华为云计算技术有限公司

参数	说明	示例
供应商名称	供应商的名称。由2~64个字符组成，包含中英文、数字、下划线、中划线、空格、竖线。	深度求索
供应商英文名称	供应商的英文名称。由2~64个字符组成，包含英文、数字、下划线、中划线、空格、竖线。	DeepSeek
描述	非必填。供应商的描述信息。由0~1000个字符组成。	深度求索（DeepSeek），是量化巨头幻方探索AGI（通用人工智能）的新组织，成立于2023年，专注于研究世界领先的通用人工智能底层模型与技术，挑战人工智能前沿性难题。
选择认证方式	选择在智能体、工作流或通过API调用该模型服务时的认证鉴权方式。 支持以下认证方式，详情请参考**表7-7**： • **无鉴权**：模型服务无需认证即可调用。 • **Api-key**：通过API Key进行认证。 • **AK/SK**：适用于盘古大模型的AK/SK认证方式。 • **App-code**：通过APP认证方式调用。 • **自定义ApiKey**：通过自定义请求头认证参数进行认证。 • **IAM鉴权**：华为云IAM认证，通过IAM账号获取用户Token进行认证。 • **HMAC**：通过AK（Access Key ID）/SK（Secret Access Key）加密调用请求。 **关于密钥加密：**鼠标移动至“选择认证方式”右侧的 ⓘ 上，可查看密钥的加密方式。平台默认使用**KMS加密 + 默认密钥**方式对认证密钥进行加密存储，防止信息泄露。 • 加密机制：使用用户选择的密钥派生DEK进行加密，默认密钥为kms-agentarts/default。 • 费用说明：实例不收费，每月提供20,000次免费调用额度。解密时支持缓存，但敏感数据较多时可能超出免费次数。如需更换加密方式，具体请参考**更换密钥的加密方式**。超出后按实际调用次数计费，详见**密码安全中心计费说明**。	Api-key

版权所有 © 华为云计算技术有限公司

表 7-7 认证方式说明

认证方式	参数说明
Api-key	通过请求Header的Authorization字段携带Bearer <Api-key>进行认证。 Api-key：输入待接入的供应商模型服务的API Key。输入后关键信息将加密保存，设置后约2分钟生效。
AK/SK	适用于盘古大模型，通过AK/SK加密调用请求。 ● AK（Access Key ID）：输入待接入的供应商模型服务的AK。输入后关键信息将加密保存。 ● SK（Secret Access Key）：输入待接入的供应商模型服务的SK。输入后关键信息将加密保存。
App-code	通过请求Header的“X-Apig-Appcode”字段携带App-code进行认证。 App-code：输入待接入的供应商模型服务的App-code。输入后关键信息将加密保存，设置后约2分钟生效。
自定义ApiKey	通过自定义请求头参数进行认证。 1. 在“Header配置”右侧，单击“+”新增Header参数。 2. 填写Header参数信息： – 参数名称：输入调用模型服务的Header参数名称。 – 参数值：输入调用模型服务的Header参数值。

版权所有 © 华为云计算技术有限公司

认证方式	参数说明
IAM鉴权	通过华为云IAM账号获取用户Token进行认证。 **鉴权配置：** • IAM认证url：获取IAM用户Token的接口地址。例如：https://{iam_host}/v3/auth/tokens，其中{iam_host}表示IAM Endpoint地址，即承载REST服务端点的服务器域名或IP。由于不同服务在不同区域的Endpoint各不相同，详情请参见**终端节点**。 例如： 您需要获取“西南-贵阳一”区域的Token，则需使用“西南-贵阳一”区域的Endpoint（iam.cn-southwest-2.myhuaweicloud.com），并结合**获取IAM用户Token（使用密码）**API的资源路径（/v3/auth/tokens），拼接后的完整请求地址如下。 https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens • 账号名：IAM用户所属的账号名。 • 项目：服务所属的区域信息。例如：cn-southwest-2 **验证方式（二选一）：** 方式一：IAM用户名/密码 • IAM用户名：IAM用户的名称。 • IAM密码：IAM用户的登录密码。 方式二：Access Key ID / Secret Access Key • Access Key ID：访问密钥ID。 • Secret Access Key：与访问密钥ID结合使用的密钥。
HMAC	通过AK/SK加密签名调用请求。 • AK（Access Key ID）：输入待接入的供应商模型服务的AK。输入后关键信息将加密保存。 • SK（Secret Access Key）：输入待接入的供应商模型服务的SK。输入后关键信息将加密保存。

步骤5 单击“确定”完成模型供应商接入。

新建的模型供应商将显示在“自定义模型”供应商卡片列表中，默认状态为“已配置鉴权”。处于“已配置鉴权”状态的模型供应商下的模型服务支持调测和使用。

步骤6 完成模型供应商接入后，请参见**接入模型服务**接入具体的模型API，并进行调测与发布。

----结束

相关操作

在模型供应商卡片列表，支持的其他操作请参考**表7-8**。

版权所有 © 华为云计算技术有限公司

表 7-8 模型供应商信息相关操作

操作	说明
查看模型供应商信息	单击模型供应商卡片，进入供应商详情页，查看模型供应商基本信息。
移除鉴权配置	**警告** 移除鉴权配置后，该供应商下的所有模型将**无法**进行调测和使用，可能导致依赖这些模型的智能体服务中断。 在新增模型供应商时，“鉴权方式”选择“Api-key”、“AK/SK”、“App-code”、“自定义ApiKey”、“IAM鉴权”、“HMAC”时，才支持移除鉴权配置。 鼠标移动至需要移除鉴权配置的模型供应商卡片上，单击“鉴权配置”，单击“移除”。 移除鉴权配置后，模型服务为“未配置鉴权”，**未配置鉴权**的模型，不支持调测、使用。
修改模型供应商信息	鼠标移动至需要修改的模型供应商卡片上，单击“编辑”。
删除模型供应商	**警告** 删除模型供应商将永久清除该供应商的所有配置信息，且不可恢复。如果供应商下存在已发布的模型服务，需先删除相关模型服务后方可执行。 鼠标移动至需要删除的模型供应商卡片上，单击“删除”。

相关文档

接入和使用DeepSeek模型

手动接入和使用OpenAI协议模型

7.3.3 接入模型服务

本章节介绍如何在已接入的模型供应商下接入模型服务API。

为满足用户对模型的个性化与专业化需求，AgentArts支持接入由用户自建或第三方在外部环境部署的模型服务API。模型服务接入后，您可以在AgentArts中进行调测和发布，并在智能体中选择使用。

支持接入的模型类型

当前支持接入的模型类型如下：

- 文本对话（Chat）
- 文本向量化（Embeddings）
- 文本排序（Rerank）
- 图像理解

典型应用场景

- 接入企业自研的领域大模型，用于行业知识问答。

版权所有 © 华为云计算技术有限公司

- 接入第三方向量模型，构建私有知识库检索能力。
- 接入图像理解模型，实现多模态智能体应用。

选型建议

表 7-9 选型建议

模型类型	典型应用场景
文本对话	智能问答、对话机器人、文本生成
文本向量化	知识库检索、语义搜索、相似度计算
文本排序	搜索结果重排、推荐系统优化
图像理解	图像问答、多模态对话、图像分析

接口协议要求

为保证接入模型服务的质量，**模型API接入前请确保其符合对应的接口协议规范**。不同模型类型支持的接口协议如下：

- 文本对话、文本向量化、图像理解：标准OpenAI协议、阿里千问接口协议、MaaS标准API V1、MaaS标准API V2
- 文本排序：AI引擎标准协议

  说明

  - 标准OpenAI协议和AI引擎标准协议规范请参考**模型API接口规范**。
  - 阿里千问接口协议规范请参考通义千问的接口协议。
  - MaaS标准API V1接口规范请参考**MaaS标准API V1**。
  - MaaS标准API V2接口规范请参考**MaaS标准API V2**。

选型建议：优先选择与待接入模型原生匹配的协议。如果模型兼容OpenAI协议，推荐选择“标准OpenAI协议”以获得最佳兼容性。

费用说明

- 调用接入的第三方模型服务，费用由对应模型厂商收取，请参考其官网计费说明。
- AgentArts平台不对接入模型服务额外收费。
- 建议通过流控配置限制调用频率，避免意外产生高额费用。

前提条件

- 已**接入模型供应商**。
- 已获取待接入模型服务的必要信息，例如API地址、鉴权信息（API Key等）、模型名称（模型ID/模型编码）及协议类型。

版权所有 © 华为云计算技术有限公司

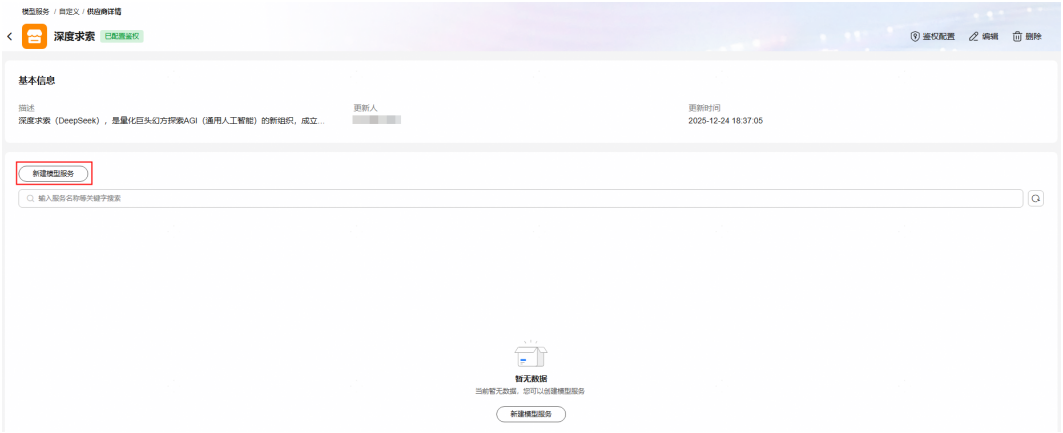
新建模型服务

步骤1 登录AgentArts智能体平台。

步骤2 在左侧导航栏，选择“开发中心 > 开发配置”。

步骤3 选择“自定义模型”页签，单击对应的模型供应商卡片，在“供应商详情”页面，单击“新建模型服务”。

图 7-9 新建模型服务



步骤4 在“新建模型服务”页面，配置参数信息，具体参数说明请参考表7-10。

表 7-10 新建模型服务参数说明

参数	说明	示例
模型服务图标	系统默认模型服务图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 在虚线框中，单击鼠标左键，上传已准备好的模型服务图标。 支持jpg、png格式图片，且不大于100KB。	系统默认图标
模型服务名称	自定义模型服务名称。由2~64个字符组成，包含中英文、数字及:._/\-，以中英文、数字开头结尾。	文本对话
模型名称	填写的模型名称必须为该模型的模型ID/模型编码，否则会导致模型不可用。 需要登录第三方模型厂商官网查看，例如，Baichuan4、deepseek-chat、glm-4-air。 如果要接入自建的模型服务，该模型名称将用于接口调用请求体的model字段。 由2~64个字符组成，包含中英文、数字及:._/\-，以中英文、数字开头结尾。	deepseek-chat

版权所有 © 华为云计算技术有限公司

参数	说明	示例
类型	选择模型类型。 • **文本对话**：文本对话模型，通常被称为对话式AI或聊天机器人，是一种经过训练能够理解和生成人类语言，并以多轮、上下文连贯的方式进行交流的人工智能系统。 • **文本向量化**：文本向量化模型的核心任务是将文本（词、句、段落或文档）转换为计算机能够理解和处理的数值形式——即高维向量（也称为“嵌入”，Embedding）。这个向量就像是文本在数学空间中的一个“坐标点”。 • **文本排序**：文本排序模型用于对一组文本对象进行相关度排序。给定一个查询（Query）和一个文本列表（如搜索引擎的结果），排序模型会根据每个文本与查询的相关程度，从高到低进行排序。 • **图像理解**：图像理解模型是一种能够对图像内容进行分析、解读和理解的人工智能模型，其核心目标是让计算机像人类一样“看懂”图像。	文本对话
模型服务API地址	填入需要接入模型的API地址信息。字符长度不大于255个字符。 格式为：https://*xxx*.com/v1/*xxx*。	https://api.deepseek.com/chat/completions
API接口协议	• 当“类型”值为“文本对话”、“文本向量化”、“图像理解”时，选择“标准OpenAI协议”、“阿里千问接口协议”、“MaaS标准API V1”、“MaaS标准API V2”。 • 当“类型”值为“文本排序”时，选择“AI引擎标准协议”。 **标准OpenAI协议**和**AI引擎标准协议**规范请参考**模型API接口规范**。 **阿里千问接口协议**规范请参考通义千问的接口协议。 **MaaS标准API V1**接口规范请参考**MaaS标准API V1**。 **MaaS标准API V2**接口规范请参考**MaaS标准API V2**。	标准OpenAI协议
流控配置	超出流控值，则触发限流，用户的请求会因为流控而失败。建议根据模型API的QPS限制和业务并发需求选择，超出限制的请求将返回限流错误。 • 无限制 • 10次/秒 • 50次/秒 • 100次/秒 • 200次/秒	无限制

版权所有 © 华为云计算技术有限公司

参数	说明	示例
选择标签	可选项。 当“模型类型”值为“文本对话”、“图像理解”时，才有此参数。 选择标签后，在应用中选择大模型时，显示在大模型右侧。 • **工具**：当该大模型支持应用调用外部工具（例如，MCP服务、插件、知识库）时，可以选择该标签。 • **思考**：该大模型具备思维推理能力时，可以选择该标签。 • **联网**：该大模型具备联网搜索能力时，可以选择该标签。	工具
是否支持关闭思维链输出	当“选择标签”选择了**思考**时，才有此参数。 • 开启：模型在调测、使用时，显示“深度思考”参数。 在模型的调测和使用过程中，“深度思考”开关的生效情况如下： – 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 – 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 – 如果模型不支持思维链输出能力，则开启、关闭均不生效。 • 关闭：模型在调测、使用时，不显示“深度思考”参数。默认**关闭**。 模型在调测、使用时，是否输出思维链，取决于模型本身是否支持思维链输出。	关闭
自定义标签	选填项。 最多支持添加5个标签。单击  ，输入标签内容，按Enter键。 添加后，在应用中选择大模型时，显示在大模型右侧。	-
模型服务描述	选填项。 模型服务的描述信息。由0~1000个字符组成。	-

步骤5 单击“确定”。

新建的模型服务，显示在模型供应商下的模型服务卡片列表中，默认状态为“未发布”。

版权所有 © 华为云计算技术有限公司

步骤6 在需要调测的模型服务卡片上，单击“ ··· > 模型调测”，具体调测操作请参考**调测模型**。

步骤7 在需要发布的模型服务卡片上，单击“ ··· > 发布”。

发布后，模型服务状态为“已发布”。

模型服务卡片展示为“已发布”，才支持在智能体中使用。

步骤8 （可选）当已接入并发布多个同类型模型服务时，可配置模型路由策略实现模型故障自动切换，提升可用性。具体操作请参考**配置模型路由策略**。

----结束

相关操作

在接入模型服务卡片列表，支持的其他操作请参考**表7-11**。

表 7-11 接入模型服务相关操作

操作	说明
查看模型服务信息	单击模型服务卡片，进入模型服务详情页，可以查看模型服务信息。
修改模型服务信息	**注意** 未发布的模型服务，才可以修改。 在需要修改的模型服务卡片上，单击“ ··· > 编辑”。
取消发布	**注意** 已发布的模型服务，才可以取消发布。 **未发布**的模型服务，不支持使用。 在需要取消发布的模型服务卡片上，单击“ ··· > 取消发布”。
删除模型服务	**警告** • 删除模型服务后，使用该模型的智能体将无法正常工作。 • 删除操作不可恢复，请确认无智能体正在使用该模型后再执行。 • 已发布的模型服务，需先取消发布才能删除。 在需要删除的模型服务卡片上，单击“ ··· > 删除”。

相关文档

- 模型服务接入后，可以调测模型服务，具体操作请参考**调测模型**。
- 已接入并发布多个同类型模型服务时，可配置模型路由策略实现模型故障自动切换，提升可用性。具体操作请参考**配置模型路由策略**。
- 模型服务接入后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

版权所有 © 华为云计算技术有限公司

常见问题

- **接入的外部模型无法使用?**

  请依次检查以下事项：

  - API Key是否正确粘贴，无多余空格或换行符。
  - DeepSeek账户是否有可用余额。
  - 模型服务API地址是否填写正确（https://api.deepseek.com/chat/completions）。
  - “模型名称”参数是否填写了正确的模型编码（如deepseek-chat），而非自定义名称。

- **API地址应该填基础路径还是完整路径?**

  需要填入模型服务商文档中给出的完整API地址，例如：

  表 7-12 模型 API 地址示例

模型服务商	API地址示例
DeepSeek	https://api.deepseek.com/chat/completions
智谱	https://open.bigmodel.cn/api/paas/v4/chat/completions

  不同模型服务商的路径结构不同（如智谱是/api/paas/v4/chat/completions，DeepSeek是/chat/completions），请以服务商官方文档为准，填写完整地址。

- **接入自建模型需要注意什么?**

  “模型名称”字段必须与模型服务商定义的模型名称完全一致，该值会直接作为接口请求体中model字段的值。随意填写会导致模型不可用。

  模型名称需要到模型服务商官网查看，注意大小写和分隔符必须完全匹配。

- **可以接入本地部署的模型吗?**

  可以，但需要确保：

  - 模型API对AgentArts平台可达（公网可访问）。
  - API符合AgentArts平台的模型对接协议。

7.3.4 模型 API 接口规范

本章节介绍**标准OpenAI协议**和**AI引擎标准协议**的接口规范。接入自定义模型服务前，请确保模型API符合对应的接口协议规范。

当前模型网关支持接入以下四种类型的模型API：

- 文本对话（Chat）
- 文本向量化（Embeddings）
- 文本排序（Rerank）
- 图像理解

各模型类型支持的接口协议如下表所示。

 版权所有 © 华为云计算技术有限公司

表 7-13 模型类型与接口协议对照

模型类型	支持的接口协议
文本对话	标准OpenAI协议、阿里千问接口协议、**MaaS标准API V1**、**MaaS标准API V2**
文本向量化	标准OpenAI协议、阿里千问接口协议、**MaaS标准API V1**、**MaaS标准API V2**
图像理解	标准OpenAI协议、阿里千问接口协议、**MaaS标准API V1**、**MaaS标准API V2**
文本排序	AI引擎标准协议

适用场景

本章节适用于以下场景：

- 需要将第三方模型服务（如DeepSeek、Moonshot等）接入AgentArts，用于确认模型API是否满足平台接口协议规范。
- 自行部署了模型推理服务，需要按平台支持的协议实现/适配API接口后再接入AgentArts。
- 模型接入或调用过程中出现异常（例如请求失败、响应字段缺失、格式不兼容等），需要对照接口规范排查请求与响应格式问题。

费用说明

调用模型API会产生相应费用，计费维度为Token数量（包括输入Token和输出Token）。具体计费规则请参考接入的模型的供应商官网。

建议在测试阶段设置合理的max_tokens参数以控制成本。

文本对话（Chat）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-14 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **表 7-15** objects	文本对话消息体。
model	是	String	文本对话使用的模型名称。

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
frequency_penalty	否	Number	频率惩罚，会根据文本中新Token的出现频率对其进行惩罚，从而降低模型重复相同内容的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● **默认值**（0.0）：不施加任何频率惩罚。模型按原本的概率分布生成文本。 ● **正值**（例如 0.5，1.0，2.0）：增加惩罚力度。值越高，模型越不愿意使用已经用过的词。使输出文本的词汇更多样化、更富有创造性，但过高的值可能导致用词生僻、语句不通顺甚至偏离主题。 ● **负值**（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于使用已经用过的词。使输出文本的词汇更集中、更稳定、更可能重复关键主题词。但过低的值会导致用词极其重复、啰嗦。 例如： 提示词为（Prompt）：“写一首关于猫的诗。” ● frequency_penalty=0（默认）：输出可能正常地重复使用“猫”、“尾巴”、“柔软”等合理词汇。 ● frequency_penalty=1.5（高惩罚）：模型会极力避免重复用词。第一句用了“猫”，第二句可能会用“毛茸伙伴”、“喵星人”、“优雅的生物”等同义词来替代，词汇非常丰富。但如果惩罚过高，可能会为了规避重复而选用不合适的词，导致诗歌变得奇怪。 ● frequency_penalty=-1.0（负惩罚）：模型不害怕重复，甚至鼓励重复。输出可能会变成：“猫，猫，可爱的

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
			猫。猫在跑，猫在跳，猫的尾巴摇啊摇。” 显得非常冗余和缺乏创意。
logit_bias	否	Map<String,Integer>	该参数接受一个JSON对象，将标记映射到从-100（禁止生成）到100（强制选择该标记）的关联偏差值。 像-1和1这样的适度值将以较小的程度改变选择标记的概率。 使用logit_bias参数时，偏差被添加到模型生成的logits之前进行抽样。
max_tokens	否	Integer	返回体允许的最大token数。
n	否	Integer	返回体中包含的choices数量，建议默认设置为1，最大限度地降低成本。 • **最小值**：1 • **最大值**：128 • **默认值**：1

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
presence_penalty	否	Number	存在惩罚，会根据文本中新Token是否出现对其进行惩罚，核心作用是降低模型再次讨论已经出现过的“话题”的可能性，从而增加模型谈论新主题的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● **默认值**（0.0）：不施加任何存在惩罚。 ● **正值**（例如 0.5，1.0，2.0）：增加惩罚力度，值越高，模型越不愿意停留在已经提及的主题上，越倾向于引入全新的想法、概念或话题。使对话或文本更容易“跑题”或转向新方向。在创意生成中，这可以带来更大的探索性。 ● **负值**（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于围绕已经出现的主题进行深入讨论，避免引入新信息，使输出内容更加集中、紧扣主题，但可能显得缺乏发散性。

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
stream	否	Boolean	布尔类型。 • 设为**true**时，返回结果为**流式**。 流式处理是一种边接收边处理、实时性强的数据处理模式。它将数据视为连续不断的“流”，允许低延迟和即时响应，广泛应用于视频播放、实时监控、大数据分析和人工智能生成内容等领域。 • 设为**false**时，返回结果为**非流式**，JSON格式结构化数据。 非流式指数据或操作不是连续、实时地传输或处理，而是一次性接收完整的输入，待完全处理后一次性返回完整的结果。这种模式常见于传统HTTP请求-响应模式、大模型API的非流式输出和语音合成等领域，强调数据的完整性和整体性，不追求即时反馈。 **默认值**：false
temperature	否	Number	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。 • **最小值**：0 • **最大值**：2 • **默认值**：1
top_p	否	Number	影响输出文本的多样性，取值越大，生成文本的多样性越强。 • **最小值**：0.0 • **最大值**：1.0 • **默认值**：1
tools	否	Array of **表 7-16** objects	可供模型调用的工具。
tool_choice	否	String	用于控制模型是如何选择要调用的函数，仅当工具类型为function时补充。 默认为auto，且当前仅支持auto。

版权所有 © 华为云计算技术有限公司

表 7-15 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
role	是	String	消息体对应的角色。 • system：表示系统提示词角色。 • user：表示用户输入角色。 • assistant：表示模型回复角色。
content	是	String	消息具体内容。
name	否	String	对话参与者的可选名称，提供给模型信息以区分相同角色的不同对话参与者。

表 7-16 FunctionCallTool

参数	是否必选	参数类型	描述
type	否	String	调用工具类型，目前仅支持function。
function	否	表7-17 object	仅当工具类型为function时补充。

表 7-17 function

参数	是否必选	参数类型	描述
name	否	String	函数名称，只能包含a-z、A-Z、0-9、下划线和中划线。最大长度为64个字符。
description	否	String	用于描述函数功能。 模型会根据这段描述决定函数调用方式。
parameters	否	Object	Json Schema对象，用于定义函数所接受的参数。

- **工具调用请求示例**
  - 流式请求示例

```
{
   "model": "my-chat-model",
   "messages": [
      {
         "role": "user",
         "content": "请帮我查询南京的天气"
```


版权所有 © 华为云计算技术有限公司

```
        }
    ],
    "tools": [
        {
            "type": "function",
            "function": {
                "name": "get_weather",
                "description": "获取给定地点的天气",
                "parameters": {
                    "type": "object",
                    "properties": {
                        "location": {
                            "type": "string",
                            "description": "地点，例如南京。"
                        }
                    },
                    "required": ["location"]
                }
            }
        }
    ],
    "max_tokens": 200,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": true
}
```

- 非流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
        {
            "role": "user",
            "content": "请帮我查询南京的天气"
        }
    ],
    "tools": [
        {
            "type": "function",
            "function": {
                "name": "get_weather",
                "description": "获取给定地点的天气",
                "parameters": {
                    "type": "object",
                    "properties": {
                        "location": {
                            "type": "string",
                            "description": "地点，例如南京。"
                        }
                    },
                    "required": ["location"]
                }
            }
        }
    ],
    "max_tokens": 200,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": false
}
```

- **非工具调用请求示例**
  - 流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
```

 版权所有 © 华为云计算技术有限公司

```
        {
            "role": "system",
            "content": " You are a helpful assistant."
        },
        {
            "role": "user",
            "content": "你好！"
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": true
}
```

- 非流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant."
        },
        {
            "role": "user",
            "content": "你好！"
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": false
}
```

响应体参数

表 7-18 响应体参数

参数	参数类型	描述
id	String	文本对话唯一标识符。
choices	Array of 表7-19 objects	返回体列表。 如果“n”大于1，则结果为多个。
created	Integer	问答发生的时间。格式为时间戳。
model	String	文本对话使用的模型名称。
object	String	固定值“chat.completion”。
usage	表7-23 object	文本对话用量统计。

表 7-19 choices

参数	参数类型	描述
index	Integer	返回多个choices时，每个choice对应的顺序。

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
message	表7-20 object	模型服务返回的具体消息体内容。
finish_reason	String	返回结束的原因。 • stop：模型达到自然停止点或提供的停止序列。 • length：达到请求中指定的最大token数。 • content_filter：由于内容过滤器的标志而省略了内容。 • tool_calls：模型选择了某个工具。

表 7-20 ChatCompletionResponseMessage

参数	参数类型	描述
content	String	返回消息体的内容，与tool_calls二选一。
role	String	返回消息体的角色。 枚举值： • user：用户输入的问题。 • assistant：大模型的答复内容。
tool_calls	Array of 表7-21 objects	工具调用消息，与content二选一。

表 7-21 ToolCall

参数	参数类型	描述
id	String	工具调用唯一标识符。
type	String	工具类型，当前仅支持function。
function	表7-22 Object	调用函数的详细信息。

表 7-22 CallFunction

参数	参数类型	描述
name	String	函数名。
arguments	String	调用函数的参数，JSON格式。

版权所有 © 华为云计算技术有限公司

表 7-23 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	回答包含的token数。
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

- **工具调用响应示例**
  - 流式响应示例

    流式返回的工具调用信息必须在一条消息内，不能分开返回。

    ```
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"role":"assistant","content":null,"tool_calls":[{"id":"call_123","type":"function","function":{"name":"get_weather","arguments":"{\"location\":\"南京\"}"}}]}"logprobs":null,"finish_reason":null}]}
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"tool_calls"}]}
    ```

  - 非流式响应示例

    ```
    {
       "id": "chatcmpl-xxx",
       "object": "chat.completion",
       "created": 1718772336,
       "model": "my-chat-model",
       "choices": [
          {
             "index": 0,
             "message": {
                "role": "assistant",
                "content": null,
                "tool_calls": [
                   {
                      "id": "call_123",
                      "type": "function",
                      "function": {
                         "name": "get_weather",
                         "arguments": "{\"location\": \"南京\"}"
                      }
                   }
                ]
             },
             "finish_reason": "tool_calls",
             "logprobs": null
          }
       ],
       "usage": {
          "prompt_tokens": 5,
          "completion_tokens": 10,
          "total_tokens": 15
       }
    }
    ```

- **非工具调用响应示例**
  - 流式响应示例

    ```
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":
    {"role":"assistant","content":""},"logprobs":null,"finish_reason":null}]}
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你好"},"logprobs":null,"finish_reason":null}]}
    ```

 版权所有 © 华为云计算技术有限公司

{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"，"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"有"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"什么"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"我"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"可以"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"帮助"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"的"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"吗"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"？"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"stop"}]}

- 非流式响应示例

{
  "id": "chatcmpl-xxx",
  "object": "chat.completion",
  "created": 1718772336,
  "model": "my-chat-model",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "你好，有什么我可以帮助你的吗？"
      },
      "finish_reason": "stop",
      "logprobs": null
    }
  ],
  "usage": {
    "prompt_tokens": 5,
    "completion_tokens": 10,
    "total_tokens": 15
  }
}

文本向量化（Embeddings）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

 版权所有 © 华为云计算技术有限公司

表 7-24 请求体参数

参数	是否必选	参数类型	描述
input	是	Array of strings	• 纯文本（string），例如，"你好" 。 • 文本列表（array），例如，["你","好"] 。 数组长度：1-2048。
model	是	String	向量化模型名称。

请求示例：

```
{
   "model": "my-embedding-model",
   "input": "你好"
}
```

响应体参数

表 7-25 响应体参数

参数	参数类型	描述
data	Array of **表 7-26** objects	向量化结果。
model	String	向量化模型名称。
object	String	固定值“list”。
usage	**表7-27** object	每次请求的用量统计。

表 7-26 Embedding

参数	参数类型	描述
index	Integer	向量在向量列表中的排序。
embedding	Array of numbers	向量数组。Float类型。
object	String	固定值“embedding”。

表 7-27 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
total_tokens	Integer	提问+回答token总数。

响应示例：

```
{
   "data": [
      {
         "index": 0,
         "embedding": [
            0.02513289265334606,
            -0.017512470483779907,
            -0.029955564066767693,
            ...
         ],
         "object": "embedding"
      }
   ],
   "usage": {
      "prompt_tokens": 5,
      "total_tokens": 5
   },
   "model": "my-embedding-model",
   "object": "list"
}
```

文本排序（Rerank）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-28 请求体参数

参数	是否必选	参数类型	描述
query	是	String	原始请求问题，基于该问题对候选文本进行排序。
top_n	是	Integer	返回排序靠前的n个结果。
docs	是	Array of strings	候选文本，文件大小限制为512MB以内。
model	是	String	排序模型名称。

请求示例：

```
{
   "model": "my-rerank-model",
   "query": "请问AI原生应用引擎提供了什么能力？",
   "docs": ["AI原生应用引擎提供了应用开发、模型网关等能力。", "AI原生应用引擎正在逐步完善、提高竞争力。"],
```

 版权所有 © 华为云计算技术有限公司

```
    "top_n": 3
}
```

响应体参数

表 7-29 响应体参数

参数	参数类型	描述
model	String	排序模型名称。
usage	**表7-30** object	每次请求的用量统计。
results	Array of **表7-31** objects	排序结果。

表 7-30 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

表 7-31 RankDocument

参数	参数类型	描述
index	Integer	文本排序后对应的序号。
document	**表7-32** object	文本。
relevance_score	Number	文本的排序分数。

表 7-32 Document

参数	参数类型	描述
text	String	文本内容。

响应示例：

```
{
    "model": "my-rerank-model",
    "usage": {
        "prompt_tokens": 5,
        "total_tokens": 5
    },
"results": [
```

 版权所有 © 华为云计算技术有限公司

```
    {
        "index": 0,
        "document": {"text": "AI原生应用引擎提供了应用开发、模型网关等能力。"},
        "relevance_score": 0.9
    },
    {
        "index": 1,
        "document": {"text": "AI原生应用引擎正在逐步完善、提高竞争力。"},
        "relevance_score": 0.5
    }
  ]
}
```

图像理解 API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-33 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **表 7-34** objects	图像理解对话消息体。
model	是	String	图像理解对话使用的模型名称。

 版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
frequency_penalty	否	Number	频率惩罚，会根据文本中新Token的出现频率对其进行惩罚，从而降低模型重复相同内容的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● **默认值**（0.0）：不施加任何频率惩罚。模型按原本的概率分布生成文本。 ● **正值**（例如 0.5，1.0，2.0）：增加惩罚力度。值越高，模型越不愿意使用已经用过的词。使输出文本的词汇更多样化、更富有创造性，但过高的值可能导致用词生僻、语句不通顺甚至偏离主题。 ● **负值**（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于使用已经用过的词。使输出文本的词汇更集中、更稳定、更可能重复关键主题词。但过低的值会导致用词极其重复、啰嗦。 例如： 提示词为（Prompt）：“写一首关于猫的诗。” ● frequency_penalty=0（默认）：输出可能正常地重复使用“猫”、“尾巴”、“柔软”等合理词汇。 ● frequency_penalty=1.5（高惩罚）：模型会极力避免重复用词。第一句用了“猫”，第二句可能会用“毛茸伙伴”、“喵星人”、“优雅的生物”等同义词来替代，词汇非常丰富。但如果惩罚过高，可能会为了规避重复而选用不合适的词，导致诗歌变得奇怪。 ● frequency_penalty=-1.0（负惩罚）：模型不害怕重复，甚至鼓励重复。输出可能会变成：“猫，猫，可爱的

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
			猫。猫在跑，猫在跳，猫的尾巴摇啊摇。” 显得非常冗余和缺乏创意。
logprobs	否	Boolean	是否返回输出Token的对数概率。
top_logprobs	否	Integer	指定在每一步生成时，返回模型最大概率的候选Token个数。 取值范围：[0,5] 仅当**logprobs**为true时生效。
max_tokens	否	Integer	返回体允许的最大token数。
presence_penalty	否	Number	存在惩罚，会根据文本中新Token是否出现对其进行惩罚，核心作用是降低模型再次讨论已经出现过的“话题”的可能性，从而增加模型谈论新主题的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 • **默认值**（0.0）：不施加任何存在惩罚。 • **正值**（例如 0.5，1.0，2.0）：增加惩罚力度，值越高，模型越不愿意停留在已经提及的主题上，越倾向于引入全新的想法、概念或话题。使对话或文本更容易“跑题”或转向新方向。在创意生成中，这可以带来更大的探索性。 • **负值**（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于围绕已经出现的主题进行深入讨论，避免引入新信息，使输出内容更加集中、紧扣主题，但可能显得缺乏发散性。
n	否	Integer	生成响应的个数。取值范围是1-4。 对于需要生成多个响应的场景（如创意写作、广告文案等），可以设置较大的n值。 默认值为1。

 版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
stream	否	Boolean	布尔类型。 ● 设为**true**时，返回结果为**流式**。 流式处理是一种边接收边处理、实时性强的数据处理模式。它将数据视为连续不断的“流”，允许低延迟和即时响应，广泛应用于视频播放、实时监控、大数据分析和人工智能生成内容等领域。 ● 设为**false**时，返回结果为**非流式**，JSON格式结构化数据。 非流式指数据或操作不是连续、实时地传输或处理，而是一次性接收完整的输入，待完全处理后一次性返回完整的结果。这种模式常见于传统HTTP请求-响应模式、大模型API的非流式输出和语音合成等领域，强调数据的完整性和整体性，不追求即时反馈。 **默认值**：false
seed	否	Integer	设置seed参数会使文本生成过程更具有确定性，通常用于使模型每次运行的结果一致。 在每次模型调用时传入相同的seed值（由您指定），并保持其他参数不变，模型将尽可能返回相同的结果。 取值范围：0到2^{31}-1。
temperature	否	Number	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。 ● **最小值**：0 ● **最大值**：2 ● **默认值**：1
top_p	否	Number	影响输出文本的多样性，取值越大，生成文本的多样性越强。 ● **最小值**：0.0 ● **最大值**：1.0 ● **默认值**：1

版权所有 © 华为云计算技术有限公司

表 7-34 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
role	是	String	消息体对应的角色。 ● system：表示系统提示词角色。 ● user：表示用户输入角色。 ● assistant：表示模型回复角色。

content	是	String	消息具体内容。
			[
			{
			"type": "image_url",
			"image_url": {
			"url": "data:image/png;base64"
			}
			},
			{
			"type": "text",
			"text": "图里有什么"
			}
			]

name	否	String	对话参与者的可选名称，提供给模型信息以区分相同角色的不同对话参与者。

- **工具调用请求示例**
  - 流式请求示例

```
{
   "model": "model-img2text",
   "messages": [
      {
         "role": "system",
         "content": " You are a helpful assistant."
      },
      {
         "role": "user",
         "content": [
            {
               "type": "text",
               "text": "图里面有什么"
            },
            {
               "type": "image_url",
               "image_url": {
                  "url": "一个图片链接"
               }
            }
         ]
      }
   ],
   "tools": [
      {
         "type": "function",
         "function": {
            "name": "get_weather",
            "description": "获取给定地点的天气",
            "parameters": {
               "type": "object",
```

 版权所有 © 华为云计算技术有限公司

```
                "properties": {
                    "location": {
                        "type": "string",
                        "description": "地点，例如南京。"
                    }
                },
                "required": [
                    "location"
                ]
            }
        }
    }
  ],
  "max_tokens": 20,
  "presence_penalty": 1.2,
  "frequency_penalty": 1.0,
  "temperature": 0.5,
  "top_p": 0.95,
  "stream": true
}
```

- 非流式请求示例

```
{
  "model": "model-img2text",
  "messages": [
    {
        "role": "system",
        "content": " You are a helpful assistant."
    },
    {
        "role": "user",
        "content": [
            {
                "type": "text",
                "text": "图里面有什么"
            },
            {
                "type": "image_url",
                "image_url": {
                    "url": "一个图片链接"
                }
            }
        ]
    }
  ],
  "tools": [
    {
        "type": "function",
        "function": {
            "name": "get_weather",
            "description": "获取给定地点的天气",
            "parameters": {
                "type": "object",
                "properties": {
                    "location": {
                        "type": "string",
                        "description": "地点，例如南京。"
                    }
                },
                "required": [
                    "location"
                ]
            }
        }
    }
  ],
  "max_tokens": 20,
  "presence_penalty": 1.2,
  "frequency_penalty": 1.0,
  "temperature": 0.5,
```

 版权所有 © 华为云计算技术有限公司

```
    "top_p": 0.95,
    "stream": false
}
```

- **非工具调用请求示例**
  - 流式请求示例

```
{
    "model": "model-img2text",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant."
        },
        {
            "role": "user",
            "content": [ { "type": "text", "text": "图里面有什么" }, { "type": "image_url", "image_url": { "url": "一个图片链接" } } ]
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": true
}
```

  - 非流式请求示例

```
{
    "model": "model-img2text",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant."
        },
        {
            "role": "user",
            "content": [ { "type": "text", "text": "图里面有什么" }, { "type": "image_url", "image_url": { "url": "一个图片链接" } } ]
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": false
}
```

响应体参数

表 7-35 响应体参数

参数	参数类型	描述
id	String	图像理解文本对话唯一标识符。
choices	Array of **表 7-36** objects	返回体列表。 如果“n”大于1，则结果为多个。
created	long	问答发生的时间。格式为时间戳。
model	String	图像理解文本对话使用的模型名称。
object	String	固定值“chat.completion”。

 版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
usage	表7-40 object	图像理解文本对话用量统计。

表 7-36 ChatNonStreamingChoice

参数	参数类型	描述
index	Integer	返回多个choices时，每个choice对应的顺序。
message	表7-37 object	模型服务返回的具体消息体内容。
finish_reason	String	返回结束的原因。 • stop：模型达到自然停止点或提供的停止序列。 • length：达到请求中指定的最大token数。 • content_filter：由于内容过滤器的标志而省略了内容。 • tool_calls：模型选择了某个工具。

表 7-37 ChatMessageResponse

参数	参数类型	描述
content	String	返回消息体的内容，与tool_calls二选一。
role	String	返回消息体的角色。 • user：用户输入的问题。 • assistant：大模型的答复内容。
tool_calls	Array of 表7-38 objects	工具调用消息，与content二选一。
audio	ChatMessageAudio	聊天消息中的音频部分。
reasoningContent	String	用于展示模型的推理过程，帮助用户理解模型的决策依据。

表 7-38 ToolCall

参数	参数类型	描述
id	String	工具调用唯一标识符。
type	String	工具类型，当前仅支持function。
function	表7-39 Object	调用函数的详细信息。

版权所有 © 华为云计算技术有限公司

表 7-39 CallFunction

参数	参数类型	描述
name	String	函数名。
arguments	String	调用函数的参数，JSON格式。

表 7-40 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	回答包含的token数。
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

- **工具调用响应示例**
  - 流式响应示例


```
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"role":"assistant","content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{}}}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{"query":"最新款iPhone发布信息
","limit":5}}}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{"query":"最新款iPhone发布信息
","limit":5}},"index":0}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"正在为您搜索最新款iPhone的发布信息……
"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"finish_reason":"tool_calls","delta":{"tool_calls":[],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"usage":{"completion_tokens":67,"prompt_tokens":35,"total_tokens":102},"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":
```


 版权所有 © 华为云计算技术有限公司

[],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:[DONE]

- 非流式响应示例

{
  "choices": [
    {
      "message": {
        "role": "assistant",
        "tool_calls": [
          {
            "id": "call_12345",
            "type": "function",
            "function": {
              "name": "image_analysis",
              "arguments": "{\"detail_level\": \"high\", \"output_format\": \"json\"}"
            }
          }
        ]
      },
      "finish_reason": "tool_calls",
      "index": 0
    }
  ],
  "created": 1753965925754,
  "id": "model-img2text",
  "object": "chat.completion",
  "usage": {
    "prompt_tokens": 142,
    "completion_tokens": 28,
    "total_tokens": 170
  },
  "request_id": "xxx"
}

- **非工具调用响应示例**
  - 流式响应示例

data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"role":"assistant","content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"这是"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"Google"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"Chrome"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"浏览器"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"的"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"图标"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}

 版权所有 © 华为云计算技术有限公司

data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"。
"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"finish_reason":"stop","delta":{"tool_calls":[],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"usage":{"completion_tokens":43,"prompt_tokens":35,"total_tokens":78},"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:[DONE]

- 非流式响应示例

{
  "choices": [
    {
      "delta": {
        "role": "assistant",
        "content": "图像整体呈现出简洁、抽象的风格，主要内容是一个灰色的圆形头像轮廓。"
      },
      "finish_reason": "stop",
      "index": 0
    }
  ],
  "created": 1753965925754,
  "id": "model-img2text",
  "object": "chat.completions",
  "usage": {
    "prompt_tokens": 142,
    "completion_tokens": 118,
    "total_tokens": 260
  },
  "request_id": "xxx"
}

7.4 调测模型

模型调测是指在平台侧对模型服务进行真实调用，并结合参数调整与效果观测，验证模型在特定场景下的可用性、稳定性与效果表现的过程。通过调测，您可以在模型正式发布或接入业务流程前，提前发现并定位常见问题（如鉴权失败、接口协议不匹配、响应超时、输出不符合预期等），确保模型在真实业务场景中能够稳定、高效地运行。

调测时，平台将您的输入参数封装为标准API请求，发送至目标模型服务，并实时返回响应结果。您可在界面上查看完整的请求参数、响应内容、耗时及错误信息。

本章介绍平台已接入的模型服务调测流程。目前支持调测的模型类型包括：文本对话、图像理解、文本向量化（向量模型）、文本排序。

模型调测通常用于以下场景：

- 模型接入后验证：新接入的模型服务需要先调测确认可调用，再进行发布和使用。
- 变更后的回归验证：当您更新API Key、变更API地址、调整模型名称（模型ID/编码）或切换接口协议后，建议重新调测验证。
- 效果与参数优化：针对同一类问题，通过调整参数等方式对比输出效果，选择更合适的配置。

 版权所有 © 华为云计算技术有限公司

- 问题排查定位：当智能体/工作流调用模型出现异常时，可通过调测复现问题并查看调用结果与错误信息，快速缩小排查范围。

费用说明

调测功能本身免费，但调测过程中向模型发送的每一次对话请求均会消耗tokens，与正式调用的计费规则一致，详情请参见**费用说明**。

前提条件

已有可调测的模型服务或路由策略，满足以下任一条件：

- 调测自定义模型服务：已完成自定义模型服务接入（包括**已接入模型供应商**并成功**接入模型服务**）。
- 调测平台预置模型服务：免费额度下可直接调测，免费额度消耗完后需对预置模型服务进行开通和鉴权配置，详情请参见**接入预置模型**。
- 调测模型路由策略：已有可用的路由策略，详情请参见**配置模型路由策略**。

调测模型服务

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“开发中心 > 开发配置”，进入“模型调测”页签。

如需直接调测单个模型，可按以下入口操作：

- 调测预置模型（资产广场入口）：在左侧导航栏，选择“资产广场”，进入“模型”页面，鼠标移动至需调测的模型卡片上，单击“模型调测”。
- 调测自定义接入模型（自定义模型入口）：在左侧导航栏，选择“开发中心 > 开发配置”。进入“自定义模型”页签，进入所需调测的“供应商详情”页面，在需要调测的模型服务卡片上，单击“ ··· > 模型调测”。

步骤3 **可选:** 在“模型调测”界面，可以调测如下几种类型的模型服务。

文本对话

1. 在“模型类型”区域选择“文本对话”，参数配置请参考**表7-41**。

 版权所有 © 华为云计算技术有限公司

图 7-10 文本对话



版权所有 © 华为云计算技术有限公司

表 7-41 文本对话类型模型参数说明

参数	说明	示例
模型服务	“模型服务A”默认展示所选的供应商模型服务。“模型服务B”为可选项。 平台支持单模型调测和双模型对比调测。在对比模式下，您可同时选择模型服务A和B，对比两者的响应效果。 可通过下拉列表选择或切换以下类型的模型服务： – **用户自主接入的模型服务**：以模型供应商维度展示。 – **平台推荐**：以模型供应商维度展示。 – **路由策略**：用户自定义创建的路由策略。	GLM-5.2
深度思考	该参数仅在以下场景中显示： – **平台推荐**：所选模型为思考模型且支持关闭深度思考时显示，例如平台推荐的Qwen3-32B。 – **用户自主接入的模型服务**：所选模型为思考模型，且在新建模型服务时已开启“是否支持关闭思维链输出”时显示。 该参数支持以下操作： – **开启**：大模型将先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成详细的思维链内容并在调测界面展示，从而提升最终输出的准确性和可靠性。 – **关闭**：大模型将跳过思维链推理过程，直接生成最终答案，响应速度更快，适用于需要快速获取结果的场景。 不同模型能力下的生效情况： – 模型能力支持思维链输出，且支持关闭：开启和关闭均生效。 – 模型能力支持思维链输出，但不支持关闭：开启生效，关闭不生效。 – 模型能力不支持思维链输出：开启和关闭均不生效。	开启
输出方式	– **流式**（默认）：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，无需等待大语言模型生成完成。 – **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。	流式
最大回复长度	设置模型生成内容的Tokens长度上限。数值越大，支持的回复越长；数值越小，回复可能被截断。取值范围为100~1,048,576，默认值为2048。	2048

 版权所有 © 华为云计算技术有限公司

参数	说明	示例
生成随机性	生成随机性（temperature）控制模型输出的随机性和创造性，取值范围0.01~1，默认值为0.5。 – 数值越高（接近1）使输出更具多样性和创新性，适合创造性任务如小说创作。 – 数值越低（接近0）使输出更加确定和一致，适合基于事实的问答和售后客服场景。 建议该参数和“核采样”只设置1个，避免同时调整后效果不可控。	0.5
核采样	核采样（Top-P）用于限制模型的选词范围。 取值范围0~1，默认值为0.5。 – 值接近0时，模型只从概率最高的前10%词汇中选择，输出更集中。 – 值接近1时，不做限制，输出更多样。 建议该参数和“生成随机性”只设置1个，避免同时调整后效果不可控。	0.5
重复语句惩罚	用于控制词汇和句式的重复率。取值范围-2.0~2.0，默认值为0。 – 正值鼓励多样化表达，减少重复。 – 负值鼓励复用已出现词汇。 – 设为0则不做干预。 建议发现AI回复频繁重复时适当调高。	0
频率惩罚	频率惩罚（frequency_penalty），正值尽量避免使用常见的单词和短语，而更倾向于生成较少见的单词。取值范围-2.0~2.0，默认值为0。	0

2. 在右侧“效果预览”区域，在对话输入框输入测试语句后按Enter键或单击，查看模型响应结果。

   单击，清除本次会话内容，可以开始新的会话。

版权所有 © 华为云计算技术有限公司

图 7-11 文本对话模型调测成功



图像理解

1. 在“模型类型”区域选择“图像理解”，参数配置请参考**表7-42**。

 版权所有 © 华为云计算技术有限公司

图 7-12 图像理解



 版权所有 © 华为云计算技术有限公司

表 7-42 图像理解类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的供应商模型服务。也可以在下拉列表切换以下模型服务： - **用户自主接入的模型服务**：以模型供应商维度展示。 - **平台推荐**：以模型供应商维度展示。	Qwen2.5-VL-72B
输出方式	- **流式**（默认）：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，无需等待大语言模型生成完成。 - **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。	非流式
上传图片	单击，可上传本地图片。支持上传JPG、PNG格式图片，且不大于4MB，最多支持上传5张。	-
提示语内容	输入提示语，对图片进行提问。	图片里有什么？

2. 单击“生成图像理解”，在右侧“效果预览”区域查看模型响应效果。

图 7-13 图像理解模型调测成功



文本向量化

1. 在“模型类型”区域选择“文本向量化”，参数配置请参考**表7-43**。

版权所有 © 华为云计算技术有限公司

图 7-14 文本向量化



表 7-43 文本向量化类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的供应商模型服务。 您也可以在下拉列表切换以下模型服务： – **用户自主接入的模型服务**：以模型供应商维度展示。 – **平台推荐**：以模型供应商维度展示。	BGE-M3
请输入文本	输入待向量化的文本，可参照以下示例： – 示例1：那是个快乐的人 – 示例2： ["那是个快乐的人", "那是个高兴的人", "那是个忧郁的人"]	那是个快乐的人

2. 单击“生成向量化”，在右侧“效果预览”区域查看模型响应效果。

版权所有 © 华为云计算技术有限公司

图 7-15 文本向量化调测成功



文本排序

1. 在“模型类型”区域选择“文本排序”，参数配置请参考**表7-44**。

 版权所有 © 华为云计算技术有限公司

图 7-16 文本排序



 版权所有 © 华为云计算技术有限公司

表 7-44 文本排序类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的模型服务。 您也可以在下拉列表切换以下模型服务： – **用户自主接入的模型服务**：以模型供应商维度展示。 – **平台推荐**：以模型供应商维度展示。	BGE-Reranker-V2-M3
待排序文本	输入待排序文本。单击 + 添加文本，最多可以添加10条。	小朋友在学校很快乐
被展示文本条数	文本排序完成后，展示的条数。取值范围为1~10，默认值为1。	3
我的问题	描述想要解决的问题。	小朋友在学校怎么样?

2. 单击“开始排序”，在右侧“效果预览”区域查看模型响应效果。

图 7-17 文本排序调测成功



步骤4 调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

----结束

常见问题

接入后调测，模型无返回结果或者报错怎么办?

API Key填写错误，如多填写了Bearer前缀。

API Key本身已过期或无效，需要到模型服务商官网确认API Key状态，必要时重新生成。

7.5 配置模型路由策略

通过配置模型路由策略，您可以将多个已接入的模型服务按规则进行编排，实现模型故障自动切换能力。当主模型因故障、超时或不可用等原因无法正常响应时，系统将根据路由策略自动切换到其他可用模型继续提供服务，从而提升业务的稳定性与可用性。

版权所有 © 华为云计算技术有限公司

当模型出现以下情况时触发切换：返回HTTP 5xx错误、响应超时（超过策略总超时时间）、返回空响应等。

路由策略工作流程：调用模型A -> 失败 -> 重试 -> 仍失败 -> 切换模型B -> ...。

例如，某在线客服系统使用GLM-5.2作为主模型，当该模型服务因维护或故障不可用时，自动切换至备用模型DeepSeek-V4-Flash，确保客服服务不中断。

路由策略创建完成后，可对路由入口进行调测验证，并在智能体中选择使用。

费用说明

路由策略本身不收费，但模型调用按各模型服务的计费规则收费，详情请参见**费用说明**。重试和故障切换产生的调用次数均正常计费，建议根据业务需求合理设置重试次数。

适用场景

- 已接入多个同类型模型服务（例如多个文本对话模型），希望在主模型不可用时自动切换，提升服务连续性。
- 业务对稳定性要求较高，希望降低单一模型服务异常带来的影响。
- 希望在不修改智能体配置的情况下，通过调整路由策略实现模型切换或扩容。

前提条件

- 已**接入模型服务**。
- 路由中引用的模型服务建议已完成调测，确保模型服务本身可正常调用。
  路由策略解决的是“调用入口的高可用”问题，模型调测用于验证“模型服务/路由入口是否可用”。
  建议先**调测**各模型服务可用性，再配置路由策略并调测路由入口，以便快速定位问题。

创建路由策略

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航，选择“开发中心 > 开发配置”。

步骤3 选择“路由策略”页签，单击“创建路由策略”。

步骤4 在“创建路由策略”页面，配置参数信息，具体参数说明请参考**表7-45**，配置完成后单击“保存”。

界面提示“创建路由策略成功！”，新建的路由策略，显示在路由策略列表中。

 版权所有 © 华为云计算技术有限公司

图 7-18 创建路由策略



表 7-45 路由策略参数说明

参数	说明	示例
策略名称	自定义路由策略的名称。由2~36个字符组成，包含中英文、数字、中划线（-）、下划线（_）、点（.），仅支持以中英文开头。	文本对话路由策略

版权所有 © 华为云计算技术有限公司

参数	说明	示例
AI模型	AI模型即模型服务。 在“模型A”下拉框中选择模型服务。仅支持选择文本对话类型的模型服务。 单击“+ AI模型”，添加模型服务。至少需添加1个模型服务，最多支持3个。 路由策略提供模型服务时，模型调用顺序为：模型A > 模型B > 模型C，当模型A无法正常工作时，可以自动依次切换为模型B、模型C。 **验证故障切换**：可在模型调测时临时将模型A设为不可用，观察是否自动切换至模型B并正常返回结果。	模型A：GLM-5.2 模型B：DeepSeek-V4-Flash 模型C：Qwen3-32B
策略总超时时间	模型路由策略的总体超时时间。取值范围为1000~1,000,000ms，默认值为10,000ms。	10000ms
模型重试次数	路由策略中单个模型服务的重试次数。取值范围为0-100次，默认值为0次。 **注意** 路由策略本身不收费，但模型调用按各模型服务的计费规则收费。重试和故障切换产生的调用次数均正常计费，建议根据业务需求合理设置重试次数。	0
策略描述	路由策略的描述信息。由1~100个字符组成。	该策略为文本对话类型的路由策略。

步骤5 在“模型调测”区域，调测模型，具体参数说明请参考**表7-46**。

版权所有 © 华为云计算技术有限公司

图 7-19 调测模型



表 7-46 模型调测参数说明

参数	说明	示例
输出方式	可选非流式、流式。 • **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 • **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，无需等待大语言模型生成完成。默认**流式**。	流式
最大回复长度	设置模型生成内容的Tokens长度上限。数值越大，支持的回复越长；数值越小，回复可能被截断。取值范围为100~32768，默认值为2048。	2048

版权所有 © 华为云计算技术有限公司

参数	说明	示例
生成随机性	生成随机性（temperature）控制模型输出的随机性和创造性，取值范围0.01~1，默认值为0.5。 • 数值越高（接近1）使输出更具多样性和创新性，适合创造性任务如小说创作。 • 数值越低（接近0）使输出更加确定和一致，适合基于事实的问答和售后客服场景。 建议该参数和“核采样”只设置1个，避免同时调整后效果不可控。	0.5
核采样	核采样（Top-P）用于限制模型的选词范围。 取值范围0~1，默认值为0.5。 • 值接近0时，模型只从概率最高的前10%词汇中选择，输出更集中。 • 值接近1时，不做限制，输出更多样。 建议该参数和“生成随机性”只设置1个，避免同时调整后效果不可控。	0.5
重复语句惩罚	用于控制词汇和句式的重复率。取值范围-2.0~2.0，默认值为0。 • 正值鼓励多样化表达，减少重复。 • 负值鼓励复用已出现词汇。 • 设为0则不做干预。 建议发现AI回复频繁重复时适当调高。	0
频率惩罚	频率惩罚（frequency_penalty），正值尽量避免使用常见的单词和短语，而更倾向于生成较少见的单词。取值范围-2.0~2.0，默认值为0。	0

步骤6 在右侧“预览调试”区域，在对话输入框输入测试语句后按Enter键或单击，查看模型响应结果。

调测区域显示模型正常返回回复内容，且响应时间在预期范围内，表示路由策略配置正确。

单击，清除本次会话内容，可以开始新的会话。

版权所有 © 华为云计算技术有限公司

图 7-20 路由调测



----结束

常见问题

接入后调测，模型无返回结果或者报错怎么办?

API Key填写错误，如多填写了Bearer前缀。

API Key本身已过期或无效，需要到模型服务商官网确认API Key状态，必要时重新生成。

相关操作

在“路由策略”列表，支持的其他操作请参考表7-47。

表 7-47 相关操作

操作	说明
查看路由策略详情	在待查看的路由策略对应的“策略名称”列下，单击路由策略名称。
修改路由策略	在待修改的路由策略对应的“操作”列下，单击“编辑”。

版权所有 © 华为云计算技术有限公司

操作	说明
删除路由策略	**警告** 删除路由策略前，请确认没有智能体或工作流正在使用该策略。删除后，引用该策略的智能体将无法正常调用模型服务。 在待删除的路由策略对应的“操作”列下，单击“删除”。

相关文档

路由策略创建完成后，用户可在智能体、工作流中使用路由策略，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

版权所有 © 华为云计算技术有限公司

8 管理使用 AgentArts 的成员

AgentArts套餐包有成员使用限制，默认采用先到先得的自动许可机制，新成员加入时，系统自动授予访问许可，自动占用可用席位，席位不够将无法访问。如果套餐变更导致超额，将自动禁止最后加入的成员许可。

当登录AgentArts平台成员数量达到上限时，会提示“成员数已达套餐上限，暂无法加入”。具有“agentarts::updateMemberPermit”权限的用户，可以释放闲置成员或购买AgentArts其他套餐包来控制访问的成员。

“成员许可”页面，展示已接入的成员，具有“agentarts::updateMemberPermit”权限的用户，可以对某个用户分配许可、释放许可、禁止访问、解除禁止访问。

⚠ 注意

用户登录AgentArts服务后，才会在“成员许可”列表中展示。

费用说明

不同套餐包的**成员数量限制**不同，具体可参考**计费说明**。

当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

前提条件

- 已**开通AgentArts服务**。
- 登录用户具有“agentarts::updateMemberPermit”权限，具体操作请参考**给IAM用户授予使用AgentArts的权限**。

管理使用 AgentArts 的成员

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航，选择“系统管理 > 成员许可”。

步骤3 在“成员许可”列表中，可以对某个用户禁止访问、释放许可、分配许可。

- 分配许可：在对应的“操作”列下，单击“分配许可”，分配许可后，此用户正常访问AgentArts。

版权所有 © 华为云计算技术有限公司

- 释放许可：在对应的“操作”列下，单击“释放许可”，释放许可后，此用户无法正常访问AgentArts。
- 禁止访问：在对应的“操作”列下，单击“禁止访问”，禁止访问后，此用户无法访问AgentArts。
- 解除禁止：在对应的“操作”列下，单击“解除禁止”，解除禁止后，此用户无法访问AgentArts。

步骤4 （可选）单击“IAM用户管理”，IAM用户管理请参考**IAM用户管理**。

----结束

 版权所有 © 华为云计算技术有限公司

9 授权管理

9.1 管理云服务授权信息

- 如果需要使用预置的MaaS模型，平台提供2,000,000tokens的免费额度，可用于模型调测、智能体试运行，**免费额度用完**或通过API接口调用智能体，则需要开通预置的MaaS模型即服务并设置鉴权，否则会导致功能不可用。
- 如果需要对日志进行集中管理和分析，则需要开通云日志服务 LTS，否则会导致功能不可用。
- 如果需要追踪和分析应用的性能问题，则需要开通应用性能管理 APM，否则会导致功能不可用。
- 如果需要实时监控和分析应用的运行状态，则需要开通应用运维管理 AOM，否则会导致功能不可用。
- 如果需要高代码智能体打包镜像并部署至运行时，则需要开通及授权依赖的容器镜像服务 SWR，否则会导致功能不可用。
- 如果需要获取出站/入站认证信息，则需要开通智能体身份 AgentIdentity，否则会导致功能不可用。
- 如果需要托管智能体，则需要开通智能体网关 AgentGateway，否则会导致功能不可用。
- 如果需要智能体安全防护，则需要开通智能体卫士AI Defender，否则会导致功能不可用。
- 如果在智能体应用托管对话时，需要访问OBS桶，则需要开通对象存储服务 OBS，否则会导致对话功能受影响，例如文件上传下载、Skill加载。

开通服务授权不会产生费用，使用后按实际使用量计费，实际扣费以各服务的账单为准。

云服务授权

开通AgentArts套餐包成功后，返回AgentArts总览页，提示开通服务及授权，在“周边依赖服务开通及授权”弹窗中，单击“开通授权”，MaaS模型即服务不在此弹框中展示。云服务授权信息可以在“授权管理 > 云服务”页面查看。

如果未开通授权的云服务，在后续使用AgentArts时需要使用，可以在“授权管理 > 云服务”页面，进行开通。

版权所有 © 华为云计算技术有限公司

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航，选择“授权管理”。

步骤3 在“云服务 > 依赖云服务开通及授权”页面，单击“一键开通及授权”，也可以在对应云服务的“操作”列下，单击“一键开通授权”或“立即开通”。

- 一键开通及授权：一键开通及授权所有依赖的必选服务，可以勾选MaaS模型即服务。
- 一键开通授权：一键开通及授权所有依赖的必选服务。
- 立即开通：仅开通云服务，不支持授权，如果需要使用依赖的云服务，则还需授权。

云服务开通及授权后，AgentArts依赖该云服务的功能，才可以正常使用。

MaaS模型即服务授权开通后，会默认开通MaaS模型即服务及其模型，开通的模型展示在**管理模型授权信息**中。

----结束

相关操作

在“依赖云服务开通及授权”页面，支持的其他操作请参考**表9-1**。

表 9-1 云服务授权相关操作

参数	说明
查看云服务开通/授权状态	在云服务对应的“服务开通状态”列下，查看云服务开通/授权状态。
查看云服务账单	在页面右上角，单击“查看账单”。

9.2 管理模型授权信息

如果未开通MaaS模型即服务，则可以在“授权管理 > 模型”页面开通。

如果已开通MaaS模型即服务，则在“授权管理 > 模型”页面，显示已开通的模型服务，可以查看模型列表、模型版本信息、模型详情，调测模型。

开通模型服务

开通服务授权不会产生费用，使用后按实际使用量计费，实际扣费以各服务的账单为准。

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“授权管理”。

步骤3 选择“模型”页签，单击“批量开通模型”。

步骤4 在弹窗中，选择需要开通的模型，根据需求勾选“后续上新模型自动开通”，勾选“我已阅读并同意上述说明，及《ModelArts Studio 服务声明》”，单击“一键开通”。

版权所有 © 华为云计算技术有限公司

将默认开通MaaS模型即服务以及选择开通的模型。模型服务开通后，模型才能继续使用。平台提供2,000,000tokens的免费额度，可用于模型调测、智能体试运行。

----结束

相关操作

在“模型”页面，支持的其他操作请参考**表9-2**。

表 9-2 相关操作

参数	说明
查看模型版本信息	在模型左侧，单击 ，查看模型版本信息。
查看模型详情	单击模型名称或在模型对应的“操作”列下，单击“查看详情”。
调测模型	**注意** 已开通的模型，才支持调测。 视频生成、图片生成不支持调测。 在模型对应的“操作”列下，单击“模型调测”。
配置鉴权	单击“鉴权管理”配置鉴权，具体操作请参考**设置模型鉴权**。

版权所有 © 华为云计算技术有限公司

10 资源与订阅

资源与订阅界面，显示订阅管理及订购服务。

资源与订阅入口

步骤1 登录**AgentArts智能体平台**。

步骤2 在左侧导航栏，选择“资源与订阅”，进入资源与订阅界面。

----结束

资源与订阅介绍

- **订阅管理**

  显示生效的套餐包、套餐资源剩余数量及资源总数量、购买CU的入口，可以查看计费说明、购买记录。

  图 10-1 资源剩余量

  从左至右、从上至下依次为：

  - 展示当前正在使用的套餐信息。
  - 单击“计费说明”，可打开计费说明帮助文档。
  - 单击“购买记录”，显示订购的套餐包。
  - 展示“购买CU”的入口。

    当账号欠费且CU按需计费订单资源被删除后，此处将展示“开通CU按需使用”。如果不开通，将导致部分功能不可用。
  - 展示模型免费试运行资源（tokens）的剩余数量及总数量。

    当模型免费试运行资源（tokens）为“0/2,000,000”时，模型调测及智能体试运行需要开通模型及设置鉴权后才可继续使用，开通的具体操作请参考**管理模型授权信息**，设置鉴权请参考**接入预置的付费模型**。

版权所有 © 华为云计算技术有限公司

模型试运行资源的免费额度仅支持模型调测及智能体试运行，智能体发布后，如果用户需要使用MaaS模型即服务的模型，需要自行开通并设置鉴权，开通的具体操作请参考**管理模型授权信息**，设置鉴权请参考**接入预置的付费模型**。

- 展示知识库存储容量（GB）的剩余数量及总数量。
- 展示成员数量（名）的剩余数量及总数量。
- 展示计算单元数的剩余数量及总数量。

- **订购服务**

  显示AgentArts的套餐包信息及正在使用中的套餐包。可以进行套餐对比、购买其他的套餐包。

  当前AgentArts团队标准版、团队高级版、企业专业版、企业旗舰版未开放，如果需要购买，请提交工单咨询购买相关事宜。

  说明

  体验版与付费套餐（团队标准版、团队高级版、企业专业版、企业旗舰版）为独立订阅体系，一个账号支持同时购买。当您同时持有体验版和付费套餐时，系统将优先启用付费套餐；待付费套餐到期后，系统将自动回退至体验版，确保服务连续使用。

  团队标准版、团队高级版、企业专业版、企业旗舰版互斥，不可同时购买，同一时间仅可持有一个套餐。套餐支持由低到高升级，版本层级依次为：团队标准版 < 团队高级版 < 企业专业版 < 企业旗舰版，各版本均可升级至任意更高版本。

  购买CU包年套餐包前，须先开通/订阅体验版、团队标准版、团队高级版、企业专业版或企业旗舰版。

版权所有 © 华为云计算技术有限公司

11 审计

11.1 支持云审计的关键操作

操作场景

平台提供了云审计服务（Cloud Trace Service，简称CTS）。通过云审计服务，可记录与AgentArts相关的操作事件，便于日后的查询、审计和回溯。

前提条件

已**开通云审计服务**。

支持审计的关键操作列表

表 11-1 云审计服务支持的 AgentArts 服务操作列表

操作名称	资源类型	事件名称
创建函数	AgentManager	create_functions
编辑函数	AgentManager	update_functions
创建依赖包	AgentManager	create_dependencies
编辑依赖包	AgentManager	update_dependencies
删除依赖包	AgentManager	delete_dependencies
部署MCP服务	AgentManager	create_mcp_deploy
删除MCP服务	AgentManager	delete_mcp_service_delete
调测MCP服务	AgentManager	create_mcp_service_tools_test
MCP服务评分	AgentManager	create_mcp_server_rating

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
修改MCP服务信息	AgentManager	update_mcp_service_modify
最终租户授权给OP账号	AgentManager	create_doAuthorization
用户接入模型服务供应商	AgentManager	create_integration_providers
更新供应商信息	AgentManager	update_integration_providers
删除供应商信息	AgentManager	delete_integration_providers
认证配置设置	AgentManager	create_provider_auths
移除供应商认证配置	AgentManager	delete_provider_auths
新增模型服务	AgentManager	create_model_services
更新模型服务	AgentManager	update_model_services
删除模型服务	AgentManager	delete_model_services
发布模型服务	AgentManager	create_model_services_online
取消发布模型服务	AgentManager	create_model_services_offline
创建模型路由策略	AgentManager	create_model_router_strategies
删除模型路由策略	AgentManager	delete_model_router_strategies
修改模型路由策略	AgentManager	update_model_router_strategies
查询统计指标	AgentManager	create_statistic_indicator
获取span列表	AgentManager	create_spans
创建评测集	AgentManager	create_evaluation_sets
查询评测集列表	AgentManager	create_evaluation_sets_list
更新评测集	AgentManager	update_evaluation_sets
删除评测集	AgentManager	delete_evaluation_sets
数据添加标签	AgentManager	create_tag_mark
数据修改标签	AgentManager	update_tag_mark
删除数据标签	AgentManager	delete_tag_mark

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
删除评估器	AgentManager	delete_evaluators
创建评估器	AgentManager	create_evaluators
浏览评估器	AgentManager	create_list_evaluators
更新评估器草稿	AgentManager	update_evaluators_draft
提交评估器版本	AgentManager	create_evaluators_versions
更新评估器	AgentManager	update_evaluators
浏览会话列表	AgentManager	create_sessions
创建标签	AgentManager	create_tag
修改标签信息	AgentManager	update_tag
查询提示词	AgentManager	create_prompt_engineering_template_list
导出模板样例	AgentManager	create_prompt_engineering_template_sample_download
删除模板	AgentManager	delete_prompt_engineering_template_delete
获取OBS桶中文件和目录列表	AgentManager	create_obs_objects
语音合成	AgentManager	create_tts
检索知识库，命中测试	AgentManager	create_knowledge_search_text
删除知识库测试记录	AgentManager	delete_knowledge_search_records
创建一个知识库	AgentManager	create_knowledge
在知识库中上传一个文档	AgentManager	create_knowledge_files
新增知识文件的切片	AgentManager	create_knowledge_files_chunks
修改知识文件的切片	AgentManager	update_knowledge_files_chunks
删除知识文件的切片	AgentManager	delete_knowledge_files_chunks
创建默认知识库配置	AgentManager	create_configurations_default_connections

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
编辑默认知识库连接	AgentManager	update_configurations_default_connections
用于测试默认知识库连接是否正常	AgentManager	create_configurations_default_connections_test_connection
修改知识库可见性	AgentManager	update_permissions
创建一个Agent	AgentManager	create_agents
复制百宝箱应用	AgentManager	create_apps_copy
创建工作流	AgentManager	create_workflows
设置工作流试运行状态	AgentManager	update_workflows_test_status
工作流添加触发器	AgentManager	create_workflows_triggers_add
工作流编辑触发器	AgentManager	create_workflows_triggers_edit
工作流删除触发器	AgentManager	delete_workflows_triggers
修改工作流	AgentManager	update_workflows
删除工作流	AgentManager	delete_workflows
发布工作流版本	AgentManager	create_workflows_versions
删除一个工作流版本快照	AgentManager	delete_workflows_versions
发布一个工作流版本通道	AgentManager	create_workflows_channels
删除一个工作流版本通道	AgentManager	delete_workflows_channels
修改一个工作流版本通道	AgentManager	update_workflows_channels
复制一个Agent	AgentManager	create_agents_copy
修改一个Agent	AgentManager	update_agents
删除一个Agent，同步删除关联的assistant	AgentManager	delete_agents
修改一个知识库	AgentManager	update_knowledge
删除一个知识库	AgentManager	delete_knowledge

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
从知识库中删除一个文档	AgentManager	delete_knowledge_files
添加触发器	AgentManager	create_agents_triggers_add
编辑触发器	AgentManager	create_agents_triggers_edit
删除触发器	AgentManager	delete_agents_triggers
创建一个工具	AgentManager	create_tools
修改一个工具	AgentManager	update_tools
删除一个工具	AgentManager	delete_tools
创建一个插件鉴权凭证	AgentManager	create_tools_credentials
删除一个插件鉴权凭证	AgentManager	delete_tools_credentials
导出工作流	AgentManager	create_workflows_export
导入工作流	AgentManager	create_workflows_import
导出Agent	AgentManager	create_agents_export
导入Agent	AgentManager	create_agents_import
解析导入文件	AgentManager	create_list
发布一个Agent版本	AgentManager	create_agents_versions
删除一个Agent版本	AgentManager	delete_agents_versions
发布一个Agent版本通道	AgentManager	create_agents_channels
删除一个Agent版本通道	AgentManager	delete_agents_channels
修改一个Agent版本通道	AgentManager	update_agents_channels
创建一个工具OpenAPI定义	AgentManager	create_tools_openapi
修改一个意图包	AgentManager	update_complex_intent
删除一个意图包	AgentManager	delete_complex_intent
创建一个意图包	AgentManager	create_complex_intent
修改一个意图分支	AgentManager	update_complex_intent_branch
删除一个意图包分支	AgentManager	delete_complex_intent_branch
创建一个意图包分支	AgentManager	create_complex_intent_branch

 版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
导出高级意图配置模板	AgentManager	create_complex_intent_export_template
导出高级意图包	AgentManager	create_complex_intent_export
导入意图包	AgentManager	create_complex_intent_import
导入意图分支	AgentManager	create_complex_intent_branch_import
修改自定义对象	AgentManager	update_objects
删除自定义对象	AgentManager	delete_objects
创建自定义对象	AgentManager	create_objects
创建一个知识库分层规则	AgentManager	create_knowledge_rule_regex
修改一个知识库分层规则	AgentManager	update_knowledge_rule_regex
删除知识库分层规则	AgentManager	delete_knowledge_rule_regex
开启一个知识库	AgentManager	update_knowledge_start
停用一个知识库	AgentManager	update_knowledge_stop
创建一个FAQ问答对	AgentManager	create_knowledge_faq
修改一个FAQ问答对	AgentManager	update_knowledge_faq
删除一个FAQ问答对	AgentManager	delete_knowledge_faq
批量删除FAQ问答对	AgentManager	create_knowledge_faq_batch_delete
创建知识库任务	AgentManager	create_knowledge_tasks
在知识库中上传一个FAQ文档	AgentManager	create_knowledge_faq_files
删除知识FAQ文件	AgentManager	delete_knowledge_faq_files
新增知识FAQ文件的切片	AgentManager	create_knowledge_faq_files_chunks
修改知识FAQ文件的切片	AgentManager	update_knowledge_faq_files_chunks
删除知识FAQ文件的切片	AgentManager	delete_knowledge_faq_files_chunks

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
从知识库中批量删除文档	AgentManager	create_knowledge_files_batch_delete
上传头像	AgentManager	create_upload_avatar
修改一个数据源	AgentManager	update_datasource
删除一个数据源	AgentManager	delete_datasource
创建一个数据源	AgentManager	create_datasource
批量删除数据源	AgentManager	create_datasource_batch_delete
发布一个工具版本	AgentManager	create_tools_versions
删除一个工具版本定义	AgentManager	delete_tools_versions
根据工具ID创建工具OpenAPI定义	AgentManager	create_tools_openapi
删除知识库任务	AgentManager	create_knowledge_tasks_batch_delete
创建第三方知识库连接	AgentManager	create_connections
删除第三方知识库连接	AgentManager	delete_connections
编辑第三方知识库连接	AgentManager	update_connections
用于测试第三方知识库连接是否正常	AgentManager	create_connections_test_connection
新增第三方知识库	AgentManager	create_external
最终租户授权给OP账号	AgentManager	create_doAuthorization
修改结构化信息	AgentManager	update_structured_messages
添加结构化信息	AgentManager	create_structured_messages
批量删除结构化信息	AgentManager	delete_structured_messages
上传消息模板	AgentManager	create_structured_messages_upload_file
导出消息模板	AgentManager	create_structured_messages_export_file
结构化信息模板	AgentManager	create_structured_messages_export_template
创建一个插件	AgentManager	create_plugins

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
创建一个工具	AgentManager	create_tools
根据插件ID创建工具OpenAPI定义	AgentManager	create_plugins_openapi
删除一个插件	AgentManager	delete_plugins
删除一个插件版本定义	AgentManager	delete_plugins_versions
删除一个工具	AgentManager	delete_tools
导出插件	AgentManager	create_plugins_export
导入插件	AgentManager	create_plugins_import
修改一个插件	AgentManager	update_plugins
修改一个工具	AgentManager	update_tools
发布一个插件版本	AgentManager	create_plugins_versions
查询租户全量统计指标	AgentManager	create_statistic_indicator_domain
查询运营统计指标	AgentManager	create_statistic_indicator_operation
查询TOP指标	AgentManager	create_statistic_indicator_top
查询当前登录用户是否为白名单	AgentManager	create_statistic_config
创建实验任务	AgentManager	create_evaluation_experiment_submit
查询实验列表，带上实验状态	AgentManager	create_evaluation_experiment_list
删除历史实验	AgentManager	delete_evaluation_experiment_delete
查询实验数据结果列表	AgentManager	create_evaluation_experiment_data_detail
查询实验的输出指标	AgentManager	create_evaluation_experiment_metric
执行Agent，带会话ID	AgentManager	create_agents_conversations
执行Agent，不带会话ID	AgentManager	create_agents_conversations
执行Workflow	AgentManager	create_workflows_conversations

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
Agent对话上传文件	AgentManager	create_agents_upload_file
创建API Key	AgentManager	create_api_auth_api_keys
创建知识库OBS配置	AgentManager	create_obs_configs
根据OBS配置执行动作	AgentManager	create_obs_configs_execute
展示用户OBS指定路径下的文件对象名集合	AgentManager	create_obs_bucket_objects
更新会话变量	AgentManager	update_agents_conversations_variables
函数插件绑定函数	AgentManager	create_plugins_bind_function
删除函数	AgentManager	delete_functions
创建副本	AgentManager	create_prompt_tasks_copy
创建提示词优化任务	AgentManager	create_prompt_tasks
创建提示词优化任务草稿	AgentManager	create_prompt_tasks_draft
删除提示词优化任务	AgentManager	delete_prompt_tasks
暂停提示词优化任务	AgentManager	create_prompt_tasks_pause
继续提示词优化任务	AgentManager	create_prompt_tasks_resume
全量更新提示词优化任务（只能更新草稿）	AgentManager	update_prompt_tasks
向优化任务批量添加数据用例	AgentManager	create_prompt_tasks_data_batchItems
向优化任务添加单条数据用例	AgentManager	create_prompt_tasks_data_items
删除优化任务中的单条用例	AgentManager	delete_prompt_tasks_data_items
通过文件导入批量添加数据用例	AgentManager	create_prompt_tasks_data_import
更新优化任务中的单条用例	AgentManager	update_prompt_tasks_data_items
上传图片，获得临时URL	AgentManager	create_prompt_tasks_upload_picture

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
更新提示词	AgentManager	update_prompt
创建一个提示词	AgentManager	create_prompt
保存模板	AgentManager	create_prompt_tasks_prompt_save
提示词一键优化	AgentManager	create_prompt_generate
导入模板	AgentManager	create_prompt_engineering_template_import
导出模板	AgentManager	create_prompt_engineering_template_download
创建卡片	AgentManager	create_cards
更新卡片信息	AgentManager	update_cards
删除卡片	AgentManager	delete_card
删除一个卡片版本定义	AgentManager	delete_card_versions
发布一个卡片版本	AgentManager	create_card_versions
数据点赞点踩	AgentManager	create_tag_vote
上传一个写作模板	AgentManager	create_agents_uploadDeepResearchTemplate
数据清理	AgentManager	create_cleannotify
更新模型调测状态	AgentManager	create_model_test_status
批量删除用户的变量记忆	AgentManager	create_agents_memories_variables_batch_delete
重置用户的变量记忆	AgentManager	create_agents_memories_variables_reset
根据conversation_id删除会话	AgentManager	delete_agents_conversations_history
继续执行任务	AgentManager	create_workflows_tasks
修改任务信息	AgentManager	update_workflows_tasks
创建实验委托	AgentManager	create_evaluation_experiment_create_agency
创建智能体实例	AgentManager	create_deploy
模型调测	AgentManager	create_chat_completions
文本转语音	AgentManager	create_agents_audio_tts
语音转写	AgentManager	create_agents_audio_transcriptions

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
增加分析事件	AgentManager	create_agents_analytics_event
一句话语音识别	AgentManager	create_asr_short_audio
批量部署MCP服务	AgentManager	create_mcp_deploy_batch
一键添加到评测集	AgentManager	create_trace_evaluation_sets_items_batch_add
同步第三方模型服务	AgentManager	create_model_services_sync
代理解调-创建代理	lumina-console-backend	backend-createAgency
代理解调-创建代理角色	lumina-console-backend	backend-createAgencyRole
代理解调-删除代理角色	lumina-console-backend	backend-deleteAgencyRoles
查询隐私协议	lumina-console-backend	backend-query-agreement
签署隐私协议	lumina-console-backend	backend-create-agreement
订单回调-服务开通(支付完成)	lumina-console-backend	backend-orderPayCompleted
订单冻结解冻回调	lumina-console-backend	backend-changeResourceStatus
资源续费（订单操作）	lumina-console-backend	backend-renewalResource
一键开通及授权	lumina-console-backend	backend-enableCloudServiceList
临时订购接口	lumina-console-backend	backend-developerSpaceSubscribe
展示MCP列表	lumina-console-backend	backend-listExternalServers
展示MCP详情	lumina-console-backend	backend-queryExternalServiceDetail
根据SkuCode查询CBC报价，销售周期	lumina-console-backend	backend-queryResource
根据SkuCode查询变化的CBC报价	lumina-console-backend	backend-queryRMSResource

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
RMS调用-查询Console资源列表	lumina-console-backend	backend-queryRMSResourceList
清理租户数据（ROS调用）	lumina-console-backend	backend-cleanupResources
更新清理租户数据状态（定时）	lumina-console-backend	backend-updateStatusTask
开始清理子模块租户数据（定时）	lumina-console-backend	backend-cleanModuleTask
查询话单	lumina-console-backend	backend-getSdrBilling
一键开通免费版	lumina-console-backend	backend-oneClickSubscribe
SKU订购	lumina-console-backend	backend-subscribeOrder
变更下单	lumina-console-backend	backend-changeOrder
实例同步接口	lumina-console-backend	backend-syncInstance
资源同步接口	lumina-console-backend	backend-syncResource
Console同步信息到服务	AgentManager	create_common_console_subscribe_sync
License消费接口	AgentManager	create_common_intern_license_report
KMS加密接口	AgentManager	create_common_intern_kms_encrypt
KMS解密接口	AgentManager	create_common_intern_kms_decrypt
配置KMS	AgentManager	create_common_intern_kms_config
创建runtime	runtime	createRuntime
批量查询runtime	runtime	listRuntimes
查询单个runtime及指定版本详情	runtime	getRuntime
删除runtime	runtime	deleteRuntime
更新runtime	runtime	updateRuntime
批量查询版本	runtime	listVersions
创建入站网络	ingressNetwork	createIngressNetwork
查询入站网络列表	ingressNetwork	listIngressNetworks

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
删除入站网络	ingressNetwork	deleteIngressNetwork
创建runtime端点	runtimeEndpoint	createEndpoint
批量查询runtime端点	runtimeEndpoint	listEndpoints
查询单个runtime端点	runtimeEndpoint	getEndpoint
更新runtime端点	runtimeEndpoint	updateEndpoint
删除runtime端点	runtimeEndpoint	deleteEndpoint
开通服务	subscription	CreateCoreSubscription
查询开通状态	subscription	ShowCoreSubscription

11.2 在 CTS 事件列表查看云审计事件

场景描述

云审计服务能够为您提供云服务资源的操作记录，记录的信息包括发起操作的用户身份、IP地址、具体的操作内容的信息，以及操作返回的响应信息。根据这些操作记录，您可以很方便地实现安全审计、问题跟踪、资源定位，帮助您更好地规划和利用已有资源、甄别违规或高危操作。

本节介绍如何在云审计服务控制台查看或导出最近7天的操作记录。

什么是事件

事件即云审计服务追踪并保存的云服务资源的操作日志，操作包括用户对云服务资源新增、修改、删除等操作。您可以通过“事件”了解到谁在什么时间对系统哪些资源做了什么操作。

什么是管理类追踪器和数据类追踪器

管理追踪器会自动识别并关联当前用户所使用的所有云服务，并将当前用户的所有操作记录在该追踪器中。管理追踪器记录的是管理类事件，即用户对云服务资源新建、修改、删除等操作事件。

数据追踪器会记录用户对OBS桶中的数据操作的详细信息。数据类追踪器记录的是数据类事件，即OBS服务上报的用户对OBS桶中数据的操作事件，例如上传数据、下载数据等。

约束与限制

- 管理类追踪器未开启组织功能之前，单账号跟踪的事件可以通过云审计控制台查询。管理类追踪器开启组织功能之后，多账号的事件只能在账号自己的事件列表页面去查看，或者到组织追踪器配置的OBS桶中查看，也可以到组织追踪器配置的CTS/system日志流下面去查看。组织追踪器的详细介绍请参见**组织追踪器概述**。

版权所有 © 华为云计算技术有限公司

- 用户通过云审计控制台只能查询最近7天的操作记录，过期自动删除，不支持人工删除。如果需要查询超过7天的操作记录，您必须配置转储到对象存储服务（OBS）或云日志服务（LTS），才可在OBS桶或LTS日志流里面查看历史事件信息。否则，您将无法追溯7天以前的操作记录。
- 用户对云服务资源做出创建、修改、删除等操作后，1分钟内可以通过云审计控制台查询管理类事件操作记录。
- 数据类事件的审计日志，不支持在CTS控制台查看。若您想要查看数据类事件的审计日志，可按以下步骤操作：

  在配置数据类追踪器时，请开启“转储到LTS”开关。完成此项配置后，CTS记录的数据类事件将会自动转储至LTS中。之后，您便可以在LTS日志流里查看数据类事件的审计日志。如需了解详细操作流程，请参考**配置云审计事件转储至LTS并查看**。

前提条件

1. **注册华为云并实名认证。**

   如果您已有一个华为账户，请跳到下一个任务。如果您还没有华为账户，请参考以下步骤创建。

   a. 打开**华为云官网**，单击“注册”。

   b. 根据提示信息完成注册，详细操作请参见**注册华为账号并开通华为云**。

   注册成功后，系统会自动跳转至您的个人信息界面。

   c. 参考**实名认证**完成个人或企业账号实名认证。

2. **为用户添加操作权限。**

   如果您是以主账号登录华为云，请跳到下一个任务。

   如果您是以IAM用户登录华为云，需要联系CTS管理员（主账号或admin用户组中的用户）对IAM用户授予CTS FullAccess权限。授权方法请参见**给IAM用户授权**。

在 CTS 事件列表查看审计事件

步骤1 登录**CTS控制台**。

步骤2 单击左侧导航栏的“事件列表”，进入事件列表信息页面。

步骤3 在列表上方，可以通过筛选时间范围，查询最近1小时、最近1天、最近1周的操作事件，也可以自定义最近7天内任意时间段的操作事件。

步骤4 事件列表支持通过高级搜索来查询对应的操作事件，您可以在筛选器组合一个或多个筛选条件。

表 11-2 事件筛选参数说明

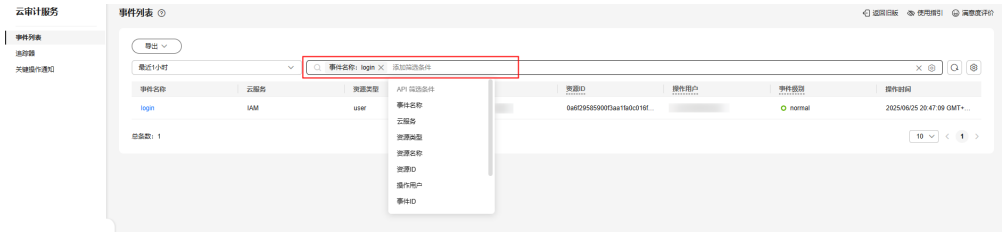
参数名称	说明
是否只读	下拉选项包含“是”、“否”，只可选择其中一项。 • 是：筛选只读操作事件，例如查询资源操作。当用户在“配置中心”页面开启了只读事件上报后，并触发了只读事件，才支持选择该选项。 • 否：筛选非只读操作事件，例如创建资源操作、修改资源操作、删除资源操作。

版权所有 © 华为云计算技术有限公司

参数名称	说明
事件名称	操作事件的名称。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 各个云服务支持审计的操作事件的名称请参见**支持审计的服务及详细操作列表**。 示例：updateAlarm
云服务	云服务的名称缩写。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 示例：IAM
资源名称	操作事件涉及的云资源名称。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 当该事件所涉及的云资源无资源名称或对应的API接口操作不涉及资源名称参数时，该字段为空。 示例：ecs-name
资源ID	操作事件涉及的云资源ID。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 当该资源类型无资源ID或资源创建失败时，该字段为空。 示例：*{虚拟机ID}*
事件ID	操作事件日志上报到CTS后，查看事件中的trace_id参数值。 输入的值需全字符匹配，不支持模糊匹配模式。 示例：01d18a1b-56ee-11f0-ac81-******1e229
资源类型	操作事件涉及的资源类型。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 各个云服务的资源类型请参见**支持审计的服务及详细操作列表**。 示例：user
操作用户	触发事件的操作用户。 下拉选项中选择一个或多个操作用户。 查看事件中的trace_type的值为“SystemAction”时，表示本次操作由服务内部触发，该条事件对应的操作用户可能为空。 IAM身份与操作用户对应关系，以及操作用户名称的格式说明，请参见**IAM身份与操作用户对应关系**。
事件级别	下拉选项包含“normal”、“warning”、“incident”，只可选择其中一项。 • normal代表操作成功。 • warning代表操作失败。 • incident代表比操作失败更严重的情况，如引起其他故障等。

版权所有 © 华为云计算技术有限公司

参数名称	说明
企业项目ID	资源所在的企业项目ID。 查看企业项目ID的方式：在EPS服务控制台的“项目管理”页面，可以查看企业项目ID。 示例：b305ea24-c930-4922-b4b9-******1eb2
访问密钥ID	访问密钥ID，包含临时访问凭证和永久访问密钥。 查看访问密钥ID的方式：在控制台右上方，用户名下拉选项中，选择“我的凭证 > 访问密钥”，可以查看访问密钥ID。 示例：HSTAB47V9V*******TLN9



步骤5 在事件列表页面，您还可以导出操作记录文件、刷新列表、设置列表展示信息等。

- 在搜索框中输入任意关键字，按下Enter键，可以在事件列表搜索符合条件的数据。
- 单击“导出”按钮，云审计服务会将查询结果以.xlsx格式的表格文件导出，该.xlsx文件包含了本次查询结果的所有事件，且最多导出5000条信息。
- 单击按钮，可以获取到事件操作记录的最新信息。
- 单击按钮，可以自定义事件列表的展示信息。启用表格内容折行开关，可让表格内容自动折行，禁用此功能将会截断文本，默认停用此开关。

----结束

相关文档

- 关于事件结构的关键字段详解，请参见**事件结构**和**事件样例**。
- 您可以通过以下示例，来学习如何查询具体的事件：
  - 使用云审计服务，审计最近两周内云硬盘服务的创建和删除操作。具体操作，请参见**安全审计**。
  - 使用云审计服务，定位现网某个弹性云服务器在某日上午发生的故障，以及定位现网创建弹性云服务器操作失败的问题。具体操作，请参见**问题定位**。
  - 使用云审计服务，查看某个弹性云服务器的所有的操作记录。具体操作，请参见**资源跟踪**。

版权所有 © 华为云计算技术有限公司

12 附录

12.1 开发常见报错

AgentArts包括智能体管理、组件库、开发配置、运营运维、资源与订阅等模块。详细展开来说包括：单智能体、工作流、多智能体、公共、组件（插件、MCP、卡片、提示词、知识库）、模型接入、配置管理(意图、数据源)、提示词工程、环境管理、License控制、记忆管理、导入导出管理等，下面我们将针对各个模块的错误码进行说明。

单智能体错误码

表 12-1 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101001	单智能体应用名称重复。	单智能体应用名称已存在。	请修改单智能体应用名称后重试。
AgentArts.02101002	指定的触发器在工作流中不存在，可能已被删除或未正确配置。	触发器不存在。	请确认传入的triggerId是否正确，检查该工作流是否确实存在该触发器。
AgentArts.02101003	当前添加的触发器类型不被支持。	触发器类型错误。	请确认触发器类型是否为TIMER或EVENT。
AgentArts.02101004	当前工作流下的触发器数量已达到最大限制。	触发器数量已达到上限。	每个工作流最多支持添加三个触发器，请删除部分触发器后再尝试添加。
AgentArts.02101005	当前未找到单智能体应用的图标配置或图标文件缺失。	单智能体应用图标不存在。	请检查并重新设置单智能体应用的图标，确保图标已正确上传或配置。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101006	上传的单智能体应用图标文件大小超过了系统限制。	单智能体应用图标超过了上传最大尺寸。	单智能体应用图标最大支持上传200KB，请压缩或更换尺寸较小的图标文件。
AgentArts.02101007	请求的智能体应用未找到或已被删除。	智能体应用不存在。	请确认该应用是否存在。
AgentArts.02101008	导入的单智能体应用文件格式错误或内容不完整。	单智能体应用导入失败。	请检查导入文件是否为合法导出的JSON格式，并确保文件内容完整无损。
AgentArts.02101009	智能体名称不符合命名规范、存在重复名称或导出数据序列化失败。	单智能体应用导出失败。	请检查智能体名称是否符合命名规则，确保名称唯一且格式正确，并确认数据未被破坏。
AgentArts.02101010	当前单智能体应用绑定的工具数量已超过系统限制的最大值。	单智能体应用绑定的工具数量超过上限。	单智能体应用最多可绑定20个工具，请移除部分工具后再进行绑定操作。
AgentArts.02101011	当前单智能体应用绑定的知识库数量已超过系统限制的最大值。	单智能体应用绑定的知识库数量超过上限。	单智能体应用绑定的知识库数量最大为3个。
AgentArts.02101012	当前单智能体应用绑定的工作流数量已超过系统限制的最大值。	单智能体应用绑定的工作流数量超过上限。	单智能体应用最多可绑定50个工作流，请移除部分工作流后再进行绑定操作。
AgentArts.02101013	当前设置的发布通道类型不被系统支持。	单智能体应用发布通道类型错误。	单智能体应用发布通道仅支持WEB_PAGE_CHANNEL和APP_STORE_CHANNEL两种类型，请检查并修改为合法类型。
AgentArts.02101014	当前设置的单智能体应用类型不在支持的枚举范围内。	不支持单智能体应用类型。	请使用支持的智能体类型，如Agent。
AgentArts.02101015	智能体应用已被人修改。	智能体应用已被更新。	请刷新页面后再试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101016	当前用户无权在指定项目中运行该单智能体应用。	单智能体应用执行权限不足。	请确认当前用户在目标项目中拥有该单智能体应用的执行权限。
AgentArts.02101017	请求中未提供合法的单智能体应用名称。	单智能体应用名称为null。	请填写单智能体应用名称，并确保参数格式正确后重试。
AgentArts.02101018	向客户端发送单智能体应用信息时发生异常，可能是连接中断或数据格式错误。	单智能体应用信息发送失败。	请检查客户端连接状态及数据格式，确保通信链路正常后重试。
AgentArts.02101019	单智能体应用执行时间超过系统设定的最大等待时间，导致运行中断。	单智能体应用运行超时。	请优化智能体逻辑或增加超时时间配置，确保在合理时间内完成执行。
AgentArts.02101020	当前单智能体应用的已发布版本数量已超过系统限制的最大值。	单智能体应用发布版本数量超过上限。	单智能体应用最多可发布100个版本，请删除旧版本后再进行发布操作。
AgentArts.02101021	智能体应用版本被删除。	智能体应用版本不存在。	请检查智能体应用版本是否存在。
AgentArts.02101022	请求的单智能体应用发布通道在系统中未找到，可能未发布或已被删除。	单智能体应用发布通道不存在。	请检查单智能体应用发布通道是否发布。
AgentArts.02101023	请求的单智能体应用发布版本在系统中未找到。	检查单智能体应用信息失败。	请检查单智能体应用是否已成功发布。
AgentArts.02101024	当前单智能体应用绑定的MCP数量已超过系统限制的最大值。	单智能体应用绑定的MCP数量超过上限。	单智能体应用绑定的MCP数量不超过5个。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101025	在删除定时触发器关联的调度作业时发生异常，可能是作业不存在或调度器状态异常。	调度器删除作业失败。	请检查调度器配置及作业状态，确保作业存在且调度器正常运行后再重试操作。
AgentArts.02101026	调度器在尝试立即执行任务时发生异常，可能是任务不存在或调度器未启动。	执行任务失败。	请确认任务名称和任务组名称是否正确，并确保调度器已正常启动后再重试操作。
AgentArts.02101027	在调用获取代理令牌的接口过程中发生异常，可能是网络问题、API服务不可用或输入参数不合法。	调用获取代理令牌API时发生异常。	请检查API请求是否正确，确保网络连接正常，确认API服务可用。
AgentArts.02101028	在将工作流配置中的ID转换为UUID的过程中发生异常，可能是数据格式错误或依赖对象缺失。	转换UUID失败。	请检查工作流配置数据的完整性，确保依赖的插件、MCP、子工作流等对象已正确导入，并验证配置格式是否符合规范。
AgentArts.02101029	当前接口仅支持流式调用。	不支持非流式调用。	请在请求头中设置stream=true，以使用流式调用方式。
AgentArts.02101030	上传文件数量超过20个。	上传文件数量超出限制。	请减少上传文件数量。
AgentArts.02101031	请求中指定的Agent在当前项目和工作区中不存在，或当前用户没有访问该Agent的权限。	Agent不存在或无权限。	请确认Agent ID是否正确，且当前用户对该项目和工作区中的Agent具有访问权限。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101032	请求中指定的Agent或其版本在系统中不存在，可能已被删除或未正确发布。	当前Agent版本不存在。	请确认Agent ID和版本信息是否正确，并确保该Agent已成功发布并可用。
AgentArts.02101033	单次会话上传文件个数不超过限制个数。	单次会话超过上传文件个数。	请减少单次会话上传文件个数后重试。
AgentArts.02101034	Agent中变量名称重复。	Agent中变量名称重复。	删除重复的变量名称。
AgentArts.02101035	Agent解析全局变量失败。	Agent解析全局变量失败。	请检查会话ID是否正确。
AgentArts.02101036	获取会话变量失败。	获取会话变量失败。	请检查资源ID、会话ID和变量名称是否正确。
AgentArts.02101037	版本名称已经存在。	版本名称已经存在。	请修改版本名称。
AgentArts.02101038	仅有op账号才能发布百宝箱。	发布百宝箱失败。	切换到op账号进行发布。
AgentArts.02101039	当前发布渠道类型不支持删除。	当前发布渠道类型不支持删除。	发布渠道类型不支持删除，当前仅应用百宝箱渠道支持删除。
AgentArts.02101040	Excel文件行数超过限制。	Excel文件行数超过限制。	Excel文件行数超过限制，请调小Excel文件行数。
AgentArts.02101041	Excel文件列数超过限制。	Excel文件列数超过限制。	Excel文件列数超过限制，请调小Excel文件列数。
AgentArts.02101042	文件格式无效。	Excel文件格式无效。	请换成Excel文件进行上传。
AgentArts.02101043	请求的应用发布通道在系统中未找到，可能未发布或已被删除。	应用发布通道不存在。	请检查应用发布通道是否发布。
AgentArts.02101044	缺少必填输入参数。	缺少输入参数。	请输入智能体名字和描述后重试。
AgentArts.02101045	大模型调用错误。	智能生成图标失败。	请稍等后重试。
AgentArts.02101046	指南数量超过最大限制。	指南数量超过最大限制。	请减少指南数量后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02101047	不支持非流式调用。	不支持非流式调用。	请以流式方式调用。
AgentArts.02101048	当前系统不支持当前类型资源添加。	不支持的关联资源类型。	请更换资源类型后重试。
AgentArts.02101049	Skill数量超过最大限制。	Skill数量超过最大限制。	请减少Skill数量后重试。
AgentArts.02101050	Skill不存在。	Skill不存在。	请确认Skill_ID是否正确且存在。
AgentArts.02101051	Skill版本无效或不存在。	Skill版本无效。	请确认当前Skill是否已启用对应版本。
AgentArts.02101052	指定的Skill或版本不存在或不符合预期配置。	Skill或版本不存在或不匹配。	请确保Skill、版本以及Skill配置正确无误。
AgentArts.02101054	智能体提示词为空。	缺少输入参数。	请输入完整有效的智能体提示词后重试。

表 12-2 错误码（后台报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103025	Agent接收到了无法识别的输入格式（如不支持的文件类型或非法的多模态数据流），导致推理引擎无法解析任务指令。	Agent输入类型不支持。	请检查传递给Agent的变量类型，确保其符合当前模型支持的输入规范（如文本、图片链接或标准JSON）。
AgentArts.103028	用于控制Agent动态选择工具的切换逻辑字典（Tool Switch Dict）格式不正确，导致Agent无法在多个工具间进行路由。	工具切换配置非法。	请核对tool_switch字段的结构，确保其为一个合法的键值对映射，且Key值与已注册的工具名称一致。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103029	尝试动态更新Agent的技能集时传入了非法对象。系统要求输入RestFul API或Function的列表。	插件更新类型错误。	请确保在更新插件列表时，传入的对象是标准的API定义或函数定义类，而非原始字符串或其他基础类型。
AgentArts.103030	Agent内部生成的任务追踪ID或外部传入的task_id长度超过了系统预设的最大阈值，可能导致数据库写入失败。	任务ID长度超限。	请缩短任务标识符的长度，或检查任务生成逻辑中是否存在重复拼接ID的异常。
AgentArts.103031	在通过Model Context Protocol (MCP)扩展Agent能力时，传入的参数类型不符合规范。	MCP插件更新类型错误。	请核对MCP接口定义，确保传入的是合法的MCP协议对象列表。

工作流错误码

表 12-3 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201001	在读取或解析工作流导入文件时发生异常，可能是文件格式错误、内容不完整或包含非法数据。	工作流导入失败。	请确保导入文件为合法的JSONL格式，且内容符合工作流配置规范。
AgentArts.02201002	组件被删除。	工作流组件ID不存在。	请检查组件是否存在。
AgentArts.02201003	参数长度不符合预期。	工作流节点执行异常。	请检查传入的数值范围参数，确保其为成对的两个数值，格式应为[起始值, 结束值]。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201004	请求中指定的工作流ID在当前项目和工作区中不存在，可能已被删除或未正确创建。	工作流或该工作流版本号不存在。	请确认工作流ID是否正确，并确保该工作流在当前项目和工作区中已存在。
AgentArts.02201005	工作流的配置信息或结构校验未通过，可能包含非法节点、格式错误或违反系统规则的内容。	工作流信息校验失败。	请检查工作流的配置内容，确保其符合系统定义的格式规范和校验规则，并修正错误后重试。
AgentArts.02201006	工作流中文名称长度不超过64字符。	工作流中文名称超出字符限制。	请减少工作流中文名称长度后重试。
AgentArts.02201007	工作流图标不能为空。	工作流图标名为空。	请上传工作流图标。
AgentArts.02201008	在生成或处理工作流导出文件时发生异常，可能是数据格式错误、资源读取失败或系统内部错误。	工作流导出失败。	请确认导出的工作流配置完整且合法。
AgentArts.02201009	工作流可能已被他人更新或操作过于频繁。	当前工作流不是最新。	请刷新页面后再操作。
AgentArts.02201010	账号权限不足。	工作流版本无权限删除。	更换更大权限的账号。
AgentArts.02201011	当前配置的工作流发布通道类型不合法。	工作流发布通道类型错误。	工作流发布通道仅支持WEB_PAGE_CHANNEL和APP_STORE_CHANNEL两种类型。
AgentArts.02201012	当前工作流嵌套的子工作流层级已超过系统限制的最大深度。	工作流嵌套超过最大层级。	工作流最大嵌套深度不超过10层。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201013	工作流版本不存在。	导出依赖的工作流版本不存在。	请检查依赖工作流是否存在。
AgentArts.02201014	当前工作区中已存在相同中文名称的工作流。	工作流中文名称重复。	请修改工作流中文名称后重试。
AgentArts.02201015	当前工作区中已存在相同英文名称的工作流。	工作流英文名称重复。	请修改工作流英文名称后重试。
AgentArts.02201016	工作流非游离节点过多。	工作流非游离节点数量超过最大限制。	工作流非游离节点个数不超过150个。
AgentArts.02201017	工作流版本与当前部署版本不一致，无法进行校验操作。	工作流校验接口异常。	请确保请求中传入的工作流版本与当前工作流的最新版本一致后再进行校验。
AgentArts.02201018	输入内容涉及敏感内容。	工作流内容审核错误。	请重新输入非敏感对话。
AgentArts.02201019	控制器绑定的工作流节点类型或结构不符合规范，无法正确解析或执行。	控制器绑定工作流节点不合法。	请检查绑定的工作流节点是否为合法的数组类型，并确保其schema结构完整且符合预期格式。
AgentArts.02201020	当前执行请求的projectId与智能体所属projectId不一致。	工作流执行权限不足。	请确保executeParams中的projectId与智能体所属projectId一致。
AgentArts.02201021	当前操作的工作流版本已过期，可能已被其他用户修改或保存了新版本。	不是最新的工作流。	请刷新网页后重试。
AgentArts.02201022	工作流执行时间超过了系统设定的超时限制，导致运行被中断。	工作流运行超时。	请优化工作流执行逻辑以减少耗时，或在执行参数中适当增加超时时间。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201023	根据提供的任务ID未能查询到对应的有效异步任务记录。	工作流异步任务ID不存在。	请确认传入的taskId、workflowId、projectId和workspaceId是否正确，确保任务已存在且未过期。
AgentArts.02201024	当前尝试恢复的任务状态不是"PENDING"（阻塞状态），无法进行恢复操作。	工作流异步任务需要为阻塞状态。	确保该任务当前处于PENDING状态后再进行resumeTask操作，或检查任务状态流转逻辑是否符合预期。
AgentArts.02201025	他节点正在处理异步任务，锁已被占用。	其他节点正在处理异步任务。	请确认系统中是否已存在正在运行的异步任务处理节点。
AgentArts.02201026	在尝试异步执行工作流任务时发生异常，可能是任务执行逻辑错误、资源不足或线程池提交失败等原因导致。	执行异步工作流任务失败。	请检查executeWorkflow方法的执行逻辑是否正常，确认线程池taskAsyncPoolSize配置是否合理，并查看日志中具体的异常信息以进一步排查问题。
AgentArts.02201027	当前用户不是该工作流的创建者，因此没有权限修改其属主信息。	此工作流并非由您创建，您无权对其进行修改。	请确认操作用户是否为工作流的创建者。
AgentArts.02201029	根据提供的工作流ID和版本ID未能查询到对应的有效版本信息，可能版本已被删除或未正确发布。	未找到工作流版本。	请确认workflowId和versionId是否正确。
AgentArts.02201030	该工作流尚未完成发布流程。	工作流尚未发布。	请先在工作流编辑页面完成发布操作。
AgentArts.02201031	任务型工作流屏蔽了节点。	任务型工作流不允许添加节点。	请删除任务型工作流中已屏蔽的节点。
AgentArts.02201032	当前请求的projectId所属租户不再允许发布工作流。	该租户没有发布工作流权限。	请确认当前操作的projectId是否为系统允许发布工作流的项目，或联系管理员配置租户发布权限。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201033	在处理工作流输出节点的引用类型时，解析引用内容失败。	输出参数未正确引用变量。	请检查输出节点中引用字段的配置内容。
AgentArts.02201034	在解析工作流的复杂意图节点时，未能找到与该节点正确连接的目标节点。	该节点没有被连接。	请检查工作流中复杂意图节点的连接关系，确保其通过有效的边连接到目标节点。
AgentArts.02201037	上传工作流图标过大。	工作流图标大小超过限制。	请压缩或裁剪图标图片，确保其大小不超过系统设定的限制后再进行上传。
AgentArts.02201038	处理工作流字段值时，其类型不被当前系统支持。	工作流信息校验失败，不支持的工作流字段值类型。	请检查工作流字段值的配置类型，确保其为系统支持的合法类型，并根据实际需求进行修正。
AgentArts.02201039	在转换工作流字段的JSON Schema时，发现其嵌套层级超过了系统设定的最大深度限制。	工作流信息校验失败，JSON Schema超过最大深度。	请简化JSON Schema的嵌套层级，确保其深度不超过限定深度。
AgentArts.02201040	当前项目中已发布的自定义节点数量已达到系统设定的最大限制，无法继续新增自定义节点。	自定义节点数量超过最大限制。	请删除部分已有的自定义节点。
AgentArts.02201041	当前项目没有权限执行发布或取消发布自定义节点的操作。	项目ID没有权限发布或取消发布自定义节点。	请确认操作权限，仅OP账户允许进行自定义节点的发布或取消发布操作。
AgentArts.02201042	在配置内容审核规则时，某一类型的敏感词组数量超过了系统设定的最大限制。	内容审核错误，类型大小超过限制。	请检查对应类型的敏感词配置，减少敏感词组数量，确保不超过限制数量后再进行保存。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201043	在配置内容审核规则时，某一类型中的敏感词组内容长度超过了系统设定的最大限制。	内容审核错误，类型长度超过限制。	请检查对应类型的敏感词组内容，缩短其字符长度，确保不超过限定字符后再进行保存。
AgentArts.02201044	在配置内容审核规则时，输入回复文本的长度超过了系统设定的最大限制。	输入审核错误，类型长度超过限定长度。	请检查对应类型的输入回复文本内容，缩短其字符长度，确保不超过限定字符后再进行保存。
AgentArts.02201045	在配置内容审核规则时，输出回复文本的长度超过了系统设定的最大限制。	输出审核错误，内容长度超过限定长度。	请检查对应类型的输出回复文本内容，缩短其字符长度，确保不超过限定字符后再进行保存。
AgentArts.02201046	在配置内容审核规则时，替换词的长度超过了系统设定的最大限制。	替换词审核错误，类型长度超过限定长度。	请检查对应类型的替换词内容，缩短其字符长度，确保不超过限定长度字符后再进行保存。
AgentArts.02201047	子工作流不存在或未发布。	子工作流不可用。	请确认子工作流ID是否正确，并确保该工作流已发布且属于当前工作区。
AgentArts.02201048	在处理工作流分支字段值时，传入的字段值类型不被支持。	工作流分支字段值类型不受支持。	请检查字段值的类型配置。
AgentArts.02201049	提问器节点字段配置时，发现了一个不被支持的验证器类型。	不受支持的验证器类型。	请检查字段的验证器配置，支持的类型有LENGTH_LIMIT_VALIDATE、NUMBER_RANGE_VALIDATE、ENUM_LEGALITY_VALIDATE、DATE_TIME_FORMAT、COMMON_VALIDATE。
AgentArts.02201050	在解析日期时间格式配置时，传入的参数格式不正确或为空。	提问器节点解析日期时间格式失败。	请确保日期时间格式配置为一个非空字符串，并符合标准的日期时间格式。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201051	在解析长度限制配置时，传入的参数不是合法的整数或格式不正确。	提问器节点解析长度限制格式失败。	请确保长度限制配置为一个非空字符串，并且可以被正确转换为整数。
AgentArts.02201052	在解析数字参数时，传入的字符串无法被转换为合法浮点数。	提问器节点解析数字格式失败。	请确保传入的数字参数为合法的数字字符串。
AgentArts.02201053	在解析整数参数时，传入的字符串无法被转换为合法的整数。	提问器节点解析整数格式失败。	请确保传入的整数参数为合法的整数字符串。
AgentArts.02201054	在解析枚举值时，传入的字符串格式不正确，无法被正确拆分为多个枚举项。	提问器节点解析枚举格式失败。	请确保枚举值为以逗号分隔的字符串列表。
AgentArts.02201055	在解析数字范围参数时，传入的参数个数不正确，或参数格式不合法。	提问器节点解析数字范围参数失败。	请确保传入的范围参数为两个合法的数字。
AgentArts.02201056	在解析通用验证器参数时，参数格式不正确或缺少必要字段。	提问器节点设置格式无效。	请确保传入的参数为一个合法的字符串列表，并符合该验证器的格式要求。
AgentArts.02201057	在处理工作流表达式时，使用了一个不被支持或未定义的操作符。	未知的操作符。	请检查操作符是否正确。
AgentArts.02201059	客户端在工作流执行过程中主动断开了流式接口的连接，导致执行流程被中断。	客户端断开流式接口请求。	请确保客户端在工作流执行期间保持连接稳定，避免在执行过程中手动关闭或中断请求。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201060	循环节点的循环次数设置不正确。	工作流循环节点按次数循环范围应该为1-1000之间。	请检查循环节点的配置，将循环次数设置为 1 到 1000 之间的整数后再进行保存。
AgentArts.02201061	您尝试访问的工作流不在当前项目或工作区的权限范围内。	无权限使用该工作流。	请检查您是否有权限访问该工作流，或者确认工作流ID是否正确。
AgentArts.02201062	绑定知识库数量超过了允许的最大限制。	知识库节点绑定知识库过多。	请减少选择的知识库数量至限定个数以内。
AgentArts.02201063	在执行工作流过程中发生异常，可能由于参数校验失败、资源不可用、内部服务调用失败或线程执行异常等原因导致。	运行工作流失败。	请检查工作流配置是否正确，确认输入参数是否合法，查看相关服务是否正常运行。
AgentArts.02201064	节点的异常处理配置缺失或格式不正确。	节点异常处理配置不合法。	请检查节点名称的配置项中是否包含合法的异常处理配置。
AgentArts.02201065	节点超时时间范围：0.1-900。	节点超时时间设置不合法。	请检查节点的超时时间设置是否合法。
AgentArts.02201066	在当前工作流中引用的子工作流ID不存在。	子工作流不存在。	请检查当前工作流中引用的子工作流ID是否正确。
AgentArts.02201067	在解析工作流节点时，节点引用的资源不存在或解析过程中发生异常。	工作流节点引用的资源不存在或异常。	请检查节点的配置，确认其引用的资源是否已正确创建并存在于当前项目中，同时确保资源格式和类型符合要求。
AgentArts.02201068	当前工作区不存在指定ID的数据库资源。	数据库不存在。	请确认数据库ID是否正确，并检查该数据库资源是否已被删除。
AgentArts.02201069	当前账号未开通IAM委托。	异步执行任务需账号开启IAM委托。	请开通IAM委托权限。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201070	当前任务支持超时。	异步执行任务超时。	请增加异步任务执行时长。
AgentArts.02201071	异步执行任务失败。	异步执行任务失败。	请依据报错信息检查应用配置。
AgentArts.02201072	代码节点已不支持沙箱执行方式。	代码节点已不支持沙箱执行方式。	代码沙箱不再提供服务，建议使用高可靠、高性能、高弹性、高安全的FunctionGraph执行方式。
AgentArts.02201073	请求的工作流应用发布通道在系统中未找到，可能未发布或已被删除。	工作流应用发布通道不存在。	请检查工作流应用发布通道是否发布。
AgentArts.02201074	当前发布渠道类型不支持删除。	当前发布渠道类型不支持删除。	发布渠道类型不支持删除，当前仅应用百宝箱渠道支持删除。
AgentArts.02201075	数据源执行SQL失败。	数据源执行SQL失败。	请检查数据源状态是否连通。
AgentArts.02201076	查询节点查询字段不能为空。	查询节点查询字段不能为空。	请检查查询字段是否为空。
AgentArts.02201077	工作流节点异常重试次数设置不合法。	节点异常重试次数配置不合法。	请检查节点异常重试次数配置，最大重试次数为3。
AgentArts.02201078	流式输出节点不支持重试、超时、返回设定内容、执行异常流程。	流式输出节点异常设置不合法。	请检查流式输出节点是否设置了重试、超时、返回设定内容、执行异常流程。
AgentArts.02201079	大模型节点配置为非流式输出格式非json，异常处理方式只支持"中断流程"。	大模型节点异常处理方式配置不合法。	请检查大模型节点是否输出配置为：非流式非json，异常处理方式配置了非"中断流程"。
AgentArts.02201080	工作流节点不支持配置重试。	工作流节点重试设置不合法。	请检查工作流节点是否设置了重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201081	节点异常处理类型配置不合法，合法的选项有：中断流程，返回设定内容，执行异常流程。	节点异常处理方式配置不合法。	请检查节点异常处理方式配置是否合法。
AgentArts.02201082	工作流未试运行成功不能发布。	工作流未试运行成功不能发布。	请先调试工作流成功，再发布。
AgentArts.02201084	获取执行ID调试事件时发生异常。可能原因： 1. 工作流实例不存在或已删除； 2. 参数提取事件处理失败； 3. 调试事件数据转换或排序过程中出现错误。	获取调试事件失败。	方案1：确认执行ID是否正确，检查该执行是否存在； 方案2：重新发起请求，触发工作流重新执行； 方案3：检查系统日志，查找具体的异常堆栈信息； 方案4：联系技术团队，提供执行ID以便排查问题。
AgentArts.02201085	工作流包含受限节点。	工作流包含受限节点。	请删除工作流受限节点。
AgentArts.02201088	提问器节点配置参数长度超出最大限制。	提问器节点配置参数长度超出最大限制。	请减少配置参数的长度。
AgentArts.02201089	该接口禁止访问。	该接口禁止访问。	更换访问接口。
AgentArts.02201090	解析第三方工作流文件失败。	解析第三方工作流文件失败。	请排查工作流导出文件是否正确。
AgentArts.02201091	目标服务地址受到访问限制。	工作流访问地址受限。	请检查工作流中节点访问策略。
AgentArts.02201092	不支持的工作流类型。	不支持的工作流类型。	请检查导入文件的工作流类型是否支持。
AgentArts.02201093	无权限查询数据库。	无权限查询数据库。	请确认当前数据源配置与查询的数据库保持一致。
AgentArts.02201094	SQL转换失败。	SQL转换失败。	请检查SQL语句。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02201095	运行时环境启动中，请稍后重试。	运行时环境启动中。	请稍后重试。
AgentArts.02201096	运行时环境不存在。	运行时环境不存在。	• 试运行提示这个错误码请联系客服处理。 • API调用提示这个错误码可参考如下操作重试，如重试失败请联系客服处理。 1. 查找到智能体运行时的实例，单击“删除”，删除智能体运行时的实例。查找智能体运行时实例方法以单智能体为例，参考**管理应用**中的“查看智能体运行时数据”。工作流、多智能体复制ID时切换到对应页签即可。 2. 重新发布部署智能体，具体操作参考如下： – 单智能体：**发布并部署应用为API服务**。 – 工作流：**发布并部署工作流**。 – 多智能体：**部署多智能体应用为API**。
AgentArts.02201097	使用的沙箱资源超过配额限制。	使用的沙箱资源超过配额限制。	请减少使用频率，并稍后重试。

表 12-4 错误码（后台报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101501	工作流启动失败，定义的全局变量未能获取到有效的初始赋值。	开始组件全局变量缺失。	请检查工作流触发时的输入参数，确保所有在开始节点定义的全局变量均已传入或设有默认值。
AgentArts.101531	结束节点在工作流运行结束时尝试初始化输出逻辑失败。	结束组件初始化失败。	请检查结束组件的输出结构配置，确保其引用的变量在整个工作流上下文中是可达的。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101532	结束节点在尝试按照预设模板格式化最终输出数据时发生异常。	结束组件模板组装错误。	请检查结束组件的Response模板语法，确保引用的占位符与实际输出变量名匹配。
AgentArts.101533	在流式（Streaming）输出模式下，结束节点无法正确聚合或转发最后的流片段。	结束组件流数据处理错误。	请检查客户端连接状态，并确认前置节点的流式输出是否完整闭合。
AgentArts.101561	LLM节点尝试加载模型配置或建立连接失败。	LLM组件初始化失败。	请检查工作流中该节点的模型选择、API Key配置及Endpoint连通性。
AgentArts.101562	LLM节点的提示词模板组装失败，可能存在引用的变量未定义或语法错误。	LLM组件模板组装错误。	请核对提示词中的变量占位符，确保其已在节点输入中正确映射。
AgentArts.101563	在与大模型交互过程中，流式响应中途断开或返回非法格式。	LLM获取流式输出错误。	请检查网络环境的稳定性，或确认模型服务商是否存在并发配额限制及超时问题。
AgentArts.101591	脚本组件在准备运行环境或解析输入参数时发生异常。	代码组件初始化失败。	请检查代码节点的输入映射配置，确保所有引用的变量已正确连接。
AgentArts.101592	无法向代码执行沙箱发起请求，可能是沙箱服务未启动、网络不可达或接口鉴权失败。	代码沙箱请求失败。	请检查沙箱服务的运行状态及网络配置，并确认系统环境变量中的沙箱端点地址是否正确。
AgentArts.101593	用户编写的代码在沙箱内执行时报错。	代码沙箱执行错误。	请根据sandbox_msg中的报错信息（如语法错误、索引越界或超时等）检查并修正您的代码逻辑。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101594	沙箱服务内部发生崩溃或遇到了未处理的系统异常，导致执行非正常中断。	代码沙箱非预期错误。	请尝试重新运行工作流。如果问题持续，请联系系统管理员检查沙箱服务的资源占用及日志信息。
AgentArts.101595	代码组件在执行过程中发生了非预期的运行时异常。	代码调用执行异常。	请检查代码中的逻辑，特别是涉及外部调用、第三方库引用或大规模数据处理的部分。
AgentArts.120007	代码节点执行完毕后，返回的字典中不包含画布上定义的输出变量名。	代码返回值Key缺失。	请检查Python/JS脚本的return语句，确保返回的字典键名与节点“输出设置”中的定义严格一致。
AgentArts.120008	代码节点的返回对象类型非法。系统要求返回一个dict（字典），以便进行变量映射，但实际收到了其他类型。	代码返回值类型错误。	请修改代码逻辑，确保最终结果以{"key": value}的格式返回，不要直接返回字符串或列表。
AgentArts.101651	消息节点在准备发送逻辑或加载配置时发生异常。	消息组件初始化失败。	请检查消息节点的输入参数映射，确保消息类型（如文本、卡片）配置完整。
AgentArts.101652	消息组件未配置必选的消息模板内容，导致无法生成待发送的消息主体。	消息模板缺失。	请在组件配置中填写消息内容模板，或选择一个已有的模板进行引用。
AgentArts.101653	在将变量渲染到消息模板时发生错误，通常是由于引用的占位符变量不存在。	消息模板组装错误。	请检查模板中的变量引用语法，确保所有占位符对应的变量已在该节点的输入列表中定义。
AgentArts.101654	同步调用消息组件发送接口时发生未捕获的运行时异常。	消息组件调用非预期错误。	请查看系统后台日志排查网络连接或第三方消息网关的返回状态。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101098	提示词执行引擎在提交渲染后的最终指令阶段发生非预期中断。	意图识别提示词执行异常。	请确认提示词引擎状态，并检查完整提示词长度是否超过了模型上下文窗口限制。
AgentArts.101097	将业务参数转换为LLM消息序列（Role/Content结构）时发生格式错误。	模型输入消息转换失败。	请检查对话历史记录的组织形式，确保Role定义符合该模型的要求。
AgentArts.101096	请求大模型推理服务进行意图判定时，服务端返回错误或网络连接超时。	意图识别模型调用失败。	请确认模型API状态、网络连通性以及API Key的配额是否充足。
AgentArts.101095	在对用户原始输入进行清洗、格式化或截断处理时发生逻辑错误。	意图识别输入预处理失败。	请检查输入数据是否为空或包含异常特殊字符，确保满足组件的处理要求。
AgentArts.101094	系统在根据意图列表和用户输入变量渲染提示词模板时发生异常。	意图识别模板创建失败。	请核对模板中的占位符是否与组件输入的字段名一致，检查是否有非法语法字符。
AgentArts.102053	模板语法存在错误（如大括号未闭合、使用了非法转义字符），导致Jinja2或String格式化引擎无法渲染最终结果。	模板格式化错误。	请检查提示词文本中的语法格式，确保占位符使用正确且没有语法逻辑冲突。
AgentArts.103004	规划引擎在请求大模型进行任务规划时，收到错误响应或连接超时。	模型调用失败。	请检查后端模型服务的可用性、API Key有效性及网络连通性。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101042	提问组件接收到的响应格式不符合预期（例如期望结构化数据但收到纯文本），导致无法继续后续逻辑。	提问组件响应类型非法。	请检查反问逻辑中的输出约束，或在提示词中明确要求模型返回特定的数据类型。
AgentArts.101043	提问组件为了补全关键信息与用户进行的交互轮数超过了系统设定的阈值，为防止无限循环已强制中断。	超过最大追问次数。	请优化提问策略以更高效地收集信息，或在组件配置中调大最大对话轮数。
AgentArts.101044	组件配置了不支持的问题生成方式，导致系统无法正确构造向用户提问的文案。	问题构建方法非法。	请检查组件配置中的“提问方式”选项，确保选择了系统当前支持的构建引擎或模板。
AgentArts.101045	在设置为“直接收集用户回复”的模式下，未配置发送给用户的问题描述，导致交互无法开始。	直接采集模式下问题内容为空。	请在该组件的配置项中填写具体的提问文案，用于引导用户进行输入。
AgentArts.101047	提问组件关联的安全护栏（Rails）模块无法加载配置或初始化引擎失败。	护栏初始化失败。	请检查安全策略配置是否符合语法规范，并确认相关依赖库已正确安装。
AgentArts.101048	在对用户输入进行安全检查（自定义输入护栏）时发生异常，导致无法判断内容是否合规。	输入护栏执行失败。	请检查自定义输入护栏的脚本逻辑或后端检测服务的可用性。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101049	在组件执行过程中的特定安全检查环节触发了自定义护栏逻辑异常。	自定义执行护栏错误。	请排查自定义校验规则的代码是否存在逻辑漏洞或执行超时。
AgentArts.101050	系统内置的默认安全护栏在处理组件输出结果时发生非预期崩溃。	默认执行护栏错误。	请检查模型输出是否包含特殊不可见字符，并确保系统内置护栏规则库为最新版本。
AgentArts.101059	提问组件在进行“自我反思”时模型调用失败或逻辑中断。	提问组件反思逻辑失败。	请检查反思步骤关联的模型状态，确保其提示词能够引导模型输出正确的判定结论。
AgentArts.120000	在执行问答（QA）类节点或组件时，预设的备选选项配置为空，导致模型或系统无法进行选择。	问答选项列表为空。	请检查问答节点的配置，确保“选项”列表中至少包含一个有效的备选项。
AgentArts.120001	问答节点配置了系统无法识别的检索或匹配策略。	不支持的问答策略。	请在节点设置中重新选择支持的匹配策略（如“精确匹配”或“语义匹配”）。
AgentArts.120002	系统尝试根据索引获取选项或数据时，索引值超过了数组或列表的实际长度。	索引超出范围。	请检查前置逻辑生成的数据量，确保索引访问在有效区间内。
AgentArts.120003	问答组件在处理输入数据时，未能找到指定的参数键。	输入键缺失。	请检查节点间的连线映射，确保上游节点正确输出了变量。
AgentArts.101741	API组件在尝试加载接口配置、初始化鉴权信息或准备执行环境时发生异常。	API组件初始化失败。	请检查API组件的连接器配置，确保目标接口的访问地址及基本认证参数正确无误。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101742	API组件接收到的输入参数不符合目标接口的要求，如缺少必填字段或参数类型不匹配。	API组件参数校验错误。	请根据接口文档核对并修正该组件的参数映射，确保传入数据格式合法。
AgentArts.101743	API组件的输入端（Inputs）配置缺失，导致无法构造有效的请求报文发送至目标服务器。	API组件输入未定义。	请在组件配置面板中定义输入变量，并确保其已正确连接至前置节点的输出数据。
AgentArts.101744	目标API成功响应，但在尝试解析响应体（Response Body）或将结果映射到工作流变量时失败。	API组件输出处理错误。	请检查响应解析逻辑（如JSON Path），确保其与目标接口实际返回的数据结构保持一致。
AgentArts.101745	在执行网络请求期间发生底层错误，包括但不限于连接超时、DNS解析失败或收到5xx系列服务器错误。	API组件执行异常。	请确认外部服务是否运行正常，并检查网络防火墙或代理服务器是否允许编排引擎访问该接口地址。
AgentArts.105001	插件系统触发了未捕获的运行时异常，通常由于插件包逻辑错误或与当前引擎版本不兼容导致。	插件未知异常。	请检查插件运行日志，定位具体的崩溃堆栈信息，并确认插件的依赖库是否完整。
AgentArts.105002	尝试上传或注册一个标识符（Plugin ID）已存在的插件。	插件已存在。	请更改插件名称或版本号，或使用“更新插件”接口对现有插件进行覆盖安装。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.105003	工作流在调用插件时，无法在插件库或文件系统中检索到对应的执行文件或配置。	插件未找到。	请确认该插件是否已在插件管理页面成功发布，并检查ID引用是否正确。
AgentArts.105004	调用插件时传入的动态参数不符合插件定义的规范（如缺少必填项或类型不匹配）。	插件参数校验失败。	请根据插件定义的输入规范调整工作流中的输入变量，确保参数合法。
AgentArts.105006	插件管理服务在读写插件元数据、权限或状态时发生数据库连接错误。	插件数据库异常。	请检查后端数据库服务状态及连接池配额，确认插件配置表未被锁定。
AgentArts.105009	在对插件描述进行语义向量化以便进行搜索推荐时，向量模型服务返回错误。	插件向量推理失败。	请检查Embedding模型服务的可用性，确认插件的描述文本长度未超过模型限制。
AgentArts.105010	调用外部API插件时，目标服务器返回了身份验证错误（如401 Unauthorized）。	API鉴权失败。	请重新检查插件配置中的API Key、Bearer Token或OAuth2凭证是否有效且未过期。
AgentArts.105011	外部API响应了非200/201的异常状态码（如403，404，500等）。	API响应状态码错误。	请根据返回的具体HTTP状态码排查第三方服务逻辑，确保接口地址和请求方法正确。
AgentArts.105012	外部API在预设时间内未返回数据。可能是由于网络延迟或三方服务处理耗时过长。	API请求超时。	请在插件设置中适当调大超时阈值，或检查目标服务的网络连通性。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.105013	插件返回的数据量超过了系统设定的安全负载上限，可能导致内存溢出。	响应体过大。	请限制API返回的数据条数，或在插件中启用分页/流式处理。
AgentArts.105014	系统尝试通过代理服务器访问外部插件API时，代理服务器拒绝连接或无法解析目标。	代理连接失败。	请检查全局代理配置（Proxy Server/Port/Credentials）是否正确，并确认内网穿透状态。
AgentArts.101801	系统无法根据中间表示（IR）构造子工作流实例。这通常是因为引用的子工作流已被删除、未发布或ID不存在。	子工作流构建失败	请确认子工作流是否存在且处于已发布状态，并重新保存父工作流以刷新引用。
AgentArts.101803	父工作流中定义的子工作流节点参数配置不完整或不符合子工作流的输入要求。	子工作流配置校验失败。	请进入该组件的参数映射界面，确保所有子工作流要求的必填输入项均已正确映射。
AgentArts.101804	子工作流在运行过程中由于内部节点报错导致流程中断。	子工作流执行异常。	请查看该子工作流的具体执行日志，定位内部具体哪个节点触发了异常。
AgentArts.101805	递归调用冲突：检测到子工作流ID与当前父工作流ID相同。系统禁止自循环引用以防止死循环。	非法的子工作流引用。	请更换引用的子工作流，避免在工作流中直接或间接地调用自身。
AgentArts.101811	循环组件配置了系统无法识别的循环模式。	不支持的循环类型。	请检查循环节点的类型设置，确保其为系统支持的模式（如按次数循环或按列表项迭代）。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101812	在执行循环时，关键配置字段为空，导致循环引擎无法确定起始条件或迭代对象。	循环组件必填项缺失。	请在循环组件配置中填写缺失的字段，例如迭代列表或终止条件。
AgentArts.101813	字段的输入类型与循环模式要求的类型不符（例如：列表循环要求输入为list，但实际收到了string）。	循环组件字段类型错误。	请检查前置节点传递给循环组件的变量类型，确保数据格式满足循环逻辑的要求。
AgentArts.101831	尝试对不同类型的数据进行聚合（例如将数字和对象强行合并到同一个列表），导致类型冲突。	聚合值类型不一致。	请确保所有参与聚合的数据项类型一致，或在聚合前使用脚本节点统一数据格式。
AgentArts.101832	系统无法执行指定的聚合策略（例如无效的合并算法或摘要算法）。	不支持的聚合策略。	请在聚合组件配置中选择有效的策略，如“追加到列表”、“对象合并”或“字符串拼接”。
AgentArts.101850	组件要求的必填输入字段未能获取到任何数据（None或空字符串）。	输入字段不能为空。	请检查工作流的上游节点是否输出了该字段，或在当前节点为该字段配置默认值。
AgentArts.101862	循环次数配置超出了系统允许的范围[1,1000]，旨在防止因配置失误导致的死循环或资源耗尽。	循环次数非法。	请将循环次数调整为1到1000之间的整数。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101870	系统接收到无法识别的MCP组件类型可能是协议版本不匹配或引用了未注册的扩展协议。	不支持的MCP类型。	请检查MCP服务器的协议版本，并确保在编排引擎中正确配置了对应的协议处理器。
AgentArts.101872	在执行协议交互时，关键字段为空。	MCP组件必填项缺失。	请完善MCP组件的配置信息，确保所有标识外部服务或资源的必填参数均已填写。
AgentArts.101873	在与MCP服务器进行通信或执行远程指令时发生异常	MCP执行错误。	请检查外部MCP服务的运行状态、网络连通性以及该服务生成的详细错误日志。
AgentArts.101874	尝试调用MCP协议中未定义或当前服务器未实现的方法	不支持的MCP方法。	请核对MCP服务器支持的功能列表（Capabilities），确保调用的方法在协议规范内。
AgentArts.101875	参数的输入值类型与协议预定义的类型不符。	MCP参数类型错误。	请检查工作流中传递给MCP组件的变量类型，确保其符合协议定义的规范。
AgentArts.100703	执行SQL指令时发生数据库层面异常。	SQL执行失败。	请根据MySQL返回的具体错误原因排查是否涉及表结构缺失、数据冲突或连接超时。
AgentArts.100704	构建的SQL命令未能通过预检查，可能是语法错误或包含非法字符。	SQL语法校验失败。	请核对生成的动态SQL语句，确保表名、字段名拼写正确且符合语法规范。
AgentArts.101623	分支表达式中包含了系统禁止的安全敏感字符或不受支持的特殊符号。	表达式包含非法字符。	请移除表达式中的非法字符，仅使用标准的操作符、数字、字符串及已定义的变量名。
AgentArts.101624	分支逻辑在判定条件的真伪时触发了底层异常。	条件评估异常。	请根据详细的错误提示检查逻辑分支的配置，确保复杂条件的嵌套逻辑清晰且无冲突。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101681	意图识别组件在加载大语言模型配置或建立服务连接时发生错误	意图识别LLM初始化失败	请检查该节点选择的模型名称、API Key以及对应的模型推理服务状态。
AgentArts.101093	模型返回的结果无法解析为预期的意图标签，通常由于输出格式不符合规范	意图识别输出解析错误	请检查提示词中的输出格式约束，并确保模型具备生成结构化数据的能力。
AgentArts.120005	全局变量的默认值类型与定义的变量类型（如String、Integer、Object）不匹配。	全局变量类型错误。	请在全局变量管理页面核对该变量的类型定义，并确保默认值符合该格式。
AgentArts.101901	加载组件的具体配置时失败，原因可能是该组件已下线或版本配置不匹配。	组件配置加载失败。	请更新该节点的组件版本，或在组件库中重新确认该组件的发布状态。
AgentArts.101902	工作流的中间表示（Intermediate Representation）未能通过语法或逻辑一致性校验。	IR校验错误。	请根据详细提示修复画布中的逻辑冲突（如循环依赖或必填端口未连线）。
AgentArts.101903	工作流或节点ID包含了系统禁止的特殊字符，目前仅支持数字、字母和下划线。	无效的ID格式。	请修改ID名称，移除空格、连字符（-）或其他特殊符号。
AgentArts.101904	启动工作流时传入的初始参数不完整，必须提供完整的IR（定义）或workflow_state（状态恢复数据）。	初始化数据类型错误。	请检查调用接口的Payload，确保提供了正确的启动参数。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101905	节点的运行时间超过了预设的上限，系统为防止资源长期占用已自动终止该任务。	组件执行超时	请检查该节点的业务逻辑是否存在死循环或长耗时操作，或者在组件设置中调大执行超时时间限制。
AgentArts.101906	工作流在启动阶段无法正确加载运行环境、解析IR定义或分配执行资源。	工作流初始化失败。	请根据报错原因排查工作流配置的完整性，并确认编排引擎的系统资源是否充裕。
AgentArts.101907	执行实例已被手动中止或由于系统级指令强制关停。	工作流执行被终止。	如果这不是预期的手动操作，请检查是否有外部监控服务触发了强制停止逻辑。
AgentArts.101912	引擎尝试从后端存储检索异步任务的运行结果时失败。	获取异步结果失败。	请确认任务存储服务的连通性，并检查任务结果是否已过期被清理。
AgentArts.101913	尝试停止异步任务时，由于任务已完成、任务已不存在或调度器拒绝操作而失败。	取消异步任务失败。	请确认任务的实时状态，如果任务仍在运行但无法取消，请检查调度引擎的响应能力。
AgentArts.101914	工作流异步引擎无法连接到Celery。	Celery代理初始化失败。	请检查中间件服务的运行状态及连接字符串配置，确保网络防火墙已放行对应端口。
AgentArts.101920	执行节点尝试在异步工作流调度中心注册自身身份时发生错误。	ApsWorker注册失败。	请检查工作节点的网络环境及集群配置，确保各节点ID唯一且心跳同步正常。
AgentArts.101921	ApsWorker在监听Redis消息发布订阅（Pub/Sub）频道时发生连接中断或数据解析错误。	Redis监听器异常。	请确认Redis服务的稳定性，并检查是否存在大并发消息导致监听缓冲区溢出。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101922	异步工作节点在尝试识别本地主机标识时失败，导致无法完成服务注册。	获取本地主机名错误。	请检查运行环境的系统配置，确保/etc/hosts或相关的DNS配置允许正确解析主机名。
AgentArts.101923	异步执行任务时发现Worker ID不合法或已失效，导致任务无法被正确指派或认领。	Worker ID异常。	请重启该节点的Worker服务，确保系统自动分配一个有效的唯一标识。
AgentArts.101924	执行流程进入了预设的“异常结束”节点，该节点主动触发了工作流的非正常终止逻辑。	异常终点组件触发终止。	这是由于业务逻辑触碰了预设的报错路径（Error Path），请根据业务流程排查为何会流转至此异常终点。
AgentArts.101931	工作流在尝试进行流式数据推送到前端或下游时，发现注册的回调处理实例（Callback Instance）已销毁、未初始化或类型不匹配。	工作流流式回调实例无效。	请检查流式调用接口的上下文生命周期，确保在整个执行期间回调对象保持活跃，且没有被提前回收。
AgentArts.101032	工作流拓扑结构定义存在逻辑断点或不合法，导致执行引擎无法定位后续待执行的节点。	工作流定义失效。	请检查工作流画布中的节点连接线是否完整，确保没有孤立节点，并重新保存/发布工作流以刷新定义。
AgentArts.101039	工作流运行至具体的某个节点时，组件内部逻辑抛出异常。该错误通常由组件特有的运行环境或业务逻辑触发。	组件执行错误。	请查看对应节点的输入输出日志，根据报错详情定位并修复具体组件的配置问题。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101040	工作流引擎在调度或管理执行生命周期时遇到系统级错误。这可能涉及内存溢出、线程中断或数据库状态同步失败。	工作流执行异常。	请尝试重新触发工作流任务，如果错误持续出现，请联系系统管理员检查后台服务运行状态。
AgentArts.101046	工作流中存在逻辑回环，且运行时的总跳转次数或循环轮数已超过系统全局设定的硬性上限。此机制用于防止无限循环耗尽计算资源。	工作流循环执行超限。	请优化工作流逻辑，确保循环具备明确的跳出条件，或检查是否存在逻辑设计死循环。
AgentArts.101051	工作流在触发异常回调时，未能成功捕获或序列化当前执行栈的错误上下文。	回调异常信息采集失败。	请检查回调记录器的存储空间或外部日志系统的连接状态，确保系统有权限写入异常追踪数据。
AgentArts.101052	在对组件进行单步跟踪或调试模式执行时发生错误。	组件单步调试异常。	请检查调试器连接状态及组件的上下文信息。
AgentArts.101053	当前组件类型（如某些内置原子组件或加密组件）不支持单步调试功能。	组件类型不支持调试。	请确认组件定义中是否开启了调试支持，或尝试直接运行而非调试模式执行。
AgentArts.101054	单个组件的IR定义不符合规范，解析失败。	组件IR非法。	请检查组件的配置文件或DSL编译结果，确保其IR结构完整。
AgentArts.101057	组件的初始化检查失败，其输出字段必须与输入字段保持一致（如状态透传组件）。	组件输入输出一致性校验失败。	请调整组件定义，确保指定的输入输出字段名和类型完全匹配。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101058	尝试在全局上下文中写入数据，但已存在的全局项类型不是字典（dict），导致无法合并或追加。	全局上下文类型冲突。	请检查全局状态管理逻辑，确保在整个编排流程中全局变量的基准类型是一致的。
AgentArts.101621	条件分支节点中的表达式不符合逻辑运算语法规范，导致解析器无法识别。	分支表达式语法错误。	请检查表达式中的括号是否闭合、操作符（如==，!=，>，<）是否书写正确，以及变量名是否合法。
AgentArts.101622	在执行表达式时发生运行时错误，通常是由于引用的变量为None或类型不匹配（例如尝试将字符串与数字进行比较）。	表达式计算错误。	请确保表达式中引用的所有变量在执行到该节点时均已有值，且数据类型支持所选的操作符。
AgentArts.101041	规划器组件配置了系统无法识别的规划算法类型，导致任务拆解引擎无法初始化。	规划器类型非法。	请检查Planner节点的“规划模式”设置，确保选择了当前版本支持的合法类型。
AgentArts.109510	系统无法为当前的对话会话创建上下文处理器，可能是由于自定义处理逻辑脚本加载失败。	处理器创建失败。	请检查上下文处理流的插件或中间件配置，确保相关类名或函数路径正确。
AgentArts.101908	任务节点执行的循环次数超过了系统或节点设定的硬上限，通常是为了防止逻辑死循环。	任务达到最大迭代次数。	请检查循环逻辑的终止条件是否正确，确保在预期内能正常退出循环。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101911	尝试对同一个执行实例进行并发异步操作，或分布式环境下任务锁定冲突。	异步执行冲突。	请确保不要对同一执行ID同时发起多个状态变更请求，并检查分布式锁（例如Redis Lock）的状态。
AgentArts.102308	在对生成的IR进行自我检查和反思修正（Self-Reflection）时，修正逻辑报错。	工作流IR反思修正失败。	优化反思阶段的提示词设置，使其能够给出更明确的纠错指令。
AgentArts.102303	在自然语言转为Agent的第一阶段，系统尝试理解用户模糊意图并引导澄清时发生错误。	意图澄清预处理失败。	请调整预处理阶段的提示词指令，使其能够更健壮地处理非结构化用户输入。
AgentArts.102302	将IR转换为可在华为云平台运行的特定DSL描述时失败。	IR转换为云端DSL失败。	请检查中间表示逻辑是否过于复杂，超出了目标云平台DSL的支持范围。
AgentArts.102305	模型虽然设计了结构，IR JSON时失败。	工作流IR生成失败。	请检查后端生成模板，确保其生成的JSON结构符合规范。
AgentArts.102304	模型无法根据用户需求设计出合理的节点拓扑结构（如节点间连线冲突或逻辑环路）。	工作流结构设计失败。	请尝试简化用户需求描述，或者增加可用工具/插件的描述信息以便模型更好地拆解。
AgentArts.102307	在NL2Agent动态构建工作流时，当前生成的会话状态或工作流状态不匹配，导致流程中断。	工作流状态机异常。	请清除当前生成缓存，尝试启动一个新的生成会话。
AgentArts.102306	NL2Agent专用模型在调用过程中返回错误。	NL2Agent模型服务异常。	检查模型服务的并发限额及健康状况。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.101956	系统在启动时扫描插件目录或建立组件索引失败。	组件池初始化失败。	请检查插件目录权限、配置文件路径以及是否有损坏的第三方组件包导致扫描中断。
AgentArts.101951	尝试在组件池中注册组件，但该类型名称已被其他组件占用。	组件类冲突。	请检查自定义组件的命名空间或类型定义，确保每个扩展组件在全局范围内具有唯一的类型标识。
AgentArts.101953	组件池注册接口期望接收到一个类对象（Class），但实际传入了实例、函数或其他非类类型的数据。	注册对象类型错误。	请确保在调用注册方法时传递的是组件类本身，而非已实例化的对象或方法引用。
AgentArts.101952	组件类型名称不符合系统命名规范（如包含特殊字符或关键字冲突）	无效的组件类名称。	请修改组件类的注册名称，仅使用合法的字符组合，并避开系统的保留字。
AgentArts.101955	组件池找到了类定义，但在调用构造函数创建组件实例时抛出异常。这通常由组件构造函数中的逻辑错误或缺少必要的构造参数引起。	组件实例化失败。	请检查组件类的__init__方法代码，确保其能够独立于特定执行上下文正确创建对象。
AgentArts.101954	工作流引擎尝试加载节点，但在当前运行环境的组件池中未找到对应的注册信息。	组件类未找到。	请确认该自定义组件是否已正确导入并执行了注册逻辑，或者检查IR中定义的类型名称是否拼写正确
AgentArts.109300	记忆引擎的IR配置文件存在逻辑错误或格式不合法，导致无法正确映射记忆分片与检索策略。	记忆IR配置错误。	请检查记忆模块的DSL或JSON配置文件，确保记忆类型（如语义记忆、情节记忆）的参数定义符合规范。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.120006	当前工作流包含的节点总数超过了系统允许的最大限制。	工作流节点超限。	建议将复杂的工作流拆分为多个子工作流（Sub-workflow），或精简冗余节点以优化执行效率。

多智能体错误码

表 12-5 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02301001	系统中未找到有效的控制器节点信息。	多智能体缺失控制器节点。	请检查配置文件和节点分组逻辑，确保控制器节点存在并正确分组。
AgentArts.02301002	当前控制器中未配置任何节点，无法进行多智能体流程构建。	多智能体节点为空。	请在控制器中添加至少一个有效的节点，确保节点列表不为空。
AgentArts.02301003	当前控制器配置中存在重复的子工作流ID，导致无法正常构建流程。	多智能体存在重复的子工作流。	请检查工作流配置，确保每个子工作流的ID在控制器中是唯一的。
AgentArts.02301004	控制器节点的配置信息为空，无法进行后续流程校验和执行。	多智能体控制器节点配置为空。	请确保控制器节点的configs字段已正确填写，且格式符合规范。
AgentArts.02301005	当前传入的工作流类型不被系统支持。	多智能体不支持的工作流类型。	请检查工作流类型配置，确保使用的是系统预定义的合法工作流类型。
AgentArts.02301006	指定的意图工作流版本在系统中不存在，无法加载对应的工作流定义。	意图工作流配置无效。	请确认工作流ID和版本ID的配置是否正确，并确保该版本已成功发布到系统中。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02301007	当前意图工作流的输入或输出节点未正确配置，导致无法获取有效字段信息。	意图工作流的输入或输出不能为空。	请检查意图工作流的输入和输出节点配置，确保其包含合法且非空的字段定义。
AgentArts.02301008	当前控制器中配置了多个意图类型的工作流节点。	存在重复的意图节点。	请确保控制器中仅配置一个意图类型的工作流节点，避免重复定义。
AgentArts.02301009	意图工作流中的query参数类型必须为字符串且为必填项，当前配置不符合要求。	意图工作流参数查询无效。	请检查意图工作流的输入参数配置，确保query字段类型为string且标记为required。
AgentArts.02301010	意图工作流中的minScore参数类型必须为数字且不能为必填项，当前配置不符合要求。	意图工作流输入参数minScore无效。	请检查意图工作流的输入参数配置，确保minScore字段类型为number且未标记为required。
AgentArts.02301011	意图工作流中缺少必需的输入参数，无法完成校验。	意图工作流缺少必需的输入参数。	请确保意图工作流的输入参数中包含必须的输入参数。
AgentArts.02301012	意图工作流的输出参数intent_id类型必须为整数。	意图工作流输出参数intent_id无效。	请检查意图工作流的输出参数配置，确保intent_id字段类型为integer。
AgentArts.02301013	意图工作流中缺少必需的输出参数intent_id。	意图工作流缺少输出参数。	请确保意图工作流的输出节点中包含intent_id参数，其类型为integer。
AgentArts.02301014	意图工作流输入参数messages中的子参数类型必须为字符串且为必填项。	意图工作流输入参数messages的子参数无效。	请检查messages参数中的子参数，确保其类型为string且标记为required。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02301015	意图工作流中messages参数缺少必需的子属性，无法完成消息结构校验。	意图工作流message参数缺少属性。	请确保messages参数中包含role和content两个必需的子属性，并正确配置其类型和必填状态。
AgentArts.02301016	意图工作流输入参数intents中的子参数类型必须为字符串且为必填项。	意图工作流参数intents的子参数无效。	请检查intents参数中的子参数，确保其类型为string且标记为required。
AgentArts.02301017	意图工作流中intents参数缺少必需的属性，无法正确解析意图信息。	意图工作流intents参数缺少属性。	请确保intents参数中包含id和name两个必需属性，并正确配置其类型为string且为required。
AgentArts.02301018	意图工作流中包含了不允许的交互类型节点。	意图工作流不应包含交互节点。	请检查意图工作流中的节点类型，移除所有交互类节点。
AgentArts.02301019	意图工作流中引用的子工作流版本不存在。	未找到子工作流版本。	请检查子工作流的配置，确认所引用的id和version_id正确且已发布对应版本。
AgentArts.02301020	当前多智能体类型为controller，与已有智能体名称冲突。	多智能体名称冲突。	请检查智能体名称配置，确保controller类型的多智能体名称具有唯一性且不与已有智能体重复。
AgentArts.02301021	当前controller配置中存在重复绑定的子智能体。	绑定的子智能体重复。	请检查controller节点的agents配置，确保每个子智能体的id唯一且不重复。
AgentArts.02301022	当前controller节点绑定的子智能体数量已超过系统支持的最大限制。	绑定的子智能体达到最大限制。	请减少绑定子智能体数量。
AgentArts.02301023	指定的智能体ID对应的版本不存在，无法完成子智能体绑定操作。	当前子智能体版本不存在。	请确认所绑定的子智能体ID和版本号是否正确，并确保该版本已发布或存在。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02301024	多智能体导出过程中发生异常。	多智能体导出失败。	请检查多智能体的DSL配置是否完整有效。
AgentArts.02301025	当前controller节点绑定的子智能体嵌套深度已超过系统支持的最大限制。	绑定子智能体达到最大深度限制。	请减少绑定子智能体深度。
AgentArts.02301026	当前多智能体全局配置输入参数默认值超出最大长度限制。	多智能体全局配置输入参数默认值达到最大长度限制。	请减少输入参数默认值的长度。
AgentArts.02301027	当前多智能体全局变量默认值超出最大长度限制。	多智能体全局变量默认值达到最大长度限制。	请减少全局变量默认值的长度。
AgentArts.02301028	多智能体模型配置或意图识别为空。	多智能体模型配置或意图识别为空。	请配置模型或意图识别工作流。
AgentArts.02301029	控制器只能挂载一个Plan&Execute模式的子Agent。	控制器挂载单智能体失败。	请在控制器下只挂载一个Plan&Execute模式的子Agent。
AgentArts.02301030	Plan&Execute模式的子Agent只能挂在顶层控制器。	绑定子控制器失败。	请将Plan&Execute模式的子Agent挂在顶层控制器。
AgentArts.02301031	多智能体不能绑定自身。	绑定子多智能体失败。	绑定其他多智能体后重试。

版权所有 © 华为云计算技术有限公司

表 12-6 错误码（后台报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103100	控制器在调用底层原子技能尝试从用户输入或上下文提取参数时发生异常，通常是由于模型未按规范要求输出或变量映射失败。	参数提取原子能力调用失败。	请检查参数提取器的配置，并核对模型是否能够准确识别当前上下文中的关键槽位信息。
AgentArts.103101	在执行参数补全或修正逻辑时，控制器无法正确构造修正指令或接收到模型反馈的错误信息。	参数修正原子能力调用失败。	请验证“参数修正”环节的提示词引导词，确保模型在参数缺失时能正确触发反问或补充逻辑。
AgentArts.103102	控制器在管理执行流时收到了预料之外的中断信号，可能源于用户的主动干预或系统级的高优先级强制抢占。	控制器执行中断。	请检查当前控制器的并发策略，并确认是否存在其他模块触发了非法的中断指令。
AgentArts.103103	由于全局意图频繁触发，控制器执行的跳转次数超过了系统设定的阈值，旨在防止陷入意图切换死循环。	全局意图跳转超限。	请优化全局意图的触发阈值或关键词权重，确保意图识别的精准度，避免多个意图配置过于接近。
AgentArts.103104	控制器在意图检测模块执行期间发生内部崩溃或返回了非预期的分类结果。	意图识别错误。	请检查意图分类器所依赖的分类模型及样本库，确保意图定义列表不为空且格式合法。
AgentArts.103105	意图识别成功，但在调度该意图所绑定的对应工作流执行时发生故障。	意图关联工作流执行错误。	请进入该意图对应的工作流画布进行调试，确认起始节点的输入参数与控制器传递的意图上下文兼容。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103200	尝试启动一个已经在运行中的多智能体协作任务。系统禁止对同一个Runner实例进行重复初始化。	多智能体运行器已启动。	请检查调用逻辑，确保在发起新任务前旧任务已结束，或直接通过状态查询接口获取当前进度。
AgentArts.103201	在多智能体分发任务或进行路由跳转时，系统无法在当前群组中找到标识的Agent成员。	目标成员不存在。	请核对协作网络中的Agent注册名称，确保分发逻辑或路由配置中指向的成员ID正确无误。
AgentArts.103202	多智能体流转进入了人工干预节点或需要外部反馈的环节，当前消息处理进程已挂起，等待用户输入。	消息处理挂起。	请通过人工审核接口或对话框提供所需的反馈数据，以恢复智能体协作流。
AgentArts.103203	多智能体在进行跨Agent通讯或推理过程中收到了强制中断信号，导致当前的协作事务链条断裂。	协作处理被中断。	请检查是否存在节点超时或系统资源抢占导致的任务强制释放。
AgentArts.103204	尝试对一个已经停止、崩溃或处于闲置状态的运行器发送控制指令（如挂起、终止）。	运行器未在运行。	请先验证运行器的实时状态，确保指令发送对象处于活跃的运行周期内。
AgentArts.103205	在运行器完成初始化并正式启动之前，尝试调用其执行接口，导致非法操作异常。	运行器尚未启动。	请确保在执行run()或next()方法前，已正确完成了运行器的配置与start()调用。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103232	在尝试销毁运行空间时，无法正常关闭Agent间通讯的消息总线。	消息队列关闭失败。	请确认是否存在残留的死锁进程正在持续写入队列，必要时请手动重置消息代理服务。
AgentArts.103233	控制器尝试终止正在运行的协作任务，但由于某个Agent处于不可中断的阻塞调用中，导致任务停止指令超时或失效。	停止运行任务失败。	请检查Agent工具调用是否设置了合理的超时时间，确保所有长时间运行的节点都能响应中断信号。
AgentArts.103234	运行空间检测到所有Agent已进入空闲状态，尝试自动关停环境以节省资源时发生异常。	空闲停止执行失败。	检查运行空间的垃圾回收（GC）逻辑，确保没有非托管资源在空闲期阻碍销毁进程。
AgentArts.103235	在每次迭代后校验“终止条件”时，由于上下文变量解析错误或逻辑表达式非法，导致无法判断协作是否该结束。	停止条件校验失败。	请核对协作配置中的终止表达式语法，并确保表达式引用的变量在上下文中确实存在。
AgentArts.103236	控制器在执行特定的stop_when触发器逻辑时发生错误。与一般条件校验不同，这通常涉及更复杂的代码片段执行失败。	执行Stop-When逻辑错误。	请调试stop_when的自定义函数逻辑，确保其在各种边界场景（如空响应或异常返回）下均能稳定运行。
AgentArts.103261	这通常是因为该成员的提示词模板损坏或输入参数不符合其预设规范。	成员消息处理失败。	请检查该成员的独立配置，并核对前置Agent传递给它的上下文数据是否完整。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.103262	成员在执行推理任务的生命周期中抛出未捕获异常。这可能涉及模型接口调用失败、插件执行崩溃或内存溢出。	消息执行期间异常。	查看该成员的运行日志以定位具体崩溃位置，确保其依赖的外部服务和插件环境正常。
AgentArts.103291	尝试对消息进行操作（如更新状态或强制中止）时，发现该消息已离开处理队列，或处于非活动状态（如已完成、已丢弃）。	消息未处于处理状态。	请检查消息ID是否过期，并确认任务调度器是否已意外移除了该消息。
AgentArts.103300	尝试对一个已处于Active状态的群组实例执行start()命令。	智能体群组已启动。	在发起新一轮群组协作前，请先确认旧任务是否已通过stop()释放资源，或直接使用当前活跃实例。
AgentArts.103301	在群组组件未完成初始化或未进入运行状态时尝试调用其协作接口。	群组未初始化。	请检查群组定义的IR结构，确保在调用执行方法前已成功执行了init()过程。
AgentArts.103302	群组在协调多个Agent间的流转逻辑（如投票、轮询或分发）时发生系统级错误。	群组执行异常。	重点排查群组的选择器（Selector）逻辑，确保其能根据当前上下文正确计算出下一个执行的Agent。
AgentArts.103303	向群组添加Agent成员时失败。	成员注册失败。	请检查Agent成员的唯一性，确保没有重复将同一个Agent实例以不同角色注册到同一个群组中。

版权所有 © 华为云计算技术有限公司

组件错误码

表 12-7 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401001	提交的工具信息格式不符合规范。	工具信息格式错误。	请检查工具信息格式并修改。
AgentArts.02401002	工具信息中的输入参数格式有误。	入参格式错误。	请检查入参定义并修改。
AgentArts.02401003	工具信息中的输出参数格式有误。	输出参数格式错误。	请检查出参定义并修改。
AgentArts.02401004	当前工具名称已被占用。	工具名称已存在。	请重新命名工具。
AgentArts.02401005	查询的工具在系统中不存在。	工具不存在。	请确认工具ID或名称是否正确。
AgentArts.02401006	工作流校验组件未能通过验证。	工作流校验失败。	请检查工作流配置及组件状态。
AgentArts.02401007	当前用户没有修改该工具的权限。	无修改权限。	请联系管理员检查您的权限配置。
AgentArts.02401008	当前用户没有删除该工具的权限。	无删除权限。	请联系管理员检查您的权限配置。
AgentArts.02401009	在导入插件过程中发生错误。	插件导入失败。	请检查插件包是否完整或重试。
AgentArts.02401010	在导出插件过程中发生错误。	插件导出失败。	请检查网络状态或重试。
AgentArts.02401011	上传或处理的文件没有任何内容。	文件内容为空。	请检查文件内容是否正确。
AgentArts.02401012	上传的文件大小超出了系统限制。	文件大小超限。	请压缩文件或上传更小的文件。
AgentArts.02401013	上传的文件扩展名或类型不在支持列表中。	文件类型不支持。	请上传支持的文件格式。
AgentArts.02401014	尝试创建的依赖项已经存在。	依赖项已存在。	请勿重复创建，或检查依赖项名称。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401015	系统中不存在指定的依赖项。	依赖项未找到。	请检查依赖项ID或名称。
AgentArts.02401016	当前用户缺少FunctionGraph相关权限。	缺少FunctionGraph权限。	请联系管理员检查是否拥有FunctionGraph权限。
AgentArts.02401017	创建依赖项过程中发生系统错误。	创建依赖项失败。	请检查输入信息或稍后重试。
AgentArts.02401018	输入的数据无法解析或格式不正确。	数据格式错误。	请检查数据格式是否符合规范。
AgentArts.02401019	指定的名称已被占用。	名称已存在。	请更换其他名称。
AgentArts.02401020	指定的参数范围无效。	范围无效。	请检查参数范围设置并重试。
AgentArts.02401021	调用函数图接口时发生异常。	请求函数图失败。	请检查网络连接或稍后重试。
AgentArts.02401022	更新依赖项信息时发生错误。	更新依赖项失败。	请检查输入信息或稍后重试。
AgentArts.02401023	指定的依赖项正被1个函数使用，无法直接操作。	依赖项被占用。	请先删除关联的函数或解除关联关系。
AgentArts.02401024	执行删除操作时发生错误。	删除依赖项失败。	请确认依赖项未被占用并重试。
AgentArts.02401025	函数的内存大小设置不能超过2048MB。	内存设置超限。	请调整内存大小至2048MB以内。
AgentArts.02401026	在解析文本编码时发生UTF-8异常。	UTF-8解析异常。	请确保输入内容为合法的UTF-8编码。
AgentArts.02401027	当前环境不支持Python运行时。	不支持Python。	请选择Node.js作为运行时环境。
AgentArts.02401028	无法获取函数的代码内容。	获取代码失败。	请检查函数是否存在或代码是否已上传。
AgentArts.02401029	Java11函数需要codeType和codeFileName参数，但未提供或不受支持。	Java11参数缺失。	请补充相关参数并确认环境是否支持Java11。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401030	请求FunctionGraph接口时发生错误。	请求FunctionGraph接口失败。	请稍后重试或联系技术支持。
AgentArts.02401033	所选的运行时功能在当前环境不受支持。	运行时不支持。	请更换其他运行时环境。
AgentArts.02401034	更新函数代码时发生错误。	代码更新失败。	请检查代码包格式或稍后重试。
AgentArts.02401035	该函数名称已被其他函数占用。	函数名称已存在。	请使用唯一的函数名称。
AgentArts.02401036	无法验证工具信息。	检查工具信息失败。	请检查该节点的配置信息。
AgentArts.02401037	插件权限验证失败。	插件权限错误。	请确认您是否有权访问该插件。
AgentArts.02401039	系统无法解析上传的文件内容。	文件解析失败。	请检查文件是否损坏或格式是否正确。
AgentArts.02401040	文件的内部格式不符合要求。	文件格式错误。	请参照文档上传标准格式的文件。
AgentArts.02401041	超出最大导出数量限制。	导出数量超限。	请减少导出数量或分批导出。
AgentArts.02401042	指定的版本信息不存在。	版本不存在。	请检查ID和版本号是否正确。
AgentArts.02401043	系统无法获取OBS临时访问URL。	获取OBS URL失败。	请检查OBS配置或网络状态。
AgentArts.02401044	设置的Visibility值无效。	可见性值无效。	请设置合法的可见性值。
AgentArts.02401045	系统无法生成OpenAPI文档。	OpenAPI生成失败。	请检查接口定义或配置信息。
AgentArts.02401046	试图发布一个尚未进行测试的组件。	发布未经测试组件。	请先完成组件测试后再发布。
AgentArts.02401047	插件身份验证未能通过。	插件鉴权失败。	请检查鉴权信息是否正确。
AgentArts.02401048	未找到该插件的有效凭证。	插件凭证缺失。	请为插件配置或重新生成凭证。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401049	该插件已存在凭证，无法重复创建。	凭证已存在。	请直接使用现有凭证或修改配置。
AgentArts.02401050	工具的中文名称已被占用。	中文名已存在。	请更换工具中文名称。
AgentArts.02401051	工具的英文名称已被占用。	英文名已存在。	请更换工具英文名称。
AgentArts.02401052	请求的MCP服务未找到。	MCP服务不存在。	请检查MCP服务ID或状态。
AgentArts.02401053	无法更新记忆参数信息。	更新记忆参数失败。	请检查参数配置或稍后重试。
AgentArts.02401054	无法读取记忆参数信息。	获取记忆参数失败。	请检查数据是否存在或系统状态。
AgentArts.02401055	无法获取文件的输入流。	获取文件流失败。	请检查文件是否可读或路径是否正确。
AgentArts.02401056	文件内容解析过程中发生错误。	解析文件失败。	请检查文件完整性或格式。
AgentArts.02401057	指定路径下未找到文件。	文件不存在。	请检查文件路径是否正确。
AgentArts.02401058	无法解析提供的OpenAPI文件。	OpenAPI解析失败。	请检查OpenAPI文件格式是否符合规范。
AgentArts.02401059	无法从URI中提取Path信息。	路径获取失败。	请检查URI格式是否正确。
AgentArts.02401060	MCP服务发生相关内部错误。	MCP服务错误。	请检查MCP服务状态或日志。
AgentArts.02401061	无法构建文件输入流。	构建文件流失败。	请检查文件源是否可用。
AgentArts.02401062	图片进行Base64编码时发生错误。	Base64编码失败。	请检查图片格式或大小。
AgentArts.02401063	OCR服务无法识别图片内容。	OCR识别失败。	请保证图片清晰度或检查OCR服务状态。
AgentArts.02401064	提供的urls参数不符合要求。	URL参数校验失败。	请检查URL格式及有效性。
AgentArts.02401065	表中未找到指定的API。	API未找到	请检查API名称是否正确。
AgentArts.02401066	适配IR失败，指定的MCP服务不存在。	适配IR失败(服务缺失)。	请检查MCP服务ID是否正确。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401067	适配IR失败，MCP服务中的工具不存在。	适配IR失败(工具缺失)。	请检查工具名称和ID对应关系。
AgentArts.02401068	MCP服务名称已被占用。	MCP名称已存在。	请更换MCP服务名称。
AgentArts.02401069	对应ID的MCP服务不存在。	MCP服务不存在。	请检查MCP服务ID。
AgentArts.02401070	该MCP服务的配置信息已存在。	MCP配置已存在。	请直接修改现有配置。
AgentArts.02401071	未找到相关的MCP配置信息。	MCP配置不存在。	请先创建配置。
AgentArts.02401072	操作的对象MCP服务不存在。	MCP服务缺失。	请检查服务是否存在。
AgentArts.02401073	MCP服务URL参数缺失。	URL不能为空。	请填写MCP服务的URL地址。
AgentArts.02401074	当前状态不允许更新MCP服务URL。	禁止更新URL。	如需修改，请重新创建服务。
AgentArts.02401075	MCP服务名称为必填项。	MCP名称不能为空。	请输入MCP服务名称。
AgentArts.02401076	MCP服务英文名称为必填项。	英文名不能为空。	请输入MCP服务英文名称。
AgentArts.02401077	MCP服务描述信息为必填项。	描述不能为空。	请输入服务描述。
AgentArts.02401078	MCP服务的可见性参数设置错误。	可见性设置无效。	请检查可见性配置。
AgentArts.02401079	该服务由他人创建，您权限不足。	无操作权限。	请联系服务创建者或管理员。
AgentArts.02401080	MCP配置错误，未提供认证密钥。	缺少认证密钥。	请在配置中补充认证密钥。
AgentArts.02401081	无法配置自定义MCP服务。	自定义配置失败。	请检查自定义配置参数。
AgentArts.02401082	无法配置公共MCP服务。	公共配置失败。	请检查公共服务配置权限或参数。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401083	MCP配置错误，字段不能为空。	配置字段为空。	请补充缺失的配置字段。
AgentArts.02401084	MCP配置错误，缺少必要字段。	缺少必要字段。	请补全所有必填字段。
AgentArts.02401085	获取工具列表失败。	获取MCP工具失败。	请根据错误码排查问题。
AgentArts.02401086	插件不存在或当前无权限使用。	插件无法使用。	请检查插件ID及您的权限。
AgentArts.02401087	检测到非法的API调用请求。	非法API调用。	请停止非法调用并检查请求来源。
AgentArts.02401088	请求中的参数无效。	参数无效。	请检查该参数的值是否正确。
AgentArts.02401089	远程调用服务失败。	远程调用失败。	请检查远程服务状态或网络连接。
AgentArts.02401090	用户无权访问请求的资源。	无资源访问权。	请申请相应的资源访问权限。
AgentArts.02401091	调用时缺少必要的输入参数。	缺少输入参数。	请补全所有必填参数。
AgentArts.02401092	该服务正在创建中，无法执行删除操作。	无法删除创建中服务。	请等待创建完成后再试。
AgentArts.02401093	该服务正在删除中，请勿重复操作。	无法删除“删除中”的服务。	请等待删除过程结束。
AgentArts.02401094	从FunctionGraph删除服务失败。	FunctionGraph删除失败。	请检查FunctionGraph服务状态或联系管理员。
AgentArts.02401095	操作的MCP服务在系统中不存在。	MCP服务不存在。	请检查服务标识是否正确。
AgentArts.02401096	MCP服务配置不是合法的JSON格式。	配置格式错误。	请检查并修正配置的JSON格式。
AgentArts.02401097	修改对象属性时发生错误。	属性修改失败。	请检查输入值或系统状态。
AgentArts.02401098	尝试创建的MCP服务名称已经存在。	MCP服务名称已存在。	请检查是否重复创建。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401099	从FunctionGraph创建MCP服务未成功。	FunctionGraph创建服务失败。	请检查FunctionGraph配置及日志。
AgentArts.02401100	等待应用程序更新超时。	更新等待超时。	请稍后重试或检查后台任务。
AgentArts.02401101	查询MCP的工具列表时发生错误。	查询工具列表错误。	请检查MCP服务健康状态。
AgentArts.02401102	MCP服务配置内容格式不符合要求。	配置格式不正确。	请参照文档修正配置格式。
AgentArts.02401103	调用MCP的工具时发生错误。	工具调用错误。	请检查工具参数及服务日志。
AgentArts.02401104	调用FunctionGraph服务接口失败。	FunctionGraph接口调用失败。	请检查网络或FunctionGraph服务状态。
AgentArts.02401105	部署的MCP服务总数超过了最大限制。	部署数量超限。	请删除闲置服务或申请扩容。
AgentArts.02401106	查询MCP的工具列表错误。	查询工具列表失败。	请稍后再试或联系管理员。
AgentArts.02401107	当前用户没有使用MCP服务的权限。	无MCP服务权限。	请联系管理员申请权限。
AgentArts.02401108	提供的镜像格式不合法。	非法镜像格式。	请检查镜像地址及格式。
AgentArts.02401109	镜像大小超过了系统允许的限制。	镜像大小超限。	请使用更小的镜像或优化镜像层。
AgentArts.02401110	操作授权过程失败。	授权失败。	请检查授权配置或重试。
AgentArts.02401111	取消授权操作失败。	取消授权失败。	请稍后重试或联系管理员。
AgentArts.02401112	无法查询授权信息。	查询授权失败。	请检查网络或系统状态。
AgentArts.02401113	IAM创建或查询机构失败。	IAM操作失败。	请检查IAM权限及配置。
AgentArts.02401114	系统发生未预期的内部错误。	内部错误。	请联系管理员进行排查。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401115	工具数量超过了允许的限制。	工具数量超限。	请删除多余工具或申请扩容。
AgentArts.02401116	解析请求头(Headers)时发生异常。	Headers解析异常。	请检查请求头格式是否规范。
AgentArts.02401117	插件展示名称已存在。	插件展示名称重复。	请更换插件展示名称。
AgentArts.02401118	插件名称已存在。	插件名称重复。	请更换插件名称。
AgentArts.02401119	MCP名称格式错误：不能以_、 - 或数字开头，长度需2-64字符。	名称格式无效。	请检查名称格式，允许的内容包括： • 中文、英文、数字 • 符号：下划线、连字符、括号、点号、空格。
AgentArts.02401121	插件认证配置不可用或已失效。	认证配置不可用。	请检查并更新认证配置。
AgentArts.02401122	插件鉴权请求超时。	鉴权超时。	请检查网络状况或稍后重试。
AgentArts.02401123	插件鉴权时发生网络错误。	鉴权网络错误。	请检查网络连接。
AgentArts.02401124	此函数已被本插件绑定。	函数已绑定。	请勿重复绑定。
AgentArts.02401125	指定操作的函数不存在。	函数不存在。	请检查函数ID是否正确。
AgentArts.02401126	删除函数时发生异常。	删除函数错误。	请稍后重试。
AgentArts.02401127	函数已绑定插件，无法删除。	删除函数失败(已绑定)。	请先解绑插件再删除函数。
AgentArts.02401128	插件类型一旦创建禁止修改。	禁止修改插件类型。	如需改变类型，请重新创建插件。
AgentArts.02401129	未找到指定的函数模板。	函数模板不存在。	请检查模板ID或使用其他模板。
AgentArts.02401130	插件ID已存在。	ID已存在。	请修改插件英文名以生成新ID。
AgentArts.02401131	不支持复制API类型的插件。	禁止复制API插件。	请手动创建新的API插件。
AgentArts.02401132	Score值无效，必须在1-10之间。	分数范围错误。	请调整Score值到合法范围内。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401138	工具数量超出系统限制。	工具数量超限。	请清理旧工具或联系管理员。
AgentArts.02401139	异常流程的重试次数必须在0-3之间。	重试次数无效。	请调整重试次数。
AgentArts.02401140	无权限查看该ID对应的插件。	无权查看插件。	请联系所有者开通权限。
AgentArts.02401141	插件和工具的输入参数名重复。	参数名重复。	请修改参数名称以避免冲突。
AgentArts.02401142	系统不支持对应的MCP安装方式。	安装方式不支持。	请选择支持的安装方式。
AgentArts.02401143	未指定MCP安装方式。	安装方式为空。	请补充安装方式信息。
AgentArts.02401144	第三方MCP仅支持SSE或streamable_http部署方式。	部署方式受限。	请调整部署方式设置。
AgentArts.02401145	插件IAM认证URL参数配置错误。	IAM认证URL错误。	请检查认证URL配置。
AgentArts.02401146	MCP配置格式验证失败，JSON结构错误或缺少必填字段。	MCP格式校验失败。	请检查配置JSON。
AgentArts.02401147	FunctionGraph资源配额超出，无法创建服务。	FunctionGraph配额不足。	请检查配额限制或申请增加配额。
AgentArts.02401148	FunctionGraph服务执行出错。	FunctionGraph执行错误。	请稍后再试或联系管理员。
AgentArts.02401149	与MCP服务连接成功，但获取工具列表时服务端返回异常。	获取MCP工具错误。	请查看目标MCP服务的运行日志以排查应用层错误。
AgentArts.02401150	无法解析主机地址或建立基础网络连接。	网络连接失败。	请检查服务域名拼写是否正确，以及DNS解析是否正常。
AgentArts.02401151	目标服务地址受到访问限制。	地址受限。	请检查访问策略。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401152	JSON格式不正确或类型不匹配，无法将buildOutputList解析为工具输出列表。	输出列表解析失败。	请确保JSON符合List<IsOutputList>格式，字段名称和类型匹配。
AgentArts.02401153	字符串格式不正确或字段类型不匹配，无法解析为对象。	对象解析失败。	请确保字符串是有效的JSON格式。
AgentArts.02401154	RequestInfo字段的JSON内容格式错误或不完整。	RequestInfo解析失败。	请确保requestInfo符合ToolInfo或ToolRequestInfo定义。
AgentArts.02401155	实际URL与预期的URL前缀不匹配。	URL前缀不匹配。	请检查并修正URL前缀设置。
AgentArts.02401156	IntfType参数格式不正确，无法解析为List<ToolIntfType> 类型。可能是JSON格式错误、字段缺失或类型不匹配	无法将buildIntfType解析为ToolInfo。	请检查intfType参数的JSON格式是否符合预期，确保其包含有效的ToolIntfType对象列表，并且每个对象至少包含toolId和intfType字段。此外，确保JSON字符串未被截断或包含非法字符。
AgentArts.02401157	输入的JSON字符串格式不正确，无法解析为ToolTestStatus对象列表。	未能将buildTestStatus解析为ToolInfo。	请检查输入的JSON字符串是否符合ToolTestStatus对象列表的格式要求，确保字段名称和类型与类定义一致。
AgentArts.02401158	OutputSchema的格式不符合List<ToolOutputSchema>的JSON结构	无法将buildOutputSchema解析为ToolInfo。	检查outputSchema的输入内容是否符合预期的JSON格式。
AgentArts.02401159	输入的JSON字符串格式错误或内容无法映射到List<IsInputList> 类型，导致解析失败。	无法将buildInputList解析为ToolInfo。	请确保传入的JSON字符串符合标准格式，并且其内容与IsInputList类的字段匹配。建议在调用buildInputList方法前增加输入校验逻辑，或使用McpJsonUtils.fromJson方法进行更健壮的解析。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401160	输入的JSON字符串格式不正确，无法被解析为ToolInputSchema列表。可能是JSON语法错误、字段缺失或类型不匹配。	无法将filterToolInputSchema解析为ToolInfo。	请检查传入的inputSchema字符串是否为合法的JSON格式，并确保其包含符合ToolInputSchema类定义的字段（如toolId和inputSchema）。 建议使用JSON校验工具进行验证，并在开发环境中添加异常捕获和日志输出以辅助调试。
AgentArts.02401161	系统与目标MCP服务建立连接超时，可能原因：网络不通、代理被拦截或路径错误。	MCP连接超时。	请检查： • 防火墙/安全组是否放通； • 如需代理，请确认白名单及认证配置； • URL路径是否正确。
AgentArts.02401162	目标服务器拒绝连接，通常是因为服务进程未启动或端口配置错误。	连接被拒绝。	请检查目标主机服务进程是否运行，以及端口是否开启监听。
AgentArts.02401163	MCP服务目标地址路径错误(404 Not Found)。	路径错误(404)。	请检查服务路径是否正确。
AgentArts.02401164	MCP服务认证失败，无权访问。	认证失败(401/403)。	请检查认证凭证是否有效。
AgentArts.02401165	创建MCP API网关失败。	APIG创建失败。	请稍后重试或联系管理员。
AgentArts.02401166	调用函数依赖包查询接口时发生异常。	查询依赖包失败。	请稍后重试或检查相关服务。
AgentArts.02401167	调用函数依赖包版本查询接口时发生异常。	查询依赖版本失败。	请稍后重试或检查相关服务。
AgentArts.02401168	操作失败，因为选择的依赖包版本不存在或已被删除。	依赖包版本不存在。	请确认选择的依赖包版本是否正确且存在。
AgentArts.02401169	导入OpenAPI文件为空。	导入OpenAPI文件为空。	请确保上传的OpenAPI文件非空。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401170	导入OpenAPI文件Server为空。	导入OpenAPI文件中Server为空。	请确保上传的OpenAPI文件Server非空。
AgentArts.02401171	导入OpenAPI文件Server中URL为空。	导入OpenAPI文件中Server中URL为空。	请确保上传的OpenAPI文件Server中URL非空。
AgentArts.02401172	解析过程中失败，可能是OpenApi文件中某些参数为空解析失败。	解析OpenApi文件失败。	请检查Parameters、RequestBody、Response等参数。
AgentArts.02401173	系统无法解析目标MCP服务的域名，通常是因为域名拼写错误或DNS不可达。	无法解析主机地址。	请仔细检查服务配置中URL的域名部分拼写是否正确。
AgentArts.02401174	MCP认证配置不可用或已失效。	认证配置不可用。	请检查并更新认证配置。
AgentArts.02401175	调用函数详情查询接口时发生异常。	获取函数失败。	请确认选择的函数是否正确且存在。
AgentArts.02401176	获取函数的metadata信息失败。	获取函数metadata失败	请确认选择的函数是否正确且存在。
AgentArts.02401177	OAUTH鉴权信息配置越界。	OAUTH鉴权信息配置错误。	请确认OAUTH鉴权信息长度在1~255。
AgentArts.02401178	APIKEY鉴权信息配置越界。	APIKEY鉴权信息配置错误。	请确认APIKEY鉴权信息长度在164以内。
AgentArts.02401179	获取预置插件免费额度失败。	获取预置插件免费额度失败。	请确认该插件是否有免费额度，检查redis是否正常。
AgentArts.02401180	插件更新发布失败，插件未发布或未找到已发布版本。	插件更新发布失败。	在插件详情页检查是否已发布，并更新智能体中插件版本。
AgentArts.02401181	APIkey的名称仅支持输入字母、数字、下划线、短横线。	APIkey的名称校验错误。	在插件配置鉴权中，检查APIkey类型的名称输入是否规范。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02401182	参数实际输入参数类型与输出参数类型不匹配。	函数输入参数类型和输出参数类型不匹配。	请检查并修正函数输入参数类型和输出参数类型。
AgentArts.02401183	应用的运行时信息为空，导致空指针异常。	应用的运行时信息缺失。	请检查应用的运行时信息是否正确。
AgentArts.02401184	内部服务访问失败。	内部服务访问失败。	内部服务访问失败，请联系运维人员，定位排查。
AgentArts.02401185	创建MCP运行时失败。	创建MCP运行时失败。	创建MCP运行时失败，请稍后重试，或联系运维人员，定位排查。
AgentArts.02401186	删除运行时失败。	删除MCP依赖的运行时失败。	删除MCP运行时失败，请稍后重试，或手动删除运行时后再次执行MCP删除。

模型错误码

表 12-8 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501001	未找到指定的模型。	缺少模型配置。	请检查模型配置信息或使用其他模型。
AgentArts.02501002	供应商名称已存在。	供应商名称重复。	请修改参数名称以避免冲突。
AgentArts.02501003	模型名称已存在。	模型服务名称重复。	请修改参数名称以避免冲突。
AgentArts.02501004	路由策略名称已存在。	路由策略名称重复。	请修改参数名称以避免冲突。
AgentArts.02501005	模型调用异常。	模型调用异常。	请检查模型配置是否正确。
AgentArts.02501006	模型调用地址不满足要求。	模型调用地址不满足要求。	请更换调用模型地址。
AgentArts.02501007	无权限调用该模型。	模型调用被安全风控拦截。	当前请求的项目ID与配置的项目ID不匹配，导致无权限执行操作。
AgentArts.02501008	模型服务提供方认证信息无效。	模型服务提供方认证信息无效。	请检查模型服务配置的认证信息。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501009	模型提供方不存在或不属于该项目。	提供方不存在或不属于该项目。	请检查模型服务提供方配置信息是否正确或您的权限。
AgentArts.02501010	Provider_id和metadata_id同时为空。	Provider_id和metadata_id不能同时为空。	请检查参数provider_id和metadata_id。
AgentArts.02501011	Provider_id和auth_id同时为空。	Provider_id 和auth_id不能同时为空。	请检查参数provider_id和auth_id。
AgentArts.02501012	模型元数据信息为空。	模型的元数据不存在。	请检查metadata_id参数是否正确。
AgentArts.02501013	模型服务信息为空。	模型服务不存在。	请检查模型id参数是否正确。
AgentArts.02501014	无权操作该模型服务。	用户无权操作该模型服务。	请检查projectId、workspaceId及您的权限。
AgentArts.02501015	模型认证信息为空。	模型认证数据不存在。	请检查模型认证信息配置及配置id是否正确。
AgentArts.02501016	无权限删除认证数据。	用户无权限删除认证数据。	请检查projectId、workspaceId及您的权限。
AgentArts.02501017	认证元数据为空。	认证元数据不存在。	请检查providerId是否正确。
AgentArts.02501018	无权限操作该提供方。	用户无权限操作该提供方。	请检查projectId、workspaceId及您的权限。
AgentArts.02501019	\	模型服务发布状态已完成，无需重复操作。	\
AgentArts.02501020	该模型服务已经上线/下线，无法修改。	不允许修改该模型服务。	请检查模型服务状态，确保其处于可修改状态。
AgentArts.02501021	请求参数中存在重复的模型ID。	请求体无效，存在重复的模型。	请检查请求参数的模型ID是否重复。
AgentArts.02501022	请求参数中模型ID为空。	请求体无效，模型为空。	请检查请求参数的模型ID是否为空。
AgentArts.02501023	部分模型不存在。	请求体无效，部分模型不存在。	请检查请求参数的模型ID是否真实有效。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501024	模型服务认证配置不存在。	模型服务认证配置不存在。	请检查模型是否存在认证配置。
AgentArts.02501025	模型服务已发布，无法删除。	模型服务已发布，无法删除。	请先下线模型，再删除。
AgentArts.02501026	路由策略不存在。	路由策略不存在。	请检查路由策略配置是否存在。
AgentArts.02501027	WorkspaceId为空。	WorkspaceId为空，请在查询参数或请求头中提供。	请在查询参数或请求头中提供workspaceId。
AgentArts.02501028	模型服务提供方认证信息无效，AK_SK认证配置不正确。	模型服务提供方认证信息无效，AK_SK认证配置不正确。	请检查模型AK_SK认证配置是否真实有效。
AgentArts.02501029	模型服务提供方认证信息无效，API_KEY认证配置不正确。	模型服务提供方认证信息无效，API_KEY认证配置不正确。	请检查模型API_KEY认证配置是否真实有效。
AgentArts.02501030	模型服务提供方认证信息无效，APP_CODE认证配置不正确。	模型服务提供方认证信息无效，APP_CODE认证配置不正确。	请检查模型APP_CODE认证配置是否真实有效。
AgentArts.02501031	请求参数无效，stream不是布尔值。	请求参数无效，stream不是布尔值。	请检查请求参数stream的类型是否是布尔类型。
AgentArts.02501032	模型策略信息为空。	获取模型策略失败。	请检查模型策略信息。
AgentArts.02501033	请求模型服务超时。	请求模型服务超时。	请检查模型URL配置是否正确。
AgentArts.02501034	请求模型服务内部错误。	请求模型服务内部错误。	请检查模型配置是否正确。
AgentArts.02501035	调用第三方模型服务失败，HTTP响应为空。	调用第三方模型服务失败，HTTP响应为空。	请检查第三方模型配置是否正确。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501036	调用第三方模型服务失败，模型响应体解析对象异常。	调用第三方模型服务失败，模型响应体解析对象异常。	请检查第三方模型配置是否正确。
AgentArts.02501037	模型服务不可用。	模型服务不可用。	请检查第三方模型配置是否正确。
AgentArts.02501038	\	模型策略下无可用模型。	\
AgentArts.02501039	请求次数过多。	请求被流量控制拦截，请稍后重试。	请稍后重试。
AgentArts.02501040	模型类型不符合要求。	模型类型不符合要求。	请检查模型类型是否正确。
AgentArts.02501041	无权限操作该认证元数据。	用户无权限操作该认证元数据。	请检查projectId、workspaceId及您的权限。
AgentArts.02501042	供应商存在接入模型服务，无法删除。	供应商存在接入模型服务，无法删除。	请先去除供应商已接入的模型，再进行删除。
AgentArts.02501043	供应商鉴权配置不存在。	供应商鉴权配置不存在，请先设置供应商鉴权配置。	请先设置供应商鉴权配置。
AgentArts.02501044	模型服务API地址不正确。	模型服务API地址不正确。	请检查模型服务API地址是否正确。
AgentArts.02501045	供应商鉴权配置不可用。	供应商鉴权配置不可用。	请检查供应商配置的鉴权信息是否有效。
AgentArts.02501046	无权限使用当前模型。	无权限使用当前模型。	请检查是否开通ModelArts预置服务或是否有权限使用当前模型。
AgentArts.02501047	账号费用不足。	账号费用不足。	账号费用不足，请处理账号费用。
AgentArts.02501048	模型服务不存在。	模型服务不存在。	请检查模型服务是否存在。
AgentArts.02501049	调用第三方模型服务失败，请查看请求响应查询详细原因。	调用第三方模型服务失败请查看请求响应查询详细原因。	请查看请求响应查询详细原因。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501051	模型免费Token额度已用完。	模型免费Token额度已用完。	请升级额度。
AgentArts.02501052	套餐配额用完。	套餐配额用完，禁止接入模型。	请检查配额限制或申请增加配额。
AgentArts.02501053	模型路由策略不正确。	模型路由策略不正确，请检查路由策略中的模型是否被删除或者取消发布。	请检查路由策略中的模型是否被删除或者取消发布。
AgentArts.02501054	第三方模型服务免费额度已用完。	第三方模型服务免费额度已用完。	请检查配额限制或申请增加配额。
AgentArts.02501055	该问题不符合安全合规要求，暂时无法回答。	抱歉，该问题不符合安全合规要求，暂时无法回答。	请检查问题是否合规。
AgentArts.02501057	模型接入URL无效。	模型接入URL无效。	请检查模型接入URL是否正常。
AgentArts.02501058	模型或者鉴权配置不可用。	模型或者鉴权配置不可用，请修改供应商鉴权配置或者修改模型服务。	请修改供应商鉴权配置或者修改模型服务。
AgentArts.02501059	自定义标签数量超出限制。	自定义标签数量超出限制。	请检查自定义标签数量。
AgentArts.02501060	无效的流控参数配置。	无效的流控参数配置。	请检查流控参数配置是否正常。
AgentArts.02501061	IAM和无认证的鉴权类型不允许删除。	该鉴权类型不支持删除鉴权配置。	修改供应商鉴权类型。
AgentArts.02501062	API key的创建未通过检查。	创建API key失败。	请检查当前账号所能创建的API key数量或者名称是否重复。
AgentArts.02501063	API key的鉴权与当前用户不匹配。	API key认证错误。	请检查该API key是否属于当前用户。
AgentArts.02501064	API key的鉴权与当前用户不匹配	API key名称重复。	请确认是否有当前API key的操作权限。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02501065	当前用户可创建API key数量已达到限制。	API key数量达到最大限制。	请删除当前用户的部分API key后重试。
AgentArts.02501066	当前环境不支持API key功能。	请求API key失败。	请检查当前账户的配置是否开启此项功能。
AgentArts.02501067	第三方模型服务参数校验拦截。	调用第三方模型参数错误。	根据错误信息调整请求参数。
AgentArts.02501068	同步模型服务providerUrl不能为空 。	同步模型服务providerUrl不能为空 。	请检查providerUrl是否有值 。
AgentArts.02501069	有路由策略引用该模型。	不能删除或者取消发布该模型，该模型已被路由策略引用。	请在路由策略中移除该模型。
AgentArts.02501070	免费模型并发调用限制	免费模型并发调用限制。	请避免并发调用模型，稍后重试。
AgentArts.02501071	任务ID不正确，或者是任务已失效。	视频生成任务不存在。	请检查任务ID是否正确，或者重新发起视频生成任务。
AgentArts.02501075	开通模型失败，可能由于账号欠费等原因，当前账号无法正常购买和使用按需计费云服务。	开通模型失败。	请查看当前账号状态，如已欠费，请您尽快充值。
AgentArts.00051037	模型服务不可用。	模型配置有问题。	• 检查账号是否欠费。 • 检查模型配置正确性：模型配置具体请参考**接入自定义模型**。接入模型最佳实践请参考**模型实践**。 • 检查网络连通性。

配置管理错误码

表 12-9 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02601001	分支数量超过了系统设定的限制。	分支数量超过了限制，请确认后重试。	减少分支数量至限制范围内后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02601002	意图包导入过程中出现系统错误。	导入意图包失败。	检查意图包文件完整性，确保文件格式正确后重试。
AgentArts.02601003	导入意图包的文件过大。	导入意图包时文件大小超出限制。	导入意图包的文件不大于5MB。
AgentArts.02601004	导出意图包时发生系统错误。	导出意图包失败。	检查系统权限和存储空间，确保导出路径正确后重试。
AgentArts.02601005	数据库账号信息已更新，需重新验证。	更新数据库账号信息后，请重新输入密码。	重新输入最新密码，确保信息准确无误。
AgentArts.02601006	资源数据库名称已存在。	重复的资源数据库名称，请检查数据库名称。	修改资源数据库名称，确保名称唯一性后重试。
AgentArts.02601007	请求的资源未找到。	访问的静态资源不存在。	检查资源路径和名字，确保资源存在后重试。
AgentArts.02601008	资源名称已存在。	资源操作失败，重复的资源名称，请检查资源名称。	修改资源名称，确保名称唯一性后重试。
AgentArts.02601009	Branch_id与现有意图冲突。	资源操作失败，branch_id与现有意图冲突。	检查branch_id和意图名称，确保无冲突后重试。
AgentArts.02601011	API模型名称已存在。	重复的资源嵌入或重排或OCR API模型名称，请检查API模型名称。	修改API模型名称，确保名称唯一性后重试。
AgentArts.02601012	模型名称不符合要求。	资源操作失败，模型名称无效。	检查模型名称格式，确保符合命名规范后重试。
AgentArts.02601013	当前认证类型不被支持。	不支持的认证类型。	更改认证类型后重试。
AgentArts.02601014	特殊字符不被允许在API密钥中	"******" 不支持用于API密钥。	请更改后重试。
AgentArts.02601015	指定的字符不被允许在IAM密码和密钥中。	特定字符不支持用于IAM密码和IAM密钥。	请更改后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02601016	数据库连接请求超出时间限制。	数据库连接超时。	检查数据库服务状态和网络连通性重试。
AgentArts.02601017	数据库连接过程中出现错误。	数据库连接失败。	检查数据库地址、端口、用户名和密码是否正确，确保数据库服务正常运行后重试。
AgentArts.02601018	定时任务添加过程中出现系统错误。	添加定时任务失败。	检查定时任务配置，确保时间格式和任务参数正确后重试。
AgentArts.02601019	意图包名称超出系统设定的字符限制。	导入的意图包名称长度超过设定字符。	缩短意图包名称后重试。
AgentArts.02601020	意图样例名称超出系统设定的字符限制。	导入的意图样例名称超过设定字符。	缩短意图样例名称之后重试。
AgentArts.02601021	意图包名称支持中英文、数字、下划线、中划线。	导入的意图包名称无效。	请修改导入的意图包名称。
AgentArts.02601022	意图名称支持中英文、数字、下划线、中划线。	导入的意图名称无效。	请修改导入的意图名称。
AgentArts.02601023	意图样例支持中英文、数字、下划线、中划线。	导入的意图样例无效。	请修改导入的意图样例。
AgentArts.02601024	意图样例不允许重复。	意图样例重复。	请检查意图样例。

提示词工程错误码

表 12-10 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701002	数据集添加失败。	系统错误，添加数据集失败。	请联系运维人员，定位排查。
AgentArts.02701003	数据集删除失败。	删除数据集失败。	请联系运维人员，定位排查。
AgentArts.02701004	ListBuckets接口调用异常。	获取OBS桶列表失败。	检查OBS服务连通性、权限及相关配置后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701005	桶名称对应的文件夹查询异常。	获取OBS桶文件夹失败。	确认桶名称是否正确，检查文件夹权限与存在性后重试。
AgentArts.02701006	数据格式不符合JSON规范，转换流程异常。	JSON转换失败。	检查待转换数据格式，修复格式问题后重试。
AgentArts.02701007	文件写入权限不足、磁盘空间不足或IO异常。	JSON文件写入失败。	检查文件写入权限、磁盘容量，确认文件路径有效后重试。
AgentArts.02701008	类型与字符串不匹配，解析规则不兼容。	JSON解析失败。	核对数据类型与字符串，修正格式后重试。
AgentArts.02701009	当前访问主体不在服务白名单内，校验不通过。	服务白名单校验失败。	申请加入服务白名单，核对访问身份信息后重试。
AgentArts.02701010	白名单服务返回响应异常，无有效结果。	获取白名单状态响应失败。	检查白名单服务状态，网络连通性后重试。
AgentArts.02701011	网络故障、服务超时或服务不可用。	请求白名单服务异常。	检查网络连接，确认白名单服务正常运行后重试。
AgentArts.02701012	请求未携带IAM Token，身份凭证为空。	IAM Token信息缺失。	补充有效的IAM Token，重新发起请求。
AgentArts.02701013	HTTP客户端配置错误、依赖缺失或资源不足。	HTTP客户端初始化失败。	检查HTTP客户端配置，补充相关依赖后重试。
AgentArts.02701014	IAM服务返回异常响应，业务校验不通过。	IAM服务响应失败。	核对请求参数，检查IAM服务状态后重试。
AgentArts.02701015	网络中断、服务超时或IAM服务故障。	请求IAM服务异常。	排查网络问题，等待IAM服务恢复后重试。
AgentArts.02701016	代码调用未实例化的对象，触发空指针。	空指针异常。	排查代码逻辑，对空对象做判空处理后重试。
AgentArts.02701017	模型服务接口调用异常，数据查询失败。	查询模型列表失败。	检查模型服务状态，核对查询参数后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701018	指定的模型标识无对应数据记录。	模型不存在。	确认模型标识正确，检查模型是否已创建。
AgentArts.02701019	请求参数格式、范围或必填项不满足规则。	参数校验失败。	核对接口参数规范，修正参数后重试。
AgentArts.02701020	项目ID对应的订单数据查询异常。	查询订单信息失败。	确认项目ID有效，检查订单服务状态后重试。
AgentArts.02701021	模型调用完成，未返回有效结果。	调用模型失败。	检查模型状态与输入参数，重新发起模型调用。
AgentArts.02701022	接口调用成功，响应体为空数据。	调用大模型接口失败。	核对接口输入参数，检查大模型服务返回逻辑。
AgentArts.02701023	接口请求发起、连接或响应接收阶段异常。	调用大模型接口失败。	检查网络与接口地址，确认大模型服务可用后重试。
AgentArts.02701024	请求超时未获取到响应，服务处理耗时过长。	调用大模型接口超时。	调大超时时间，检查大模型服务性能后重试。
AgentArts.02701025	调用流程中途中断，未完成正常响应。	大模型调用异常中断。	检查大模型API配置与服务状态，重新发起调用。
AgentArts.02701026	大模型API返回错误码或异常信息。	大模型调用失败。	核对大模型API文档，检查请求与权限配置后重试。
AgentArts.02701027	指定桶名称无对应OBS桶资源。	OBS桶不存在。	确认桶名称正确，检查OBS桶是否已创建。
AgentArts.02701028	文件路径在OBS中无对应文件资源。	OBS文件不存在。	核对文件路径，确认文件已上传至OBS。
AgentArts.02701029	当前身份无OBS资源操作权限，访问被拒绝。	OBS权限拒绝。	申请对应OBS资源操作权限，授权后重试。
AgentArts.02701030	未查询到目标OBS桶，桶资源不存在。	OBS桶不存在。	确认OBS桶信息正确，完成桶创建后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701031	OBS读取IO异常。	OBS文件读取时发生IO错误，网络或服务故障。	检查网络连接，确认OBS服务正常后重试。
AgentArts.02701032	Token项目ID与URL项目ID不匹配。	IAM Token校验失败。	核对项目ID配置，使用一致的身份凭证重试。
AgentArts.02701033	本地图片文件读取异常，上传流程中断。	图片上传失败。	检查本地文件完整性与读取权限，重新上传。
AgentArts.02701034	文件上传至OBS过程中，网络或服务异常。	OBS文件上传失败。	检查网络与OBS服务状态，重新上传文件。
AgentArts.02701035	文件上传流程异常，权限、路径或服务故障。	文件上传失败。	检查文件、路径与权限，确认服务正常后重试。
AgentArts.02701036	上传文件的类型不在系统支持范围内。	不支持的文件类型。	更换系统支持的文件类型，重新上传。
AgentArts.02701037	文件类型未在系统中配置，无法识别。	文件类型未定义。	配置文件类型，或使用已定义的文件类型。
AgentArts.02701038	文件类型与实际内容类型不一致。	文件内容类型不匹配。	核对文件后缀与实际内容，修正文件格式后重试。
AgentArts.02701039	OBS客户端配置错误、密钥无效或依赖缺失。	OBS客户端初始化失败。	检查OBS客户端配置与密钥，补充依赖后重试。
AgentArts.02701040	OBS服务波动，请求响应异常。	OBS服务不稳定。	等待OBS服务恢复稳定，稍后重试。
AgentArts.02701041	Excel表头格式、字段不符合系统规范。	Excel表头校验失败。	按照模板规范调整Excel表头，重新上传。
AgentArts.02701042	文件下载流程异常，网络、权限或文件故障。	文件下载失败。	检查网络、下载权限与文件状态，重新下载。
AgentArts.02701043	文件名包含系统禁止的字符或格式。	文件名非法。	修改文件名，使用合规命名规则后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701044	图片文件损坏、格式无效或存在安全风险。	非法图片文件。	使用完整、合规的图片文件，重新上传。
AgentArts.02701045	数据集仅支持excel、xlsx格式，当前格式不兼容。	数据集文件格式错误。	转换文件为excel或xlsx格式，检查OBS文件后重试。
AgentArts.02701046	同名或同标识数据集已创建，数据重复。	数据集已存在。	更换数据集名称，或使用已有数据集。
AgentArts.02701047	数据集数量超出500的系统配额限制。	数据集数量超配额。	清理无用数据集，或申请提升配额后重试。
AgentArts.02701048	数据集正被评估任务引用，无法执行操作。	数据集被占用。	等待评估任务完成，解除占用后重试。
AgentArts.02701049	数据集内容长度超出1000的系统配额限制。	数据集内容长度超配额。	精简数据集内容，或申请提升配额后重试。
AgentArts.02701050	Excel文件损坏、格式错误或读取权限不足。	读取数据集Excel失败。	修复Excel文件，检查读取权限后重新操作。
AgentArts.02701051	OBS文件下载网络中断、服务异常或权限不足。	下载OBS文件失败。	检查网络、OBS权限与服务状态，重新下载。
AgentArts.02701052	ZIP文件大小超出系统允许的最大限制。	ZIP文件体积过大。	压缩文件拆分，或使用更小体积的文件重试。
AgentArts.02701053	文件路径格式错误、不存在或无访问权限。	文件路径无效。	核对文件路径，确认路径有效且有权限访问。
AgentArts.02701054	文件大小超出单条目体积限制。	文件条目体积过大。	拆分文件，降低文件大小后重试。
AgentArts.02701055	文件总大小超出系统限制。	文件体积过大。	压缩或拆分文件，符合大小限制后重试。
AgentArts.02701056	OBS文件ZIP炸弹安全检测流程异常。	ZIP炸弹检测失败。	检查ZIP文件安全性，更换合规文件后重试。
AgentArts.02701057	数据集缓存文件损坏、缓存服务异常。	读取数据集缓存异常。	清理数据集缓存，重新生成缓存后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701058	系统未查询到对应数据集文件资源。	数据集文件不存在。	确认数据集文件已上传，检查文件路径后重试。
AgentArts.02701059	数据集行号不在[10，50]有效区间内。	数据集行号超出范围。	调整数据集行号，符合区间要求后重试。
AgentArts.02701060	数据集文件头行无有效字段信息。	数据集头行为空。	补充数据集头行内容，检查文件后重试。
AgentArts.02701061	数据集中包含无有效数据的空行。	数据集存在空值行。	删除数据集中空值行，整理数据后重试。
AgentArts.02701062	数据集表头字段数量超出20个限制。	数据集表头数量超限。	精简表头字段，不超过20个后重试。
AgentArts.02701063	数据集表头字段存在空值，不符合规范。	数据集表头存在空值。	补充表头空值字段，检查文件后重试。
AgentArts.02701064	数据集表头存在重复字段名称。	数据集存在重复表头。	修改重复表头名称，保证字段唯一后重试。
AgentArts.02701065	变量数据集下载服务调用或文件获取异常。	下载变量数据集失败。	检查服务状态与网络，重新下载变量数据集。
AgentArts.02701066	下载的变量数据集无有效数据内容。	变量数据集为空。	确认变量数据集已生成数据，重新下载。
AgentArts.02701067	数据集未配置预期结果，或结果数据为空。	数据集预期结果为空。	补充数据集预期结果，完善数据后重试。
AgentArts.02701068	任务ID无对应任务数据记录。	任务不存在。	确认任务ID正确，检查任务是否已创建。
AgentArts.02701069	任务ID删除流程异常，权限或状态不允许。	删除任务失败。	检查任务状态与操作权限，符合条件后重试。
AgentArts.02701070	当前提示词工程名称已存在，数据重复。	提示词工程名称重复。	更换提示词工程名称，重新创建。
AgentArts.02701071	任务名称对应的任务创建流程异常。	创建任务失败。	检查任务参数与权限，重新创建任务。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701072	提示词工程任务删除权限不足、状态异常。	删除提示词工程任务失败。	检查任务状态，获取删除权限后重试。
AgentArts.02701073	按条件查询任务列表，数据获取异常。	查询任务列表失败。	核对查询条件，检查任务服务状态后重试。
AgentArts.02701074	任务ID更新流程异常，参数或状态不合法。	更新任务失败。	检查任务ID状态与更新参数，重新操作。
AgentArts.02701075	任务数量超出系统配额限制。	提示词任务数量超配额。	清理无用任务，或申请提升配额后重试。
AgentArts.02701076	任务存在运行中的评估任务，锁定无法删除。	任务被占用无法删除。	等待评估任务结束，解除锁定后删除。
AgentArts.02701077	评估任务当前状态不支持当前操作。	评估任务状态错误。	检查评估任务状态，符合操作条件后重试。
AgentArts.02701078	同名或同条件评估任务已创建，数据重复。	评估任务已存在。	更换评估任务标识，或使用已有评估任务。
AgentArts.02701079	所选评估方法不符合任务要求，配置错误。	评估方法选择错误。	重新选择合规的评估方法，配置后重试。
AgentArts.02701080	未查询到目标评估任务，数据不存在。	评估任务不存在。	确认评估任务信息正确，检查任务是否已创建。
AgentArts.02701081	评估任务信息查询接口调用异常。	查询评估任务信息失败。	检查服务状态，核对查询参数后重新查询。
AgentArts.02701082	任务处于挂起/运行/暂停/重试状态，锁定禁止删除。	评估任务无法删除。	停止任务或等待完成，解除状态锁定后删除。
AgentArts.02701083	评估任务重试流程异常，服务或参数错误。	评估任务重试失败。	检查任务参数与服务状态，重新发起重试。
AgentArts.02701084	评估任务更新参数错误、状态不合法。	更新评估任务失败。	查看错误详情，修正参数与状态后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701085	评估任务未执行成功，无有效结果可下载。	评估任务无法下载。	等待评估任务成功完成，再执行下载操作。
AgentArts.02701086	模板数据查询、文件生成或导出流程异常。	导出预置提示词模板失败。	检查模板数据与导出权限，重新发起导出。
AgentArts.02701087	评估结果数据查询、解析流程异常。	查询评估结果异常。	检查评估任务状态，重新查询评估结果。
AgentArts.02701088	评估任务数量超出系统配额限制。	评估任务数量超配额。	清理无用评估任务，或申请提升配额后重试。
AgentArts.02701089	文件名以特殊字符开头，不符合命名规范。	导出文件失败。	修改文件名，避免以特殊字符开头后重试。
AgentArts.02701090	模板创建数据校验、存储流程异常，无法创建。	创建模板异常。	检查模板参数与权限，重新创建模板。
AgentArts.02701091	模板更新参数错误、数据存储异常。	更新模板异常。	核对更新参数，检查模板状态后重新操作。
AgentArts.02701092	模板数据查询接口调用、数据解析异常。	查询模板失败。	检查查询参数与服务状态，重新查询模板。
AgentArts.02701093	模板删除权限不足、模板被占用或状态异常。	删除模板失败。	检查模板状态与占用情况，获取权限后重试。
AgentArts.02701094	变量解析、替换或存储流程出现异常。	变量处理失败。	检查变量格式与配置，修正后重新处理。
AgentArts.02701095	未查询到符合条件的提示词模板数据。	提示词模板列表为空。	创建提示词模板，或调整查询条件后重试。
AgentArts.02701096	候选模板数量超出系统配额限制。	候选模板数量超配额。	清理无用候选模板，或申请提升配额后重试。
AgentArts.02701097	配置的变量存在重复标识，冲突异常。	变量重复。	修改重复变量标识，保证变量唯一后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701098	提示词正被评估任务引用，无法执行操作。	提示词被占用。	等待评估任务完成，解除占用后重试。
AgentArts.02701099	历史记录数据存储、写入数据库流程异常。	保存历史记录失败。	检查数据库状态与存储权限，重新保存。
AgentArts.02701100	全局变量数量超出系统配额限制。	全局变量数量超配额	清理无用全局变量，或申请提升配额后重试。
AgentArts.02701101	系统未查询到目标业务数据记录。	数据不存在。	核对查询条件，确认数据已创建后重试。
AgentArts.02701102	候选模板名称已存在，数据冲突。	候选模板名称重复。	更换候选模板名称，重新创建。
AgentArts.02701103	提示词名称已存在，数据重复冲突。	提示词名称重复。	更换提示词名称，重新创建。
AgentArts.02701104	模板数量超出系统配额限制。	模板数量超出配额。	清理无用模板，或申请提升配额后重试。
AgentArts.02701105	仅管理员租户拥有预置模板操作权限。	预置模板操作权限不足。	使用管理员租户账号，或申请对应操作权限。
AgentArts.02701106	待下载模板数量超出系统配额限制。	下载模板数量超配额。	分批下载模板，或申请提升配额后重试。
AgentArts.02701107	温度参数不在[{0}, {1}]有效区间内	模型温度参数非法。	调整温度参数至[{0}, {1}]区间内，重新配置。
AgentArts.02701108	Top_p参数不在({0},{1}]有效区间内。	模型top_p参数非法。	调整top_p参数至({0},{1}]区间内，重新配置。
AgentArts.02701109	Presence_penalty参数不在({0},{1})有效区间内	模型presence_penalty参数非法。	调整参数至({0},{1})区间内，重新配置。
AgentArts.02701110	提示词ID无对应数据记录，数据不存在。	提示词ID不存在。	确认提示词ID正确，检查提示词是否已创建。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701111	Name和manual_score参数同时为空，无有效更新项。	模板更新参数非法。	补充name或manual_score参数，重新更新模板。
AgentArts.02701112	提示词生成核心指令为空，无法执行生成。	提示词生成指令为空。	补充有效的提示词生成指令，重新发起操作。
AgentArts.02701113	模板优化流程异常。	优化模板失败。	根据错误信息修正，重新优化模板。
AgentArts.02701114	优化模板删除权限、状态或依赖异常。	删除优化模板失败。	检查模板状态与权限，符合条件后重新删除。
AgentArts.02701115	优化模板服务网络不通、服务不可用。	访问优化模板服务失败。	检查网络与服务状态，稍后重试。
AgentArts.02701116	优化任务正在运行，锁定禁止删除。	优化任务无法删除。	停止任务或等待完成，解除锁定后删除。
AgentArts.02701117	同名或同条件优化任务已创建，数据重复。	优化任务已存在。	更换优化任务标识，或使用已有优化任务。
AgentArts.02701118	优化任务ID无对应数据记录。	优化任务ID不存在。	确认优化任务ID正确，检查任务是否已创建。
AgentArts.02701119	当前任务状态不允许修改为目标状态。	修改优化任务状态非法。	检查任务状态流转规则，符合条件后修改。
AgentArts.02701120	模板优化任务服务接口调用异常。	访问模板优化接口失败。	检查接口配置与服务状态，重新发起请求。
AgentArts.02701121	优化任务创建接口调用成功，无有效响应数据。	创建优化任务响应为空。	检查接口入参与服务逻辑，重新创建任务。
AgentArts.02701124	状态码未在系统定义，无法识别。	优化任务状态码不存在。	使用系统已定义的状态码，重新操作。
AgentArts.02701125	优化任务启动、暂停、终止等操作流程异常。	优化任务操作失败。	检查任务状态与操作权限，重新执行操作。
AgentArts.02701126	模型类型码未在系统配置，无法识别。	模型类型码不存在。	配置模型类型码，或使用已定义类型码。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701127	提示词工程下有运行中的优化任务。	优化任务已存在。	等待现有任务完成，再创建新的优化任务。
AgentArts.02701128	系统未查询到目标文件，文件资源丢失或未创建。	文件不存在。	确认文件已上传，检查文件路径后重试。
AgentArts.02701129	文件大小超过系统设定的最大允许值。	文件大小超出限制。	压缩或拆分文件，符合大小限制后重试。
AgentArts.02701130	Limit或Offset参数为负、超出范围或格式错误。	分页参数无效。	修正Limit、Offset参数，符合分页规则后重试。
AgentArts.02701131	标签名称已存在，数据重复冲突。	标签名称重复。	更换标签名称，重新创建标签。
AgentArts.02701132	标签写入数据库，存储流程异常。	插入标签失败。	检查数据库状态与标签数据，重新插入。
AgentArts.02701133	标签列表数据查询、数据库访问异常。	查询标签列表失败。	检查数据库与查询参数，重新查询。
AgentArts.02701134	标签ID删除权限不足、标签被占用。	删除标签失败。	检查标签占用情况与权限，重新删除。
AgentArts.02701135	标签ID更新数据、数据库存储异常。	更新标签失败。	核对更新数据，检查数据库状态后重试。
AgentArts.02701136	行业名称已存在，数据重复冲突。	行业名称重复。	更换行业名称，重新创建行业数据。
AgentArts.02701137	行业数据被其他业务模块引用，无法操作。	行业正在使用。	解除业务引用，再执行行业相关操作。
AgentArts.02701138	行业ID更新数据、数据库操作异常。	更新行业失败。	检查更新数据与数据库状态，重新操作。
AgentArts.02701139	行业数据查询、数据库访问流程异常。	查询行业失败。	检查数据库与查询条件，重新查询行业。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701140	行业写入数据库，存储流程异常。	插入行业失败。	检查数据库状态与行业数据，重新插入。
AgentArts.02701141	行业ID无对应数据记录，数据不存在。	行业ID不存在。	确认行业ID正确，检查行业是否已创建。
AgentArts.02701142	项目ID调用getUserModels接口异常。	调用getUserModels失败。	核对项目ID，检查模型服务后重试。
AgentArts.02701143	项目ID查询用户模型列表数据异常。	获取用户模型列表失败。	确认项目ID有效，检查服务状态后重试。
AgentArts.02701144	项目ID查询预设模型列表数据异常。	获取预设模型列表失败。	确认项目ID有效，检查服务状态后重试。
AgentArts.02701145	通用列表数据查询、数据解析流程异常。	查询列表失败。	检查查询参数与服务状态，重新查询。
AgentArts.02701146	模板ID不存在。	模板ID为空。	请联系运维人员，定位排查。
AgentArts.02701147	项目ID不存在。	项目ID不存在。	请联系运维人员，定位排查。
AgentArts.02701149	请求体为空。	请求体不能为空。	请求体不能为空。
AgentArts.02701150	模型不可用。	模型状态校验失败。	选择可用的模型。
AgentArts.02701151	该任务的副本创建超过上限。	复制超出上限。	重新创建任务。
AgentArts.02701152	上传文件的格式错误。	上传数据失败。	检查上传的文件，可以参照模板文件进行修改。
AgentArts.02701153	不存在评估数据集。	评估数据不存在。	先添加评估数据集，再创建优化任务。
AgentArts.02701154	评估数据解析失败。	评估数据解析失败。	检查上传的文件，可用参照模板文件进行修改。
AgentArts.02701160	没有该轮的提示词优化任务。	无该提示词优化任务评估结果。	没有该轮的提示词优化任务。
AgentArts.02701161	不是草稿状态不能修改。	非草稿状态不能修改。	创建副本后，再对副本进行编辑调整。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701162	优化任务正在运行。	优化任务正在运行，不能删除。	等任务执行完毕，再删除。
AgentArts.02701163	测评数据集文件过大。	测评数据集文件超过大小限制。	减少测评数据集文件的大小，不能超过100MB。
AgentArts.02701164	只能上传一个json文件。	上传的json文件只能有1个。	检查上传的文件，可用参照模板文件进行修改。
AgentArts.02701165	压缩包模板文件不存在。	压缩包模板文件不存在。	请联系运维人员，定位排查。
AgentArts.02701166	下载模板失败。	下载模板失败。	请联系运维人员，定位排查。
AgentArts.02701167	定时任务注册失败。	定时任务注册失败。	请联系运维人员，定位排查。
AgentArts.02701169	模板类型错误。	模板类型错误。	模板类型只能是text和multi。
AgentArts.02701170	变量类型错误。	变量类型错误。	变量类型只能是text和multi。
AgentArts.02701171	提示词模板来源错误。	来源类型错误。	对提示词模板来源进行修改。
AgentArts.02701172	文本提示词优化任务存在多模态变量。	文本提示词优化任务不能有多模态变量。	去除或修改文本提示词中的多模态变量。
AgentArts.02701173	当前状态不能执行该操作。	优化任务执行时间未设置。	设置优化任务执行的时间。
AgentArts.02701174	当前状态不能执行该操作。	当前状态不能执行该操作。	当前状态不支持执行该操作。
AgentArts.02701175	提示词反馈优化参数错误。	提示词反馈优化参数错误。	请联系运维人员，定位排查。
AgentArts.02701176	上传图片大小过大。	上传图片大小不能超过限定大小。	上传大小小于限定大小的图片。
AgentArts.02701177	上传的提示词数据量过多。	上传数据超过限制。	提示词评估数据量不能超过500条。
AgentArts.02701178	提示词变量个数过多。	提示词变量个数不能超过50个变量。	减少提示词变量个数。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02701179	提示词变量内容过长。	提示词变量内容为空或者长度不能超过20000个字符。	减少提示词变量内容长度。
AgentArts.02701180	提示词变量名称过长。	提示词变量名称不能为空或者超过20个字符。	减少提示词变量名称长度。
AgentArts.02701181	优化任务名称过长。	任务名称长度超过限制。	减少名称长度。
AgentArts.02701189	上传的文件包含重复变量。	提示词变量名重复。	检查上传的文件变量是否重复。
AgentArts.02701190	未设置期望输出。	AgentArts_PROMPT_OUTPUT未设置。	检查上传的文件是否有期望输出变量。
AgentArts.02701191	上传的数据格式错误。	文件类型错误。	检查上传的数据是否满足json字符串格式。
AgentArts.02701192	上传的数据格式错误。	JSON文件格式错误。	检查上传的数据是否满足json字符串格式。
AgentArts.02701193	模板中存在未定义的变量或语法错误。	渲染模板失败。	请检查模板内容和传入的变量是否匹配，确保模板语法正确。
AgentArts.02701194	提示模板内容为空。	提示模板不能为空。	修改提示词模板内容。
AgentArts.02701195	上传的文件缺失变量。	提示词模板变量缺失。	检查上传的文件，是否缺失变量，可用参照模板文件进行修改。
AgentArts.02701196	导入了空数据文件。	导入数据为空。	检查上传的文件，可用参照模板文件进行修改。

知识库错误码

表 12-11 错误码说明

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03000000	未知错误。	系统内部错误。	请联系技术支持。
AgentArts.03000001	调用KooSearch错误。	系统内部错误。	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	报错场景	报错信息	排查与解决思路
AgentArts.03000002	系统LakeSearch配置信息错误。	LakeSearch服务异常。	请联系技术支持。
AgentArts.03000003	初始化OBS客户端失败。	操作失败。	请检查OBS信息并重试。
AgentArts.03000004	参数错误。	知识库配置信息错误。	请联系技术支持。
AgentArts.03000005	网络连接超时。	API请求超时。	请稍后重试。
AgentArts.03000006	构建SSL上下文失败。	系统内部错误。	请联系技术支持。
AgentArts.03000007	获取IAM上下文失败。	系统内部错误。	请联系技术支持。
AgentArts.03000008	获取IAM临时token失败。	系统内部错误。	请联系技术支持。
AgentArts.03000009	获取IAM Token失败。	系统内部错误。	请联系技术支持。
AgentArts.03000010	获取IAM临时凭证失败。	系统内部错误。	请联系技术支持。
AgentArts.03000011	读取token文件失败。	系统内部错误。	请联系技术支持。
AgentArts.03000012	缓存token失败。	系统内部错误。	请联系技术支持。
AgentArts.03000013	解析IAMToken失败。	系统内部错误。	请联系技术支持。
AgentArts.03000014	生成IAM Token失败。	系统内部错误。	请联系技术支持。
AgentArts.03000015	用户Token转换为域Token失败。	系统内部错误。	请联系技术支持。
AgentArts.03000016	系统参数配置错误。	系统内部错误。	请联系技术支持。
AgentArts.03000017	文件内容为空。	下载OBS对象失败。	请检查文件路径并重试。
AgentArts.03000018	下载OBS对象失败。	系统内部错误。	请联系技术支持。
AgentArts.03000019	无法从OBS桶中删除对象。	系统内部错误。	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	报错场景	报错信息	排查与解决思路
AgentArts.03000020	OBS文件数量超出限制。	无法从OBS获取对象信息。	请检查OBS配置。
AgentArts.03000021	KooSearch服务连接信息正确。	KooSearch服务连接错误。	请联系技术支持。
AgentArts.03000022	LakeSearch服务连接信息正确。	LakeSearch服务连接错误。	请联系技术支持。
AgentArts.03000023	获取知识库默认连接信息失败。	系统内部错误。	请联系技术支持。
AgentArts.03000024	知识源类型错误。	系统内部错误。	请联系技术支持。
AgentArts.03000025	默认知识库连接失败。	操作失败。	请联系技术支持。
AgentArts.03001000	\	参数校验失败。	请检查输入参数是否符合规则。
AgentArts.03001001	输入参数无效。	输入参数无效.	请检查输入参数是否正确。
AgentArts.03001002	无权限操作此资源。	操作失败。	请联系技术支持。
AgentArts.03001003	资源不存在。	资源不存在。	请联系技术支持。
AgentArts.03001004	知识库连接名称已存在。	操作失败。	请修改连接名称后重试。
AgentArts.03001005	知识库资源ID已存在。	操作失败。	请修改后重试。
AgentArts.03001006	数量已达到最大限制。	操作失败。	请删除未使用的资源或购买新的资源后重试。
AgentArts.03001007	文件名非法，文件名不能包含非法字符。	操作失败。	请检查后重试。
AgentArts.03001008	OCR未开启，无法上传以下文件。	操作失败。	请开启OCR功能后重试。
AgentArts.03001009	文件大小超过系统限制。	操作失败。	请检查并重试。
AgentArts.03001010	图标名称不能为空且需符合规范。	图标名称无效。	请检查并重试。
AgentArts.03001011	查询知识库许可错误。	获取许可项失败。	请检查知识库配置。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03001012	任务已在运行。	操作失败。	请不要重复操作。
AgentArts.03001016	记忆不属于该Agent。	操作失败。	请检查并重试。
AgentArts.03001017	记忆不属于当前用户。	操作失败。	请检查并重试。
AgentArts.03001018	您无权在知识库中上传文件。	操作失败。	请联系技术支持。
AgentArts.03001019	您无权在知识库中删除文件。	操作失败。	请联系技术支持。
AgentArts.03001020	知识库中不存在该文件。	操作失败。	请检查是否所有文件都属于该知识库。
AgentArts.03001021	下载的文件为空。	操作失败。	请检查并重试。
AgentArts.03001022	创建FAQ文件分块错误。	操作失败。	请检查该FAQ文件是否属于此知识库。
AgentArts.03001023	删除FAQ文件块错误。	操作失败。	请检查该FAQ文件是否属于知识库。
AgentArts.03001024	更新FAQ文件块错误。	操作失败。	请检查该FAQ文件是否属于知识库。
AgentArts.03001025	从知识搜索下载的FAQ文件为空。	操作失败。	请检查FAQ文件。
AgentArts.03001026	您无权从外部知识库中删除文件。	操作失败。	请检查知识库。
AgentArts.03001027	您无权从外部知识库中上传文件。	操作失败。	请检查知识库。
AgentArts.03001028	OBS目录深度超出限制。	操作失败。	请检查OBS目录后重试。
AgentArts.03001029	OBS配置不存在。	操作失败。	请检查OBS配置后重试。
AgentArts.03001030	查询OBS资源失败。	操作失败。	请检查OBS配置后重试。
AgentArts.03001031	检索的知识库数量超过最大限制	操作失败。	请稍后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03001032	第三方知识库并非来自同一来源。	操作失败。	请检查并重试。
AgentArts.03001033	知识库连接不存在。	操作失败。	请检查并重试。
AgentArts.03001034	创建重试文件任务出错。	操作失败。	请检查所有文件是否属于该知识库。
AgentArts.03001035	无法删除知识库连接。	操作失败。	请先移除此连接的所有外部知识库。
AgentArts.03001036	外部知识库地址或认证凭据不正确。	外部知识库连接失败。	请检查您的地址和认证凭据。
AgentArts.03001037	外部知识库中未找到可用的连接信息。	未能获得连接能力或连接器类型。	请检查外部知识库的连接信息和知识库类型。
AgentArts.03002001	文件格式不正确。	上传知识库文件失败。	请检查文件格式后重试。
AgentArts.03002002	知识库来源无效。	操作失败。	请检查并重试。
AgentArts.03002003	搜索模型未注册到KOS。	操作失败。	请联系技术支持。
AgentArts.03002004	搜索模型url错误。	操作失败。	请联系技术支持。
AgentArts.03002005	搜索模型类型不支持。	操作失败。	请联系技术支持。
AgentArts.03002006	OCR模型已存在且只能存在一个。	操作失败。	请联系技术支持。
AgentArts.03002007	文件上传超出速率限制。	操作失败。	请稍后再试。
AgentArts.03002008	由于文件中存在大量重复数据，文件已被压缩，实际文件大小已超出限制。	操作失败。	请检查文件并重试。
AgentArts.03002010	这些知识已被删除或关闭。	无法在知识库中检索。	请检查这些知识的状态。
AgentArts.03002012	当前知识库不支持操作。	操作失败。	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	报错场景	报错信息	排查与解决思路
AgentArts.03002013	获取图标失败。	操作失败。	请联系技术支持。
AgentArts.03002014	知识库已激活，当前不支持更新。	操作失败。	请检查知识库状态。
AgentArts.03002015	知识路由策略无效。	获取知识路由策略失败。	请检查知识路由策略并重试。
AgentArts.03002016	知识合并策略无效。	获取检索合并策略失败。	请检查知识合并策略并重试。
AgentArts.03002017	无法从上下文中获取connectionId。	操作失败。	请联系技术支持。
AgentArts.03002019	未找到该文件或该文件已过期。	操作失败。	请检查后再重试。
AgentArts.03002020	标签名称已存在。	操作失败。	请修改标签名称后重试。
AgentArts.03002100	知识库连接已存在。	操作失败。	请联系技术支持。
AgentArts.03002101	知识库连接状态错误。	操作失败。	请联系技术支持。
AgentArts.03002102	外部知识库连接信息错误。	外部知识库连接错误。	请检查连接信息。
AgentArts.03002103	创建知识失败。	操作失败。	请检查后重试。
AgentArts.03002104	第三方知识库连接信息错误。	查询第三方知识库错误。	请检查连接信息后重试。
AgentArts.03002105	第三方知识库连接信息错误。	查询第三方知识库标签错误。	请检查连接信息并重试。
AgentArts.03002106	外部知识库不支持文件下载。	操作失败。	请更换知识库稍后重试。
AgentArts.03002107	第三方知识库连接不存在。	操作失败。	请检查连接信息并重试。
AgentArts.03002108	RAG流连接信息错误。	RAG流连接错误。	请检查连接信息。
AgentArts.03002109	服务桥接连接信息错误。	服务桥接连接错误。	请检查连接信息。
AgentArts.03003001	JSON格式错误。	JSON格式错误。	请根据实际信息修改参数。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03003002	参数不能为空。	参数错误。	请检查并重试。
AgentArts.03003003	URL参数解析异常。	URL参数解析异常。	请检查请求参数格式。
AgentArts.03003004	URL路径非法。	URL路径非法。	请检查请求路径是否正确。
AgentArts.03003005	从知识库获取内容时发生错误。	ES查询错误。	建议检查ES节点状态。
AgentArts.03003006	无法为知识库创建索引。	ES索引创建错误。	建议您检查ES节点状态。
AgentArts.03003007	更新知识库索引时发生错误。	ES更新错误。	建议您检查ES节点状态。
AgentArts.03003008	无法删除知识库索引。	ES索引删除错误。	建议您检查ES节点状态。
AgentArts.03003009	批量删除知识库文档时发生错误。	ES批量删除错误。	建议您检查ES节点的索引状态。
AgentArts.03003010	批量创建知识库索引时发生错误。	ES批量创建索引错误。	请检查LakeSearch/KooSearch的ES可用容量。
AgentArts.03003011	无法从知识库索引中删除文档。	ES文档删除错误。	建议您检查ES节点的状态。
AgentArts.03003012	更新知识库索引中的文档时发生错误	ES文档更新错误。	建议您检查ES节点的状态。
AgentArts.03003013	执行批量操作时发生错误。	ES批量操作错误。	建议您检查ES节点的状态。
AgentArts.03003014	无法与搜索引擎建立连接。	连接ES失败。	建议您检查ES节点的状态。
AgentArts.03003015	修改索引设置时发生错误。	ES索引配置更新错误。	建议您检查ES节点的状态。
AgentArts.03003016	调用语义分析服务失败。	盘古NLP模型调用错误。	请联系技术支持解决。
AgentArts.03003017	问题重写服务调用失败。	多轮重写服务调用错误。	请联系技术支持。
AgentArts.03003018	文件内容提取服务调用失败。	文档解析服务调用错误。	请联系技术支持。
AgentArts.03003019	知识库文件不存在。	知识库操作失败。	请检查知识库文件。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03003020	无法从存储中获取文件数据。	读取文件内容失败。	请联系技术支持。
AgentArts.03003021	无法将文件元数据保存到数据库。	文件信息数据库写入错误。	请联系技术支持。
AgentArts.03003022	无法将文件上传到对象存储服务。	上传OBS错误。	请联系技术支持。
AgentArts.03003023	解析或生成JSON数据时发生错误。	JSON处理错误。	请联系技术支持。
AgentArts.03003024	文件当前正在被处理。	正在处理的文件无法删除。	请稍后再试。
AgentArts.03003025	未找到匹配的数据记录。	数据库查询结果为空。	请检查数据库是否正确。
AgentArts.03003026	执行数据库命令时发生错误。	数据库操作失败。	请联系技术支持。
AgentArts.03003027	更新数据记录时发生错误。	数据库更新错误。	请联系技术支持。
AgentArts.03003028	请求的AI模型当前无法提供服务。	模型不可用。	请联系技术支持。
AgentArts.03003029	指定的资源名称或ID已被使用。	资源已存在。	请检查资源名称或ID并重试。
AgentArts.03003030	无法从远程源获取JSON文件。	JSON文件下载错误。	请联系技术支持。
AgentArts.03003031	无法将JSON文件保存到本地存储或服务中。	JSON文件写入错误。	请联系技术支持。
AgentArts.03003032	已存在同名的知识库。	验证知识库名称失败	请使用其他名称。
AgentArts.03003033	知识库为空。	无法启用知识库。	请先上传文件。
AgentArts.03003034	知识库处于关闭状态。	当前操作不允许。	请先启用知识库。
AgentArts.03003035	会话已超时。	对话已过期。	请开始新的对话。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03003036	同步或刷新知识库内容时发生错误。	知识库更新失败。	请联系技术支持。
AgentArts.03003037	模型不存在。	删除模型错误。	请检查模型信息。
AgentArts.03003038	地址无效。	操作失败。	请检查请求的路径是否正确。
AgentArts.03003039	未找到图片或图片已过期。	操作失败。	请稍后重试。

License 控制错误码

表 12-12 错误码说明

错误码	报错场景	报错信息	排查与解决思路
AgentArts.03201001	上传License文件失败。	License保存失败。	请检查obs相关配置，确保服务能正常上传obs文件。
AgentArts.03201002	没有读取到License文件。	License缺失。	请上传有效的License文件。
AgentArts.03201003	无效的License或者License已经过期。	License无效或者超期。	请上传有效的License文件，确保License的有效期，ESN和服务匹配。
AgentArts.03201004	AgentArts相关微服务分配的cpu资源超过License限定的上限。	资源使用超过上限。	请降低AgentArts相关微服务的cpu资源分配或申请足额的License。
AgentArts.03201005	服务计算cpu资源时未能获取到Pod相关信息。	获取Pod资源失败。	请确保服务以k8s部署，并且account.yaml相关的账号权限配置已生效。
AgentArts.03201006	访问itep服务获取License信息失败。	从itep获取License资源失败。	请确保License已上传至itep证书管理平台。

版权所有 © 华为云计算技术有限公司

记忆管理错误码

表 12-13 错误码说明

错误码	报错场景	报错信息	排查与解决思路
AgentArts.034 01001	以下主题定义重复。	主题定义重复。	请删除重复的主题定义。
AgentArts.034 01002	以下主题中未定义标签。	主题下缺少标签。	请删除未定义标签的主题或者在主题下添加标签。
AgentArts.034 01003	以下标签定义重复。	标签定义重复。	请删除重复的标签定义。

导入导出错误码

表 12-14 错误码说明

错误码	报错场景	报错信息	排查与解决思路
AgentArts.035 01001	预置资源不允许导入。	预置资源不允许导入。	请在相关智能体中重新选择当前环境对应的预置资源。
AgentArts.035 01002	Metadata或dsl为空。	导入文件缺少源数据。	请检查导入文件或重新导出资源。
AgentArts.036 01001	资源不存在。	资源不存在。	资源不存在，请前往对应资源管理界面检查。
AgentArts.036 01002	资源导出异常。	资源导出异常。	资源导出异常，请前往配置页面检查。
AgentArts.036 01003	预置资源不允许导出。	预置资源不允许导出。	预置资源不允许导出。

公共、其它错误码

表 12-15 错误码说明（界面报错）

错误码	报错场景	报错信息	排查与解决思路
AgentArts.020 01001	鉴权信息已失效。	鉴权失败。	请检查登录信息后重新登录。
AgentArts.020 01002	Redis、k8s等配置错误。	服务内部错误。	请联系开发人员。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001003	接口参数校验异常，如控制器输入为空，插件JSON解析失败。	接口参数校验异常。	请重新校验接口参数是否正确。
AgentArts.02001004	图片、文件等上传到OBS失败。	OBS服务出错。	请联系开发人员。
AgentArts.02001005	CSS uni-search服务调用异常。	CSS uni-search服务调用异常。	请联系开发人员。
AgentArts.02001006	文件名不符合规范，如名称中存在".."、"\\"、"/"等。	非法文件名。	请修改文件名称后重试。
AgentArts.02001007	不支持该文件类型上传。	非法文件类型。	请修改文件类型后重新上传。
AgentArts.02001008	文件大小过大。	文件大小超过系统限制。	图片最大尺寸为5120KB，其他文件最大为131072KB。
AgentArts.02001010	项目ID不存在或与系统记录不匹配。	项目ID无效或者不匹配。	请检查项目ID是否正确，确保其在系统中已注册。
AgentArts.02001012	本空间内该用户没有相应的权限。	用户没有该空间的权限。	请联系管理员添加相应权限。
AgentArts.02001013	头像文件大小超过200KB。	头像文件大小超过系统限制。	请重新上传小于200KB的头像文件。
AgentArts.02001014	OBS桶的存储容量已达到或超过限制。	OBS桶容量告急。	请清理OBS桶中的文件或增加存储容量以恢复服务。
AgentArts.02001015	名称格式不符合要求，仅允许包含汉字、英文字母、数字、下划线、连字符(-)、英文括号()、中文括号（）、空格、英文句点(.)，长度1-64个字符。	名称查询仅允许包含汉字、英文字母、数字、下划线、连字符(-)、英文括号()、中文括号（）、空格、英文句点(.)，长度1-64个字符。	请检查名称是否符合格式要求，确保只包含允许的字符。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001016	名称格式不符合要求，必须以中文或英文字母开头，只能包含中文、英文字母、数字、下划线 _ 和连字符 -，且长度必须在2~64个字符之间.	名称格式需要满足条件：以中文、英文字母开头，只能包含中文、英文字母、数字、下划线 _ 和连字符 -，长度在2~64个字符之间。	请检查名称是否符合格式要求，确保以中文或英文字母开头，只包含允许的字符，并且长度在2~64个字符之间。
AgentArts.02001017	智能体API调用次数已用完。	智能体API调用量超过套餐额度。	请升级套餐。
AgentArts.02001018	网络延时过大。	语音识别超时。	请检查网络是否波动过大，或联系技术支持以获取进一步帮助。
AgentArts.02001019	语音识别过程中发生错误，导致识别失败。	语音识别失败。	请检查语音数据是否正确，或联系技术支持以获取进一步帮助。
AgentArts.02001020	当前使用量已超出套餐允许的最大配额。	配额超出套餐限制，请升级套餐。	请升级您的套餐以获取更高的配额限制。
AgentArts.02001021	当前套餐未启用该功能。	套餐功能不支持，请升级套餐。	请升级您的套餐以启用该功能。
AgentArts.02001022	上传的文件为空，无法进行后续处理。	文件不能为空。	请确保选择有效的文件进行上传。
AgentArts.02001023	OBS对象拷贝过程中发生错误。	OBS对象拷贝失败。	请检查源对象和目标对象的名称是否正确，确保OBS服务可用，并重试操作。
AgentArts.02001024	OBS服务不可用、权限不足、Bucket配置错误等。	上传文件到OBS失败。	请联系开发人员。
AgentArts.02001026	URL格式不正确。	非法的URL。	请检查URL格式是否准确。
AgentArts.02001027	当前IP地址被加入黑名单，访问被拒绝。	URL被黑名单拦截。	请检查您的IP地址是否在黑名单中，或联系管理员解除拦截。
AgentArts.02001028	检测到访问内网IP地址，该操作被系统禁止。	禁止访问内网地址。	请确保您的请求不涉及内网IP地址，或联系管理员获取帮助。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001029	工作流数据损坏导致DSL中的Config为空。	工作流Config为空。	请联系开发人员。
AgentArts.02001030	上传的文件不能为空。	非法文件。	请检查文件是否为空。
AgentArts.02001031	图标名称不符合UUID格式要求。	图标名称无效。	请使用符合UUID格式的图标名称，例如：123e4567-e89b-12d3-a456-426614174000.png或使用系统提供的自动命名功能。
AgentArts.02001032	无法从OBS中解析工具JSON。可能是JSON格式损坏、字段不匹配、数据类型错误、或转换工具实体失败。原始数据已损坏或不满足PluginEntity/ToolEntity的结构定义要求。	工具信息不正确。	请检查工具配置文件是否符合最新的JSON架构规范，重新打包并上传插件。
AgentArts.02001033	TTS API服务调用失败，可能由于网络问题或服务不可用。	TTS API服务异常。	请检查网络连接，确保TTS API服务正常运行。如果问题持续，请联系技术支持。
AgentArts.02001034	智能补充问题功能接口调用异常。可能是该功能被禁用、Agent状态异常、版本不兼容、Domain Token验证失败，或相关配置参数不正确。系统检测到异常的接口调用行为，阻止了请求的继续执行。	接口调用异常，queryDomain Token错误。	请确认智能补充问题功能是否已开启，检查Agent状态是否正常，验证版本兼容性，重新获取有效的Domain Token，或联系系统管理员确认该功能权限和配置状态。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001035	工具版本发布权限拒绝。系统仅允许工具创建者或系统管理员（Op租户）执行此操作。请确认您是否拥有该工具的发布权限。	您没有执行此操作的权限，因为您不是创建者。	方案1：如果您是工具创建者，请确认登录账号是否正确。 方案2：请联系工具创建者申请发布权限。 方案3：如为管理员操作，请确认当前账号是否属于Op租户权限。 方案4：检查用户权限配置是否正确，必要时联系系统管理员修复权限信息。
AgentArts.02001036	API网关API组创建失败。系统在调用API网关服务创建API组时发生异常。可能是认证令牌过期、网络连接超时、参数配置错误、API网关服务不可用、配额超额或权限不足。	API删除接口错误。	请联系开发人员。
AgentArts.02001037	文件上传请求频率超出LakeSearch服务限制。系统在调用LakeSearch服务上传知识库文档时触发了速率限制器，可能原因包括： ● 单个用户上传频率过高； ● 并发上传数量超限； ● 用户配额已用完； ● IP访问频率过高； ● 应用/项目的调用配额不足。	文件上传超出速率限制，请稍后重试。	方案1：降低上传频率，在连续上传操作之间增加延迟（如每分钟2-3个文件）。 方案2：实现分批次上传机制，将大量文件分成小批次处理。方案3：检查并升级用户配额，联系管理员调整上传限制。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001038	流式接口在执行过程中超过了系统设定的超时时间限制，导致请求被中断。	流式接口执行超时。	请检查网络连接状态和数据处理逻辑，必要时增加超时时间设置，或联系技术支持优化接口性能。
AgentArts.02001039	工作流模型配置无效导致前言生成失败。系统在解析工作流配置时发现以下问题之一： • Root级配置对象为空或缺失； • Default_model字段不存在于配置中； • Default_model值无法转换为预期的Map<String, String>格式； • Default_model字段类型不匹配（如为字符串、数组、null等非对象类型）。	模型不存在或模型配置转换失败。	方案1：检查工作流DSL配置文件是否存在字段"default_model"，确认其值是一个有效的对象格式。 方案2：验证配置文件的JSON格式是否正确， 可使用JSON格式化工具检查语法。 方案3：检查default_model字段的值格式，确保不包含特殊字符或无效数据。 方案4：检查工作流配置文件是否被正确部署到对象存储。
AgentArts.02001040	在尝试获取Redis分布式存储的bucket资源时发生异常，可能是Redis服务不可用或权限问题。	Redisson获取bucket失败。	请检查Redis服务状态、网络连接和权限配置，确保Redis服务正常运行且有足够的访问权限。
AgentArts.02001041	在获取用户认证token过程中发生异常，可能是认证服务不可用或用户信息验证失败。	获取用户token错误。	请检查用户登录状态，确认认证服务正常工作，必要时重新登录或联系系统管理员。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001042	在执行SQL查询语句时发生异常，可能是由于SQL语句语法错误、数据库连接问题或权限不足导致的。	SQL执行失败。	请检查SQL语句语法是否正确，确认数据库连接状态正常，验证当前用户是否具有相应的数据库操作权限。
AgentArts.02001043	系统在获取提供方认证元数据时返回空结果，导致模型服务提供方更新操作失败。具体原因可能是： • 指定提供方，尚未配置任何认证元数据； • 提供方认证元数据被意外删除或置为不可用状态； • 元数据服务查询出现数据库异常或SQL执行失败； • 认证配置与提供方类型不匹配导致元数据信息缺失。	获取提供方认证元数据失败。	方案1：检查提供方配置是否完整，进入提供方管理界面查看是否存在认证元数据配置。 方案2：重新配置提供方认证信息，确保选择正确的认证类型并填写完整参数。 方案3：查看系统日志，分析查询认证元数据的SQL执行日志。 方案4：清除相关缓存并重试操作，可能存在缓存不一致问题。
AgentArts.02001044	当前节点类型不在支持范围内。	节点类型不支持。	请检查节点类型是否正确，或联系技术支持以获取更多帮助。
AgentArts.02001046	解析事件数据答案时发生错误，可能由于数据格式不正确或数据内容异常。	解析事件数据答案失败。	请检查事件数据答案的格式和内容是否正确，确保其符合预期的JSON格式。
AgentArts.02001047	在获取资源时发生错误，导致agents、workflows或tools导入失败。	导入agents、workflows或tools失败。	请检查输入内容是否正确，或联系技术支持以获取进一步帮助。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001048	系统无法为指定的认证类型找到对应的认证适配器。系统在认证适配器注册表中查询该类型时返回空结果，可能原因包括： ● 该认证类型从未在系统中注册或插件未正确加载； ● 认证类型名称拼写错误（大小写敏感）； ● 认证插件初始化失败或被禁用； ● 依赖的认证服务或库不可用； ● 动态插件加载机制失败。	认证类型不支持。	方案1：检查认证类型名称是否正确，确保大小写拼写无误。方案2：查看系统日志中的插件加载失败信息，排查依赖问题。 方案3：检查认证插件包是否已正确部署并在系统路径中可用。 方案4：检查是否启用了动态插件加载功能。 方案5：查看系统配置文件中是否包含了该认证类型的正确定义。
AgentArts.02001049	在初始化HTTP客户端时发生错误，可能是由于SSL上下文配置不正确或安全随机数生成器不可用。	HTTP客户端初始化失败。	请检查SSL上下文配置，确保使用正确的协议版本（如TLSv1.2），并验证安全随机数生成器是否可用。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001050	系统无法获取请求的IAM上下文信息。系统按以下顺序尝试获取IAM上下文均失败： • 从请求属性中获取（用于处理标准HTTP请求循环）； • 从ThreadLocal线程变量中获取（用于处理异步或下游服务调用）。可能原因包括：IAM认证过滤器未正确配置或位置错误、请求属性在处理过程中被清理、多线程环境下ThreadLocal上下文传递失败、IAM认证服务不可用或认证失败、Token过期或无效、权限验证失败、网关或负载均衡导致的上下文丢失、服务间调用时的上下文传递异常、异步执行任务时的上下文维护问题。	获取请求IAM上下文失败。	方案1：检查IAM认证过滤器配置是否正确，确保位于请求处理链的合适位置。 方案2：查看网络连接状态，确认IAM服务是否可正常访问。方案3：验证用户访问Token是否有效，必要时重新生成Token。 方案4：检查IAM服务API响应状态，确认授权状态正常。 方案5：如果是异步调用问题，检查ThreadLocal上下文是否正确传递或设置。
AgentArts.02001051	CCMSClient为空，导致无法从CCMS服务获取iamSecret。	无法从CCMS服务获取iamSecret：ccmsClient为空，请检查CCMS配置。	请检查CCMS配置，确保ccmsClient已正确初始化。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001052	从CCMS服务获取的iamSecret为空，导致无法继续操作。	无法从CCMS服务获取iamSecret：iamSecret为空。	请检查CCMS服务的配置和状态，确保iamSecret能够正确获取。
AgentArts.02001053	Endpoint的URI格式不正确，导致无法解析。	无法找到endpoint。	请检查endpoint的格式是否正确，确保其符合URI标准。
AgentArts.02001054	Jackson库版本冲突导致无法加载JsonReadFeature和JsonWriteFeature类，系统将回退到使用已弃用的JsonParser和JsonGenerator。	由于版本冲突，无法加载JsonReadFeature和JsonWriteFeature类，将使用已弃用的JsonParser和JsonGenerator替代。	请升级Jackson库版本到兼容版本，或确保所有依赖的Jackson库版本一致，尽管功能正常但建议及时解决版本冲突问题。
AgentArts.02001055	在将文件转换为Base64字符串时发生错误，可能是由于文件内容无法读取或编码失败。	文件转换为Base64字符串失败。	请检查文件是否有效，确保文件内容可以被正确读取和编码为Base64字符串。
AgentArts.02001056	在获取图标图片时发生错误，可能是由于图片文件无法下载或MIME类型无法识别。	获取图标图片失败。	请检查图片文件是否有效，确保文件可以被正确下载，并确认文件扩展名是否正确以识别MIME类型。
AgentArts.02001057	免费套餐已使用完。	当前租户下的应用数量已经超过了套餐配额。	请升级套餐。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001058	Excel导出过程中发生异常，可能的原因包括： ● 导出的消息数据量过大导致内存不足； ● 消息对象中包含不可序列化的特殊数据类型； ● 文件写入权限不足或磁盘空间不够； ● Excel模板配置异常。	导出结构化消息失败。	请按以下步骤排查问题： 1. 减少单次导出的消息数量，建议每次导出不超过200条； 2. 检查消息数据是否包含复杂对象或特殊字符，确保contentMap字段为标准JSON格式； 3. 确保系统有足够的磁盘空间和文件写入权限； 4. 如为数据量大问题，请分批次导出。
AgentArts.02001060	消息模板名称违反了系统的命名规则，可能是名称过短、包含特殊字符或不允许使用的关键字。	消息模板名称不合法。	请按照以下规范修改结构化消息名称： ● 确保名称长度在2-64个字符之间； ● 移除开头和结尾的空格； ● 只使用允许的字符（中文、英文、数字、_、-、空格）； ● 避免使用特殊符号如@、#、%、&等； 示例格式："异常消息模板"、"系统通知_1"、"error message"等。
AgentArts.02001061	导入的数据条数超过限制，最多允许导入200条。	导入数据应不超过200条。	请检查导入的数据量，确保不超过200条。
AgentArts.02001062	尝试创建的消息模板名称在当前范围内已存在，违反了唯一性约束。	消息模板名称重复。	请修改消息模板名称，确保其在当前工作区或项目中具有唯一性，或者选择使用已存在的模板。
AgentArts.02001063	被授权的空间ID不在当前租户的空间集合中。	被授权空间不存在。	请检查被授权的空间ID是否正确，确保其存在于当前租户的空间集合中。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001068	智能体服务调用失败。系统在调用Agent相关服务时发生异常，可能是服务不可用、配置参数错误、网络连接超时或Agent服务内部逻辑错误。	调用AgentService错误，请确认配置正确。	请检查Agent服务的配置是否正确，确认网络连接正常，验证传入的参数是否合法。如有必要，请查看服务日志以获取详细的错误信息，或联系运维团队重启Agent服务。
AgentArts.02001069	请求中缺少API key，导致无法验证身份。	API key缺失。	请在请求头中添加有效的API key以进行身份验证。
AgentArts.02001070	提供的API key无效或与当前用户不匹配。	API key错误或者和当前用户不匹配。	请检查API key是否正确，并确保其与当前用户关联。
AgentArts.02001071	当前请求的租户ID与发布信息中的租户ID不匹配，导致无权限执行操作。	当前租户无权限执行。	请确保当前请求的租户ID与发布信息中的租户ID一致，或联系管理员获取权限。
AgentArts.02001072	当日的调用次数已超过允许的最大调用次数限制。	当日调用次数已达到上限。	请检查调用次数限制，或等待次日重置调用次数。
AgentArts.02001073	在指定时间内未能成功获取分布式锁，可能是由于锁被其他线程或进程占用。	获取分布式锁失败。	请检查分布式锁的使用情况，确保没有其他线程或进程长时间占用锁，或尝试增加获取锁的超时时间。
AgentArts.02001074	在解析JSON数据时发生错误，可能是由于JSON格式不正确或数据不完整。	JSON解析失败。	请检查JSON数据的格式和完整性，确保其符合预期的结构。
AgentArts.02001075	系统检测到该账号处于冻结状态，无法正常访问系统资源。	您的账号已被冻结。	请联系客服解冻账号或咨询相关冻结原因。
AgentArts.02001076	系统检测到当前账户未开通该功能的访问权限。	功能未开通。	请先开通该功能后再尝试使用。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001077	系统检测到创建的消息模板category字段值不是"异常"。当前消息模板功能仅支持异常类型的消息模板，其他类型的消息模板暂不支持创建或编辑。	消息模板仅支持异常类型。	请确保消息模板的category字段设置为"异常"。如果需要创建其他类型的消息模板，请确认业务需求并联系产品经理规划相关功能。当前版本请使用"异常"类型作为category值。
AgentArts.02001078	租户资源清理权限不足。	租户资源清理权限不足。	租户资源清理权限不足，请检查是否使用正确的租户。
AgentArts.02001079	消息模板未找到，请确定空间ID是否正确。	消息模板未找到。	消息模板未找到，切换正确的空间ID。
AgentArts.02001080	URL不在白名单中。	URL不在白名单中。	请确认URL已经添加到白名单。
AgentArts.02001081	未给当前租户创建KMS工作密钥。	未能获取到KMS工作密钥。	重新订阅。
AgentArts.02001082	系统内部异常。	使用KMS加解密失败。	请联系技术支持。
AgentArts.02001083	接口调用频率过高，超出了系统设定的频率限制。	调用过于频繁，请稍后再试。	请稍后重试，降低调用频率。
AgentArts.02001084	无法读取指定的资源文件，请检查资源路径或文件是否存在。	资源读取错误。	请确认资源文件路径是否正确，或检查文件是否被正确打包到资源目录中。
AgentArts.02001085	对象的结构或内容不符合JSON序列化要求，例如对象中包含无法序列化的字段（如循环引用、非可序列化类型、缺少无参构造函数等）。	无法将对象转换为JSON字符串。	检查AdditionalQuestionsConfig类的定义，确保其字段均为可序列化类型（如String、Integer、List等）。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001086	JSON字符串格式不正确，或者其内容与目标类AdditionalQuestionsConfig的结构不匹配，例如字段名不一致、字段类型不匹配、缺少必要字段等。	无法将JSON字符串转换为对象。	检查传入的JSON字符串是否为有效的JSON格式。
AgentArts.02001087	TriggerConfig对象中包含无法序列化的字段（如循环引用、非可序列化类型等）。	将列表转换为JSON时出错。	检查TriggerConfig类的字段定义，确保所有字段均可被Jackson正确序列化。
AgentArts.02001088	JSON字符串格式不正确，例如缺少引号、括号不匹配等。	将JSON转换为列表时出错。	在反序列化前，增加对JSON字符串的校验逻辑，例如使用StringUtils.isNotBlank(json)判断是否为空。
AgentArts.02001089	导入内容不允许修改。	签名验证失败。	请确认导入内容。
AgentArts.02001090	用户未订阅AgentArts云服务。	您未订阅AgentArts云服务，无法使用该功能。	请先订阅AgentArts云服务，订阅成功后可使用该功能。
AgentArts.02001091	隐私协议验证失败。	隐私协议验证失败。	请确认隐私协议。
AgentArts.02001092	上传的文件数量超过限制。	上传的文件数量超过限制个数。	减少上传的文件数量。
AgentArts.02001093	DomainId和当前用户不匹配。	用户domainId校验失败。	请使用当前用户对应的域名访问服务。
AgentArts.02001094	解析主机域名失败。	无效的主机域名。	请使用合法有效的域名访问服务。
AgentArts.02001095	zip发生嵌套。	zip嵌套。	请检查zip包是否规范。
AgentArts.02001096	zip路径深度小于等于10层。	zip路径深度小于等于10层。	请检查zip路径深度是否小于等于10层。
AgentArts.02001097	skill制品包不满足规范。	skill制品包异常。	请检查skill制品包是否满足规范。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001098	描述不能为空。	缺少必填字段。	请检查描述是否为空。
AgentArts.02001099	描述长度需小于1024个字符。	描述长度超过限制。	请检查描述长度是否小于1024个字符。
AgentArts.02001100	含有XML非法字符。	含有非法字符。	请检查是否含有非法字符XML。
AgentArts.02001101	租户信息不存在。	租户不存在。	请确认租户信息是否正确。
AgentArts.02001108	鉴权凭证错误或已过期。	鉴权失败。	请检查鉴权凭证是否正确。
AgentArts.02001109	长度需小于500个字符。	长度超过限制。	请检查描述长度是否小于500个字符。
AgentArts.02001110	技能id为空。	缺少技能id。	请检查技能id是否为空。
AgentArts.02001111	技能不存在。	缺少技能包。	请检查技能包是否存在。
AgentArts.02001112	技能版本不存在。	技能版本异常。	请检查技能版本是否存在。
AgentArts.02001113	obs路径不存在。	obs路径异常。	请检查obs路径是否存在。
AgentArts.02001114	预签名obs路径不存在。	预签名obs路径异常。	请检查预签名obs路径是否存在。
AgentArts.02001115	SKILL.md文件不存在。	SKILL.md文件异常。	请检查SKILL.md文件是否存在。
AgentArts.02001116	文件读取异常。	IO异常。	请重新上传文件。
AgentArts.02001117	当前项目下已存在同ID的技能。	技能已存在。	在重新创建之前删除已有的同ID技能。
AgentArts.02001118	同步删除高码数据失败。	同步删除高码数据失败。	请重试。
AgentArts.02001120	获取projectId失败。	获取projectId失败。	请重试。
AgentArts.02001121	文件大小超过了允许的最大值。	文件大小超出限制。	请使用较小的文件或压缩文件以减少其大小。
AgentArts.02001122	ZIP文件中的条目数量超过了允许的最大值。	文件数量超出限制。	请减少ZIP文件中的文件数量或拆分为多个较小的ZIP文件。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001123	ZIP文件中包含路径穿越漏洞，使用了不安全的文件路径。	文件名包含不安全路径。	请使用安全的文件名，不包含"../"等特殊路径字符。
AgentArts.02001124	ZIP文件中包含符号链接，存在不安全的路径。	文件包含符号链接。	请确保ZIP文件中不包含任何符号链接，只使用常规文件。
AgentArts.02001125	文件的前置元数据中包含了不在白名单中的无效字段。	元数据字段无效。	请删除所有未授权的元数据字段，只保留系统允许的字段。
AgentArts.02001126	在前置元数据中未找到必填的名称字段或该字段值为空。	名称字段缺失。	请在元数据中添加有效的名称字段，确保其不为空。
AgentArts.02001127	缺少权限。	使用KMS加解密失败。	请添加KMS相关权限。
AgentArts.02001128	细粒度权限校验不通过。	细粒度权限校验不通过。	请为该账号授权。
AgentArts.02001129	创建运行时环境失败。	创建运行时环境失败。	1. 请查看当前账号状态，如已欠费，请您尽快充值。 2. 请检查委托是否创建。
AgentArts.02001130	SKILL包的名称name字段不满足格式要求。 名称必须符合以下规则：1) 长度为1-64个字符； 2) 只能包含小写字母(a-z)、\、数字(0-9)和连字符(-)； 3) 必须以字母或数字开头和结尾，不能以连字符开头或结尾。	名称格式不符合规范。	请检查SKILL.md文件中的name字段，确保符合命名规范。 如果是目录名与name字段不一致的问题，请确保包含SKILL.md的目录名与frontmatter中的name字段完全一致。
AgentArts.02001131	当前租户异常冻结， 导致访问失败。	当前租户已被冻结无法访问。	请检查处理租户冻结问题。
AgentArts.02001132	创建代码解释器失败，请稍后重试。	创建代码解释器失败。	请检查代码解释器服务是否正常，或联系管理员。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.02001133	该第三方Skill已存在，不可重复。	第三方skill已存在。	请勿重复导入同一第三方Skill，如需更新请先删除已有版本后重新导入。
AgentArts.02001134	第三方Skill包内容为空，解析失败。	下载的文件内容为空。	请检查上传的Skill包是否完整，或重新打包后上传。
AgentArts.02001135	无法连接到第三方Skill服务，请稍后重试。	连接第三方Skill服务失败。	请检查网络连接，或稍后重试。如问题持续存在，请联系管理员。
AgentArts.02001136	第三方Skill服务返回异常响应。	第三方Skill服务返回错误。	请稍后重试，或联系管理员检查第三方服务状态。
AgentArts.02001137	当前用户处于已禁止状态，导致访问失败。	当前用户未被许可。	用户是否获取许可，或联系管理员。
AgentArts.02001138	当前用户处于待分配许可，导致访问失败。	当前用户未被许可。	请检查用户许可是否正常，或联系管理员。
AgentArts.01101102	请求缺少必填参数。	缺少必填参数。	请检查并补充所有必填参数。
AgentArts.01101103	状态值不在有效范围内。	无效的状态。	请确认状态值符合要求。
AgentArts.01101104	License信息不合法或已失效。	无效的License项。	请检查License信息是否有效。
AgentArts.01101105	当前无可用的License资源。	无License可被消费。	请检查License剩余数量或申请新的License。
AgentArts.01101106	暂无已占用的license可释放。	无license可被释放。	请确认license已被占用且可进行释放操作。
AgentArts.01101107	待释放数量超过已占用的license数量。	无充分的license释放。	请检查待释放的license数量不超过已占用数量。
AgentArts.01101108	鉴权凭证错误或已过期。	鉴权失败。	请检查鉴权凭证是否正确。
AgentArts.01000000	参数格式异常。	提供的参数不符合预期的格式或结构。	请参考API文档，检查并修正所有参数的格式、类型及结构后重试。
AgentArts.01000001	数据分析错误。	服务器无法解析请求中的JSON、XML或其他格式的数据体。	请检查请求体格式，确保它是完整、未损坏且符合语法规则的。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.010 00002	输入参数不能为空。	关键参数的值为空（null, 空字符串或空数组）。	请为必填参数提供一个有效的非空值。
AgentArts.010 00003	系统内部错误!	服务器在处理请求时发生了未捕获的异常或依赖服务故障。	此错误通常为临时性问题，请稍后重试。如持续发生，请联系技术支持并提供错误码。
AgentArts.010 00004	时间格式解析错误。	输入的时间字符串无法被解析为系统预期的日期时间格式。	请输入符合"年-月-日 时:分:秒"（例如：2024-01-01 10:30:00）格式的时间。
AgentArts.010 00005	您的权限不足，请配置对应的权限。	当前用户缺少基于属性的访问控制策略中定义的具体操作权限。	请联系管理员，为您添加执行此操作所需的权限。
AgentArts.010 00006	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。
AgentArts.010 00007	CBC校验失败	CBC校验失败	根据提示信息，进行操作。
AgentArts.010 00008	该租户不允许修改状态。	不允许修改状态。	请勿修改此账号状态。
AgentArts.010 00009	当前用户状态为不允许访问。	当前用户无访问许可。	您可以联系或者操作根账号，释放许可或解除禁止或增购成员来改变状态。
AgentArts.010 00010	成员校验失败，此成员不存在或校验异常。	成员校验失败。	请您传入正确成员信息。
AgentArts.000 00007	订购产品和升级资源不能同时为空。	操作请求中既未选择新产品订购，也未指定现有资源升级。	请至少填写一项：需订购的新产品列表，或需升级的现有资源ID。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000008	单次最多订购/升级产品，最少1个产品。	批量操作数量超出系统单次处理上限或低于下限。	请将本次操作的产品数量调整至1到上限数量之间，分批处理。
AgentArts.00000009	产品最多拥有约束限制内的属性。	为产品配置的属性数量超过了系统定义的最大限制。	请精简属性配置，删除非必要属性，确保总数不超过最大限制。
AgentArts.00000010	产品信息为空或格式错误。	请求中的产品信息对象缺失、为空或关键字段格式不正确。	请检查请求体，确保产品信息对象完整且符合JSON结构定义。
AgentArts.00000011	子产品编码为空或格式错误。	未提供子产品编码，或提供的编码格式不符合约定规则。	子产品编码为必填项，请按‘ABC-123’格式提供有效编码。
AgentArts.00000012	产品标编码无效。	提供的产品标编码无法通过校验，可能已被废弃或不存在。	请从最新的产品标签列表中获取并填写有效的标编码。
AgentArts.00000017	日志记录失败。	系统在尝试写入操作日志时遇到存储或网络错误。	请稍后重试。如问题持续，可能是日志服务异常，请联系运维。
AgentArts.00000018	规格不满足最小要求。	资源配置（如CPU、内存）低于该产品要求的最低运行规格。	请将资源配置提升至等于或高于最低要求。
AgentArts.00000019	协议版本错误。	请求头或参数中指定的API协议版本已废弃或不支持。	请查阅最新API文档，将协议版本更新为当前支持的版本号。
AgentArts.00000020	数据解析错误。	服务器无法解析请求中的JSON、XML或其他格式的数据体。	请检查请求体格式，确保其是完整、未损坏且符合语法规则的。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000021	用户类型查询失败。	调用用户信息服务时超时或返回了非预期结果。	请确认用户ID有效，并检查用户管理服务是否正常运行。
AgentArts.00000022	输入参数不能为空。	关键参数的值为空（null, 空字符串或空数组）。	请为必填参数提供一个有效的非空值。
AgentArts.00000023	发送请求失败。	系统在调用下游服务时发生网络连接错误或请求被拒绝。	请检查网络连通性及目标服务的可用性，稍后重试。
AgentArts.00000024	输入时间都为空或无效。	所有提供的时间参数均为空，或无法解析为有效的时间戳/日期。	请至少提供一个有效时间，格式为‘YYYY-MM-DD HH:MM:SS’。
AgentArts.00000025	查询时间区间大于最大范围天数。	查询的起始时间与结束时间间隔超过了系统允许的最大范围天数。	请缩小查询范围，确保时间间隔在最大范围天以内。
AgentArts.00000026	权限不足。	当前用户角色或访问令牌（Token）缺少执行此操作所需的权限点。	请联系系统管理员，为您的账户或角色添加相应操作权限。
AgentArts.00000027	系统内部错误!	服务器在处理请求时发生了未捕获的异常或依赖服务故障。	此错误通常为临时性问题，请稍后重试。如持续发生，请联系技术支持并提供错误码。
AgentArts.00000031	委托创建失败。	在IAM系统创建委托时失败，可能因名称重复、策略错误或配额不足。	请检查委托名称唯一性、策略文档语法，并确认委托数量未超限。
AgentArts.00000034	实例信息不存在。	在资源池中未查询到指定ID对应的服务实例。	请确认实例ID输入正确，并检查实例是否已被释放或迁移。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000043	资源未找到。	根据提供的资源ID，在当前区域和项目下未找到匹配的资源。	请确认资源ID、所属区域及项目ID是否正确。
AgentArts.00000044	存在未支付订单，请至费用中心-订单管理-我的订单中查看详情。	存在未支付订单，请至费用中心-订单管理-我的订单中查看详情。	请取消或支付未完成订单，以保持状态一致性。
AgentArts.00000045	同步失败，无法订购。	依赖的信息（如价目表、库存）同步未完成或异常。	请等待后台同步任务完成，或联系管理员检查同步服务状态。
AgentArts.00000047	请确认并勾选服务申明。	用户未勾选必须同意的服务等级协议或免责声明复选框。	请阅读页面上的《服务申明》内容，并在同意后勾选确认框。
AgentArts.00000075	更新委托信息失败。	更新IAM委托的信任策略、描述或有效期等信息时，服务端处理失败。	请检查待更新的委托信息是否符合IAM策略语法和约束，修正后重试。
AgentArts.00000077	订购失败。	产品在最终创建订单或分配资源环节发生失败。	请根据失败日志核对库存、配额及依赖服务状态，修复后重试。
AgentArts.00000078	系统繁忙，请稍等重试。	服务器当前请求量过高或内部处理资源暂时不足。	请等待30秒至1分钟后，重新发起操作请求。
AgentArts.00000079	认证失败。	提供的Token、AK/SK或密码等凭证无效、过期或被撤销。	请重新登录获取新凭证，或检查访问密钥是否有效且具有足够权限。
AgentArts.00000081	AI原生应用引擎话单查询失败。	查询AI原生应用引擎的用量话单时，话单服务异常或查询条件有误。	请检查查询时间范围、实例ID等条件，确认话单服务正常后重试。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000085	创建自定义角色系统失败。	在权限管理系统创建自定义角色体系时，因策略冲突或配置错误失败。	请检查角色策略文档的语法和逻辑，确保符合权限模型规范。
AgentArts.00000086	更新自定义角色系统失败。	更新已有自定义角色系统时，新策略与现有用户绑定关系冲突或格式错误。	请验证新角色策略的兼容性，并确保更新操作有相应权限。
AgentArts.00000087	删除自定义角色系统失败。	目标自定义角色系统仍被用户、用户组或策略引用，无法直接删除。	请先解除所有对该角色系统的引用（如解绑用户），再执行删除。
AgentArts.00000091	产品规格超出限制。	请求的产品规格组合（如CPU+内存+磁盘）超过了该产品系列允许的最大或最小区间。	请参考产品规格表，将配置调整至允许的范围内。
AgentArts.00000092	查询自定义角色系统失败。	根据ID或名称查询自定义角色系统时，系统未找到匹配记录或查询服务异常。	请确认角色系统ID/名称正确，并检查权限服务状态后重试。
AgentArts.00000093	时间格式解析错误。	输入的时间字符串无法被解析为系统预期的日期时间格式。	请输入符合“年-月-日 时:分:秒”（例如：2024-01-01 10:30:00）格式的时间。
AgentArts.00000095	查询license控制项失败。	从许可证管理服务查询功能控制项（如用户数、并发数）时失败。	请确认License密钥有效且未过期，并检查License服务状态。
AgentArts.00000096	产品配额不足。	当前账户针对特定产品的购买数量或资源配额已达上限。	请释放该产品下不需要的资源，或申请提升产品的购买配额。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000104	您的权限不足，请配置对应的权限。	当前用户缺少基于属性的访问控制策略中定义的具体操作权限。	请联系管理员，为您添加执行此操作所需的权限。
AgentArts.00000111	资源用量查询失败。	从计量计费系统查询资源消耗数据时，因服务超时或数据异常失败。	请稍后重试。如持续失败，请联系计费运维团队检查数据管道。
AgentArts.00000119	租户已订购主商品。	当前租户（或项目）下已经存在一个该主商品的有效订单或实例。	同一租户内主商品不可重复订购。如需更多资源，请考虑对现有实例进行扩容。
AgentArts.00000122	访问接口错误，请检查重试！	调用外部或内部接口时，发生HTTP错误、超时或返回了异常状态码。	请检查接口地址、请求参数、网络连通性及身份认证信息，修正后重试。
AgentArts.00000123	暂不支持MCP来源访问。	当前系统不支持来自类型模型控制平台（MCP）的访问请求。	请使用系统支持的MCP来源类型（如 OpenAI, Azure等）发起访问。
AgentArts.00000124	访问的MCP不存在。	请求中指定的模型控制平台ID或名称在注册中心不存在。	请检查MCP标识符是否正确，并确认该MCP已在平台完成注册。
AgentArts.00000125	名称搜索请至少输入两个字符。	搜索框输入的关键词长度少于2个字符，为避免搜索结果过多，系统设定了最小长度限制。	请输入至少2个字符的关键词进行搜索。
AgentArts.00000126	验证服务声明错误。	服务声明内容与预定义模板不匹配，或数字签名验证失败。	请确认您阅读并同意的是最新的、未被篡改的服务声明版本。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000128	公测资格校验失败。	当前用户没有公测资格，请先申请公测。	请先去公测平台申请公测:https://account.huaweicloud.com/usercenter/#/betaManagement?serviceCode=beta_agentarts
AgentArts.00000130	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。	当前用户缺少admin权限，如需进行此操作，请联系主账号添加权限，或由主账号直接完成相关操作。
AgentArts.00000131	使用云服务需要先进行实名认证。	该账户未进行实名认证。	使用云服务需要先进行实名认证。
AgentArts.00000132	CBC校验失败	CBC校验失败	根据提示信息，进行操作。
AgentArts.00000133	不能同时订购团队标准版、团队高级版、企业专业版或企业旗舰版主套餐。	不能同时订购团队标准版、团队高级版、企业专业版或企业旗舰版主套餐。	请先退订团队标准版、团队高级版、企业专业版或企业旗舰版套餐，然后再重新订购，数据不会丢失。
AgentArts.00000134	当前租户未订购主套餐。	当前租户未订购主套餐。	请先完成主套餐的订购。
AgentArts.00000135	按需开通CU失败。	按需开通CU失败。	按需开通CU失败。
AgentArts.00000136	该租户不允许订购收费版商品。	该租户不允许订购收费版商品。	请提工单给AgentArts，进行咨询。
AgentArts.00000137	当前资源不支持升级。	当前资源不支持升级。	当前资源不支持升级。
AgentArts.00000138	你已超过最大成员数量订购上限。	你已超过最大成员数量订购上限。	请升级更高规格的套餐。
AgentArts.00000140	当前用户状态为已禁止，不允许访问。	当前用户无访问许可。	您可以联系或者操作根账号，释放许可或解除禁止或增购成员来改变状态。
AgentArts.00000141	当前租户已被冻结无法访问。	当前租户异常冻结， 导致访问失败。	请检查处理租户冻结问题。

版权所有 © 华为云计算技术有限公司

错误码	报错场景	报错信息	排查与解决思路
AgentArts.00000142	您的账号已欠费，无法正常购买和使用按需计费云服务，请您尽快充值。	您的账号已欠费，无法正常购买和使用按需计费云服务，请您尽快充值。	您的账号已欠费，无法正常购买和使用按需计费云服务，请您尽快充值。
AgentArts.00000143	权限验证失败。	权限验证失败。	请检查用户权限。
AgentArts.00000144	当前用户状态为待分配许可，不允许访问。	当前用户无访问许可。	您可以联系或者操作根账号，释放许可或解除禁止或增购成员来改变状态。

表 12-16 错误码（后台报错）

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100201	在初始化Elasticsearch客户端时发生异常，可能是由于主机地址无效、协议不匹配或证书错误。	Elasticsearch初始化失败。	请核对Elasticsearch的连接地址、端口以及身份验证信息是否正确。
AgentArts.100000	资源不存在。	资源无法找到。	检查是否存在资源。
AgentArts.100001	身份验证凭据无效或已过期。	鉴权失败。	请重新登录系统。
AgentArts.100002	输入参数格式不正确或缺少必要参数。	参数校验失败。	请检查请求参数是否符合接口规范。
AgentArts.100003	第三方服务或SDK的返回异常。	SDK请求失败。	请检查SDK配置、网络连接或目标服务状态。
AgentArts.100004	调用默认加密方法时发生异常，可能是密钥无效或数据格式不支持。	加密操作失败。	请检查加密算法配置、密钥强度及待加密数据的合法性。
AgentArts.100005	调用默认解密方法时发生异常，可能是解密密钥不匹配或密文已被损坏。	解密操作失败。	请确保使用正确的解密密钥，并检查密文数据完整性。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100006	加密组件或加解密组件初始化异常，可能是由于环境缺失或配置错误。	加解密初始化失败。	请检查环境是否包含必要的密码库，并核对加解密配置参数是否正确。
AgentArts.100007	该接口目前处于预留状态，当前版本暂不支持调用。	接口功能保留，当前方法不支持。	请查阅最新API文档，确认该功能是否已开放或是否有替代接口。
AgentArts.100010	连接池中没有可用的服务器实例，可能是所有节点都已离线或达到最大连接数。	无可用连接服务器。	请检查后端服务运行状态、网络负载情况或增加服务器节点。
AgentArts.100011	尝试通过已关闭的连接发送请求或调用函数。	连接已关闭。	请检查连接状态，在调用前确保连接处于活动状态，必要时重新建立连接。
AgentArts.100012	指定的连接ID未绑定到当前服务器，或连接信息已失效。	连接不存在。	请核对SDK和ID是否正确，并确认该连接是否已成功在目标服务器上注册绑定。
AgentArts.100013	提供的ssl_context参数类型无效，不符合ssl.SSLContext或ContexConfigDict的要求。	SSL上下文类型错误。	请检查传入的SSL配置对象，确保其类型为指定的SSLContext或配置字典格式。
AgentArts.100014	在启动指定的长连接服务器实例时发生异常。	长连接服务启动失败。	请根据错误详情中的key值检查相关端口占用、权限配置或网络资源状态。
AgentArts.100015	尝试使用非整数类型的端口号启动WebSocket服务。	WebSocket端口类型错误。	请修正配置中的端口号，确保其为有效的整数（int）类型。
AgentArts.100016	尝试使用非法类型的主机地址（Host）启动WebSocket服务，期望类型为字符串（str）或空值（None）。	WebSocket主机地址类型错误。	请检查配置中的host参数，确保其为有效的字符串格式IP地址、域名或None。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100017	在建立新连接的过程中，服务端期望收到的回响（Echo）标识与实际收到的sdk_id不匹配。	WebSocket连接校验失败。	请检查握手协议逻辑，确保客户端发送的验证标识与服务端预期的sdk_id一致。
AgentArts.100018	当前连接模式强制要求启用SSL/TLS加密，但未检测到有效的ssl_context配置。	缺失SSL上下文配置。	请为该连接模式提供正确的ssl.SSLContext实例或ContexConfigDict配置字典。
AgentArts.100020	配置文件中设定的日志滚动最大字节数不是一个正整数。	日志最大字节数设置错误。	请检查并修改配置文件，确保max bytes的值是一个大于0的整数。
AgentArts.100021	缺少必要的LLM服务配置信息，导致无法初始化模型服务。	LLM服务配置缺失。	请检查配置文件中的LLM相关参数（如API Key、端点地址等）是否完整。
AgentArts.100022	LLM服务返回或请求的内容中缺少必要的格式化字段。	LLM字段格式缺失。	请核对提示词模板或解析逻辑，确保所有必填字段均已定义。
AgentArts.100023	LLM返回的响应数据不符合预定义的结构要求。	LLM响应结构错误。	检查模型输出是否符合JSON或预设格式，必要时调整提示词以增强输出稳定性。
AgentArts.100024	指定的模型来源无效或无法访问（如未授权或路径不存在）。	LLM模型来源错误。	请验证模型提供商、模型名称或本地模型路径是否正确。
AgentArts.100025	向LLM服务发起请求时发生网络异常、超时或接口返回错误。	模型请求失败。	请检查网络连接、API配额是否充足以及服务端运行状态。
AgentArts.100026	模型权重加载失败或推理后端初始化异常。	模型加载失败。	请检查显存/内存资源是否足够，以及模型文件是否完整。
AgentArts.100027	使用了不支持的模型类型或模型任务类型不匹配。	模型类型错误。	请在配置中确认模型类型（如Chat、Completion、Embedding）是否与调用方法一致。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100028	解析LLM输出结果时解码器发生异常，通常是由于非法的编码或截断导致。	模型解析器解码失败。	请检查模型的解码参数（如max_tokens）或尝试清理输出中的异常字符。
AgentArts.100029	Agent的IR字段验证未通过，内部逻辑流转数据异常。	Agent IR校验失败。	请检查Agent的提示词设计或工具链输出是否符合预期的IR规范。
AgentArts.100066	执行HTTP GET请求时发生错误。	GET请求执行失败。	请检查请求的URL是否正确、网络连通性以及目标服务器是否正常响应。
AgentArts.100040	数据不符合预定义的规范或存在逻辑冲突。	IR数据校验失败。	请检查生成的IR数据结构是否完整，并核对字段属性是否合法。
AgentArts.100042	当前系统环境无法识别或不支持该版本的IR数据。	IR版本不支持。	请检查IR版本号，并确保系统已更新至支持该版本的最新状态。
AgentArts.100044	无法从提供的JSON字符串或文件中解析出有效的IR数据，可能存在语法错误。	IR数据加载失败。	请检查JSON格式是否正确，并确保其编码为UTF-8。
AgentArts.100045	在状态存储介质（如Redis或内存）中找不到指定的执行ID（Key）。	无效的执行ID。	请确认执行ID是否正确，或检查该ID是否已过期、被清理。
AgentArts.100051	无法建立与数据库服务器的连接，可能是由于地址错误、端口不通或认证失败。	数据库连接失败。	请检查数据库服务运行状态、网络防火墙配置及连接字符串中的账号密码。
AgentArts.100052	SQL语句执行过程中发生错误，可能是由于语法错误、约束冲突或权限不足。	数据库执行异常。	请检查生成的SQL语句逻辑，并核对数据库表结构与索引配置。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100061	对象转换为可传输或可存储格式（如JSON/Protobuf）时发生异常。	序列化失败。	请检查对象中是否包含循环引用或不支持序列化的数据类型。
AgentArts.100062	无法将原始数据转换回指定的数据对象，通常是数据损坏或格式不匹配导致。	反序列化失败。	请校验输入数据的完整性，并确保其符合目标数据结构的定义。
AgentArts.100100	找不到指定的Redis服务，或者当前的连接状态异常。	Redis服务不可用。	请检查Redis服务器地址配置、端口开放状态以及网络防火墙设置。
AgentArts.100101	在建立Redis连接池或初始化客户端时发生异常。	Redis初始化失败。	请核对Redis认证密码、数据库索引（DB Index）以及客户端依赖库版本。
AgentArts.100102	向Redis写入数据时发生异常，可能是由于OOM（内存溢出）或类型不匹配。	Redis数据插入失败。	请检查Redis剩余内存空间以及所操作的Key对应的数据类型是否正确。
AgentArts.100103	尝试向Redis列表（List）末尾添加元素时失败。	Redis列表追加失败。	确保目标Key是列表类型，并检查Redis是否处于可写状态。
AgentArts.100104	从Redis列表中移除指定元素时发生错误。	Redis元素删除失败。	请检查Key是否存在以及指定的元素值是否在列表中。
AgentArts.100105	尝试设置Redis列表中特定索引位置的值时失败，可能索引越界。	Redis指定位置修改失败。	请确保指定的索引（Index）在列表范围内，且Key类型正确。
AgentArts.100106	无法从Redis列表中获取指定范围或位置的值。	Redis获取列表元素失败。	请确认Key是否为列表类型，并检查索引范围是否有效。
AgentArts.100107	尝试从列表中弹出（Pop）元素时发生异常。	Redis弹出元素失败。	请检查列表是否为空，或连接是否在操作期间中断。
AgentArts.100108	无法获取指定列表的长度信息。	Redis获取列表长度失败。	请确认该Key确实存在且为列表类型。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100109	从Redis中读取历史记录时发生异常，可能是数据格式损坏或Key已过期。	Redis获取历史数据失败。	请核对历史数据的存储键名，并确保数据序列化格式兼容。
AgentArts.100110	将历史记录写入Redis时失败，可能触发了内存阈值或连接问题。	Redis存储历史数据失败。	请检查Redis写入权限及可用内存空间。
AgentArts.100111	向Redis写入元素值时发生异常。	Redis设置元素失败。	请检查数据大小是否超过限制，或Redis是否配置为只读模式。
AgentArts.100112	尝试从Redis中删除指定的Key失败。	Redis删除元素失败。	请确认该Key是否被锁定或当前用户是否有删除权限。
AgentArts.100113	Redis集群节点格式不正确，期望格式为以逗号分隔的IP:PORT列表。	Redis集群节点配置错误。	请检查配置文件中的cluster_node参数，确保每个节点都包含正确的IP地址和端口号。
AgentArts.100114	无法从Redis中获取集群节点列表，可能是集群状态异常或权限限制。	获取Redis集群节点失败。	请检查Redis集群的运行状态，并确保当前连接账号拥有CLUSTER NODES执行权限。
AgentArts.100502	必要的OBS配置参数不正确或缺失，导致无法通过初始化检查。	OBS配置项错误。	请检查Access Key(AK)、Secret Key(SK)以及Endpoint等配置是否填写正确。
AgentArts.100503	执行Head Bucket操作失败，通常是因为存储桶不存在、名称错误或访问权限不足。	存储桶检查失败。	请确认存储桶（Bucket）名称是否正确，并检查当前账号是否拥有该桶的访问权限。
AgentArts.100504	上传文件到OBS时被服务端拒绝或发生网络中断。	OBS文件上传失败。	请根据返回的状态码检查上传权限、对象大小限制或存储空间配额。
AgentArts.100505	无法从存储桶获取对象，可能是对象已被删除或路径错误。	OBS获取对象失败。	请确保存储桶中存在该对象，并核对对象的Key路径是否完整。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.100506	在存储桶内复制对象时操作失败。	OBS复制对象失败。	请检查源对象是否存在，以及目标路径是否具有写权限。
AgentArts.100507	尝试从存储桶中删除对象时失败。	OBS删除对象失败。	请确认当前账号拥有该对象的删除权限，并核对对象标识符。
AgentArts.100300	调用向量化（Embedding）服务时发生异常，可能是网络超时、API密钥错误或输入文本过长。	Embedding服务请求失败。	请检查Embedding服务的可用性、网络连接以及请求参数是否符合模型输入规范。
AgentArts.101001	系统需要“操作配置”（action_map）是键值对的格式（比如“名称：内容”），但传入的格式不符合要求。	配置格式错误。	请检查配置内容，确认“操作配置”（action_map）部分是“名称：内容”这样的键值对格式，而非其他格式。
AgentArts.101002	在使用异步函数（async_fn）模式下，禁止同时指定包装类（wrap_cls）。	配置规则冲突。	请根据函数性质选择其一：使用wrap_cls进行同步对象包装，或使用async_fn进行纯异步函数注册。
AgentArts.101003	async_fn参数必须是一个异步函数（使用async def定义），不能传入普通同步函数。	配置函数类型错误。	请检查函数定义，确保在函数名前添加了async关键字。
AgentArts.101004	wrap_cls必须是Invokable的子类，以便引擎进行合规的生命周期管理。	配置类型非法。	请检查类继承关系，确保自定义的包装类继承自框架提供的Invokable基类。
AgentArts.101005	检测到其作用类的init方法中存在不支持的关键字参数。	配置不支持的关键字参数。	请简化类的构造函数参数，或使用框架支持的标准参数命名规范。
AgentArts.101007	在指定的引用范围或归属者中找不到可调用实体。	未找到可调用实体。	请核对编排文件中的节点名称是否与注册的组件名称严格匹配。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101008	编排逻辑中引用的变量在当前的上下文或作用域中不存在。	未找到指定变量值。	请检查数据流传递配置，确保该变量在之前的步骤中已被正确定义或赋值。
AgentArts.101009	生成的JSON结构损坏或不符合编排引擎规范。	配置数据格式错误（JSON格式非法）。	请检查编译器的输出，或验证手动修改后的内容满足规范。
AgentArts.101010	invokable_irs字段期望为一个字符串列表（每个字符串应为JSON格式），但实际传入类型不符。	集合类型非法。	请检查传入参数，确保invokable_irs是一个包含JSON字符串的列表。
AgentArts.101011	无法解析或不符合规范。	参数格式错误。	请根据提示修复对应索引处的JSON语法或结构错误。
AgentArts.101012	system_configs参数必须以列表（list）形式提供。	系统配置类型错误。	请将系统配置项封装在列表结构中后再进行传递。
AgentArts.101013	在执行删除操作时，每个Agent必须同时提供agentId和version属性。	Agent标识信息缺失。	请核对待删除列表，确保每个项都包含完整的ID和版本号信息。
AgentArts.101014	接收到了一个执行器层面的未知或不支持的状态码。	执行器状态码不支持。	请检查执行器版本是否与编排引擎匹配，或核对自定义状态码的定义。
AgentArts.101015	流式处理过程中出现了未定义的响应状态码。	未知流状态码。	请查阅流式接口协议文档，确认该状态码是否为新增类型或异常上报。
AgentArts.101016	尝试注销指定的中间表示（IR）组件时失败。	IR注销失败。	请检查该IR是否正在运行中，并根据返回的状态码确认注销权限。
AgentArts.101017	启动或执行指定的IR实例时发生错误。	IR运行异常。	请检查IR依赖的服务是否就绪，以及运行时输入的参数是否符合要求。
AgentArts.101018	无法将用户输入透传给IR实例。	用户输入发送失败。	请确认会话（Session）是否仍处于活动状态，并检查网络连接或回调地址是否有效。
AgentArts.101019	向管理中心注册新的IR组件时被拒绝。	IR注册失败。	请检查IR标识（ID）是否冲突，以及提交的IR定义是否完整且符合规范。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101020	尝试更新已存在的IR定义时发生错误。	IR更新失败。	请验证更新内容的版本号一致性，并确保当前服务允许热更新操作。
AgentArts.101021	系统中找不到IR管理器实例，导致管理操作无法路由。	IR管理器缺失。	请检查初始化配置中是否正确注册了该名称的管理器，并核对拼写是否正确。
AgentArts.101022	无法找到指定的可调用（Invokable）管理器，导致组件调用失败。	Invokable管理器缺失。	请确认执行引擎中已正确配置并启动了对应的Invokable管理服务。
AgentArts.101023	可调用实体（Invokable）的定义数据无法被正确解析，通常由于JSON语法错误或缺少必要描述字段。	Invokable JSON格式非法。	请校验Invokable定义的JSON文本，确保其符合框架的组件描述规范。
AgentArts.101024	从目录批量注册可调用实体时，路径参数必须是字符串（str）类型，当前传入类型不符。	Invokable注册路径类型错误。	请确保传入register_from_dir方法的参数是一个有效的文件夹路径字符串。
AgentArts.101025	指定的注册路径不是一个有效的目录，导致无法扫描其中的组件定义。	Invokable注册路径非法。	请检查路径是否存在，并确保该路径指向的是一个文件夹而非文件或快捷方式。
AgentArts.101026	在扫描目录并批量注册组件的过程中发生异常中断。	目录批量注册Invokable失败。	请检查目录下的所有组件定义文件是否完整，并根据错误详情排查具体哪个文件导致了中断。
AgentArts.101027	尝试将可调用实体注册到指定的插件管理器（PluginManager）时被拒绝。	插件管理器注册失败。	请检查插件管理器的状态，确认是否存在名称冲突或权限限制。
AgentArts.101028	对可调用实体执行特定操作（如启动、停止、卸载）时失败。	Invokable操作失败。	请根据返回的错误码和详细信息，检查该实体的当前状态是否允许执行该操作。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101029	用户输入中的数据类型与请求头或配置中设定的Content-Type不一致。	HTTP内容类型不匹配。	请根据错误提示修正请求头的Content-Type，或调整传入参数的格式以符合预期类型。
AgentArts.101030	编排引擎在执行过程中遇到了未捕获的非预期异常。	编排引擎未知错误。	请查看系统日志以获取详细的堆栈信息，并联系开发人员排查潜在的逻辑漏洞。
AgentArts.101034	缺少用于Invokable管理的可调用实体管理配置项，导致执行器无法初始化。	执行管理器配置缺失。	请检查application.yaml配置文件，确保已正确定义相关管理参数（如存储、路由等）。
AgentArts.101035	缺少用于AgentIR管理的工作节点（Worker）配置信息。	工作节点配置缺失。	请在application.yaml中补充必要的Agent运行时配置，以确保执行引擎能正确调度IR。
AgentArts.101036	系统无法找到模型实现类，可能是类名拼写错误或该模型未被扫描加载。	模型类未找到。	请确认该模型类已正确注册到模型仓库中，并检查其所在的路径是否在扫描范围内。
AgentArts.101037	IR管理器尝试向指定会话发送状态通知或回调时失败。	会话通知失败。	请检查会话连接状态（如WebSocket是否断开）以及回调服务地址是否能正常访问。
AgentArts.101052	在对组件进行单步跟踪或调试模式执行时发生错误。	组件单步调试异常。	请检查调试器连接状态及组件的上下文信息。
AgentArts.101053	当前组件类型（如某些内置原子组件或加密组件）不支持单步调试功能。	组件类型不支持调试。	请确认组件定义中是否开启了调试支持，或尝试直接运行而非调试模式执行。
AgentArts.101054	单个组件的IR定义不符合规范，解析失败。	组件IR非法。	请检查组件的配置文件或DSL编译结果，确保其IR结构完整。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101055	组件结束（End Component）时接收到的输入参数中'userFields'字段应为字典类型。	组件结束调用输入错误。	请检查前序组件的输出或当前组件的参数映射，确保传入的是字典格式。
AgentArts.101056	组件在运行时检测到输入参数不符合其定义。	组件运行输入校验失败。	请根据校验错误提示核对上游组件传递给该组件的参数格式和值范围。
AgentArts.101057	组件的初始化检查失败，其输出字段必须与输入字段保持一致（如状态透传组件）。	组件输入输出一致性校验失败。	请调整组件定义，确保指定的输入输出字段名和类型完全匹配。
AgentArts.101058	尝试在全局上下文中写入数据，但已存在的全局项类型不是字典（dict），导致无法合并或追加。	全局上下文类型冲突。	请检查全局状态管理逻辑，确保在整个编排流程中全局变量的基准类型是一致的。
AgentArts.101099	在初始化组件时，无法加载其特定的配置文件或参数。	组件配置加载失败。	请检查该组件关联的JSON/YAML配置文件路径是否正确，以及文件权限是否正常。
AgentArts.101100	在组件配置的'userFields'中定义了字段，但在实际传入的'inputs'数据中未找到该字段。	组件结束配置映射缺失。	请确保配置中声明要透传或处理的字段在输入数据流中确实存在。
AgentArts.101101	插件的函数集合定义非法，期望为字符串格式的列表、元组或集合。	插件函数列表类型错误。	请检查插件注册信息，确保functions字段是一个包含函数名称字符串的集合。
AgentArts.101102	在对任务插件执行特定操作（如加载、卸载）时失败。	任务插件操作失败。	请根据详细的错误信息排查插件服务状态或权限问题。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101103	更新插件时提供的组件类型非法。系统仅接受RestFul API或Function对象的列表。	更新插件类型不匹配。	请检查更新调用中的参数类型，确保传入的是标准API定义或函数对象。
AgentArts.101104	在解析任务参数或配置时发现未定义的非法字段名。	任务参数字段名错误。	请核对任务定义的规范，移除或修正配置文件中拼写错误的字段。
AgentArts.101105	动态编排中的规划器（Planner）无法加载必要的配置，导致任务拆解功能不可用。	任务规划配置初始化失败。	请检查规划器相关的application.yaml配置段，确保策略模型和参数已正确设定。
AgentArts.101106	任务执行所依赖的底层模型（如用于意图识别或规划的大模型）初始化异常。	任务模型初始化失败。	请检查模型连接凭证、资源可用性（GPU/内存）或后端服务地址。
AgentArts.101107	在任务启动阶段，必要的扩展插件未能成功加载或建立连接。	任务插件初始化失败。	请查看系统日志定位具体是哪个插件初始化超时或报错。
AgentArts.101200	在对编排模型进行合法性检查时发现校验错误。	编排校验失败。	请根据详细的描述信息修正DSL结构或组件连接逻辑，确保符合编排规范。
AgentArts.101201	组件定义的JSON根节点类型非法。系统仅接受对象（dict）或列表（list）。	组件JSON根类型错误。	请检查JSON文件的最外层结构，确保其为支持的格式。
AgentArts.101260	在对根IR执行操作时，其中的子IR发生错误。	子图操作失败。	请检查子图组件的依赖关系及运行状态，确认子IR是否已被正确定义或加载。
AgentArts.101261	尽管所有子图组件都已成功处理，但在执行根Agent IR的最后步骤（如合并或持久化）时发生错误。	根图操作失败。	请检查根组件的全局配置及状态存储介质的写入权限。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.101285	复合类型组件定义不完整：既没有指定IR构建器（Builder），也缺少可选的IR静态定义。	复合组件缺少引用。	请为复合组件提供有效的builder_spec或optional_ir路径。
AgentArts.101286	系统无法定位复合组件所声明的IR构建器逻辑。	复合组件构建器未找到。	请确认构建器类的包名和路径是否正确，并确保已包含在系统的类路径中。
AgentArts.101287	在动态加载复合组件构建器时发生Python导入异常。	构建器导入失败。	请检查构建器代码是否存在语法错误，或其依赖的第三方库是否已安装。
AgentArts.101288	加载的对象并不是一个合法的IR构建器类。	构建器类型错误。	请确保自定义构建器类实现了预期的构建接口或继承了BaseIRBuilder。
AgentArts.101289	构建器在尝试动态生成复合组件实例的IR时发生内部执行错误。	复合组件构建执行失败。	请检查构建器的build()方法逻辑，特别是其对输入参数的处理是否严谨。
AgentArts.101290	在执行操作期间，系统尝试自动注册内置的可调用实体（Invokable）时触发了内部异常。	内部内置组件注册失败。	这通常属于底层框架异常，请检查系统启动日志中关于基础库加载的部分，并确认核心内置插件包是否完整且未被篡改。
AgentArts.102001	模型推理服务响应成功但在Body中未包含任何文本内容或有效负载，通常发生在模型拒绝回答或触发安全过滤时。	LLM服务返回空值。	请检查输入提示词是否敏感，或尝试调整模型的温度值以改变生成策略。
AgentArts.102002	在调用远程模型接口时发生网络层异常，如DNS解析失败、网关超时（504）或连接被重置。	HTTP请求网络错误。	请检查模型服务商的API地址是否可访问，并确认网络防火墙或代理服务器配置正常。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102003	模型服务返回了非200状态码，或者在协议层面返回了明确的错误标识。	LLM业务逻辑错误。	请根据具体报错信息检查账户余额、Token配额或API参数是否合法。
AgentArts.102004	建立TCP连接或SSL握手时失败，通常由于目标服务器宕机或网络不稳定。	LLM服务连接失败。	请验证模型服务端点（Endpoint）是否正确，并尝试增加连接重试机制。
AgentArts.102006	定义的解析逻辑在创建实例时报错。	输出解析器初始化失败。	请检查输出解析器的配置是否符合规范，确保与预期的模型输出格式匹配。
AgentArts.102007	在处理模型返回的字符串时，解析器抛出异常。通常是因为模型输出并非合法的JSON或未包含预期的关键字。	输出解析器执行异常。	请优化提示词以引导模型输出更严谨的结构化数据，或在解析前增加数据清洗步骤。
AgentArts.102008	多模型执行协议（MEP）在分发或聚合多个模型请求时发生逻辑错误。	MEP协议返回异常。	请检查MEP路由规则配置，确保下游各个备选模型的健康状况正常。
AgentArts.102009	提示词引擎在执行过程中丢失了必需的运行时上下文引用，导致无法读取Trace ID或变量池。	运行时上下文缺失。	请检查线程池调度或异步调用链中是否正确传递了Context对象。
AgentArts.102010	在将提示词定义转换为IR时解析失败。	IR构建错误。	请检查提示词配置中的DSL语法，确保条件分支和变量引用的逻辑闭环。
AgentArts.102011	配置中指定了系统无法识别的服务提供商或协议类型。	不支持的模型服务类型。	请核对支持的服务类型列表（例如OpenAI、Anthropic、DashScope等），修正拼写或更新驱动包。
AgentArts.102012	组件在等待模型返回时接收到了来自运行环境的中断信号，可能是由于用户主动取消或工作流整体超时。	LLM执行被中断。	无需特别处理，如非预期中断，请排查前端是否存在误触取消操作或调大全局超时设置。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102050	在创建提示词变量对象时传入了错误的参数类型或配置项，导致变量引擎无法识别占位符。	变量初始化参数错误。	请检查模板中定义的变量属性（如Name、Type、Default Value），确保其符合变量定义的规范。
AgentArts.102051	提示词组装器在初始化时接收到的配置信息不完整，无法建立模板与变量之间的映射关系。	组装器初始化错误。	请核对组装器的核心配置，确保传入了有效的模板内容和变量列表。
AgentArts.102052	更新模板参数时，传入的键（Key）在模板定义中不存在，或者缺少了模板中要求的必填变量。	组装器键值对匹配错误	请根据模板定义的变量名核对输入数据。
AgentArts.102053	模板语法存在错误（如大括号未闭合、使用了非法转义字符），导致Jinja2或String格式化引擎无法渲染最终结果。	模板格式化错误。	请检查提示词文本中的语法格式，确保占位符使用正确且没有语法逻辑冲突。
AgentArts.102054	提示词管理组件在加载配置或连接后端存储时发生异常。	提示词组件初始化失败。	请检查组件的注册状态和配置文件路径，确保组件所需的底层资源已就绪。
AgentArts.102055	在执行提示词渲染或处理逻辑时发生非预期中断。	提示词组件调用失败。	请检查调用链中的变量传递过程，排查是否存在空指针或非法的逻辑跳转。
AgentArts.102056	为约束模型输出而定义的格式不符合标准规范。	格式定义非法。	请使用标准的校验工具检查结构定义，确保keywords（如properties、type）使用正确。
AgentArts.102101	尝试创建或导入模板时，系统检测到已存在相同名称或唯一标识符的模板，导致命名冲突。	提示词模板重复。	请更换模板名称或ID，或在更新现有模板时使用更新接口而非创建接口。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102102	从存储中检索到的模板数据格式损坏或关键字段（如内容、版本号）缺失，导致模板无法被正常解析。	模板数据异常。	请尝试重新保存该模板以修复受损的数据结构，或检查数据库层面的数据一致性。
AgentArts.102103	引擎尝试加载指定的模板，但在数据库或本地存储中未找到对应的记录。可能是由于模板ID错误、模板已被删除或环境迁移导致数据丢失	提示词模板不存在。	请核对模板ID是否正确，并确认该模板在当前租户或空间下是否已保存。
AgentArts.102104	系统在连接模板仓库时发生I/O错误或权限受限，导致无法读取模板列表。	模板存储加载失败。	请检查存储服务的运行状态（如MySQL/Redis），并验证服务账号的读取权限。
AgentArts.102150	大模型在执行“变异”指令（即对提示词进行同义改写或策略调整）时，未能保留原始模板中的变量占位符，导致新生成的提示词无法与业务参数匹配。	提示词变异操作失败。	请检查变异指令的提示词，要求模型在优化内容时必须严格保留所有的变量标签。
AgentArts.102151	大模型在执行“交叉”指令（即融合两个优秀提示词的优点）时，生成的混合模板丢失了父级模板中的共有占位符。	提示词交叉操作失败。	请确保参与交叉的父模板具有一致的占位符结构，并优化交叉引导词以强调结构完整性。
AgentArts.102152	在提示词评估或自优化过程中，测试用例中定义的变量占位符与原始模板不一致，导致数据无法正确注入。	测试用例占位符不匹配。	请核对测试用例的数据结构，确保其Key值与原始提示词模板中的变量名完全对应。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102153	测试用例中的消息模板内容与系统中存储的原始模板内容不符，无法建立有效的版本对应关系进行对比优化。	测试用例内容不匹配。	请刷新测试用例，确保其基于当前最新的提示词模板版本构建。
AgentArts.102154	启动优化任务时传入的配置参数（如迭代次数、样本比例或评估指标权重）不符合逻辑或超出范围。	提示词优化参数非法。	请根据接口文档核对优化参数的取值范围，并确保所有必填配置项均已正确填写。
AgentArts.102155	系统无法根据提供的Job ID检索到对应的优化任务记录，可能是任务已被物理删除或ID输入错误。	提示词优化任务不存在。	请确认任务ID的准确性，或在任务列表中重新发起优化申请。
AgentArts.102156	尝试执行的操作（如停止、重启）与任务当前所处的状态冲突（例如：无法重启一个正在运行中的任务）。	任务状态异常。	请先刷新任务状态，待当前流程结束后再尝试后续操作。
AgentArts.102157	在对新生成的提示词变体进行自动评分或效果对比时发生异常。可能是由于评估模型不可用或评估逻辑代码崩溃。	提示词优化评估失败。	请检查评估器的配置，并确保评估模型及其API Key状态正常。
AgentArts.102158	任务调度器在尝试拉起优化工作流时发生底层错误。通常涉及资源锁定、并发配额限制或初始化超时。	优化任务启动失败。	请稍后重试，或检查系统后端调度引擎的负载情况。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102159	尝试对已失败或已停止的任务进行断点续传或重跑时失败。	优化任务重启失败。	如果多次重启失败，建议直接克隆该任务并创建一个全新的优化作业。
AgentArts.102160	导入或执行优化任务时，测试用例的数据格式未能通过校验。	优化用例校验失败。	请按照指定的CSV/JSON模板格式重新整理测试用例数据，并注意字符编码规范。
AgentArts.102161	在优化评估阶段，系统无法从历史消息记录中提取出作为基准的预期结果，导致无法计算准确率或评分。	无法获取评估标签。	请检查测试数据集，确保每个测试用例都包含正确的预期输出。
AgentArts.102170	在保存优化中间过程、变体结果或评估报告时，数据库写入发生异常。	优化任务存储失败。	请检查后端存储服务的磁盘空间及连接池状态，确保系统具备持久化大规模评估数据的能力。
AgentArts.102201	请求的元模板标识符不在系统支持的预设列表内，无法匹配对应的生成逻辑。	不支持的元模板类型。	请核对元模板名称是否正确，或参考系统文档获取当前支持的元模板清单。
AgentArts.102202	尝试调用的元模板文件或配置在存储层缺失，可能已被删除或路径配置错误。	元模板不存在。	请确认元模板资源已部署至正确目录下，或联系管理员同步模板库。
AgentArts.102203	在解析用于生成提示词的工具描述（Tool/Function Definition）时发生错误，通常由于工具的JSON格式非法。	元模板工具解析失败。	请检查传入元模板的工具定义，确保其符合标准的OpenAPI或Function Calling规范。
AgentArts.102204	元模板渲染引擎在将基础元数据转化为最终提示词时发生内部崩溃或逻辑死锁。	元模板构建失败。	请排查元模板内容的复杂性，并检查系统内存资源是否足以支撑大规模模板构建。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.102205	调用元模板生成接口时，未提供该模板必需的控制参数。	元模板输入项缺失。	请根据元模板的要求补全输入参数，确保所有必填Key均已赋值。
AgentArts.102206	为元模板指定的下游推理模型类型尚未适配模板生成引擎。	LLM配置类型不支持。	请切换至主流的LLM类型（如OpenAI、Azure或系统内置驱动），或联系开发团队增加模型适配。
AgentArts.102207	在执行提示词自动生成的推理过程中，模型返回了错误信息或连接中断。	LLM生成结果失败。	请检查后端LLM服务的状态、配额以及请求超时设置。
AgentArts.102208	在对现有模板进行程序化修改或打补丁（Patching）时发生错误，通常与文件权限或并发锁定有关。	模板编辑失败。	请确认当前用户具备该模板的编辑权限，并检查是否有其他任务正在占用该资源。
AgentArts.102209	提示词引擎试图通过错误的存储协议访问数据。	存储类型不匹配。	请核对系统配置中的Storage Backend设置，确保与物理存储类型保持一致。
AgentArts.102211	提示词引擎触发了未预见的代码逻辑异常。	捕获到未知异常。	请将详细的错误堆栈反馈给技术支持团队，并尝试重启相关服务模块。
AgentArts.102213	在处理用户提供的针对提示词效果的反馈数据时发生异常。	优化反馈处理异常。	请检查反馈数据的输入参数结构，确保反馈ID与优化任务ID正确对应
AgentArts.102214	系统在尝试根据Badcase数据反向优化提示词时失败，可能是由于Badcase样本量不足或格式不兼容。	优化Badcase异常。	请增加Badcase样本的多样性，并确保每个样本都包含明确的错误分类标签。
AgentArts.103001	规划器组件的元数据或运行参数不完整，导致无法通过预执行检查。	规划器配置校验失败。	请检查规划器节点的设置界面，确保所有必填项已正确填写，且逻辑连接无误。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.103002	规划器未关联具体的提示词（Prompt）模板，导致引擎无法构造任务拆解指令。	提示词名称缺失。	请在规划器配置中选择一个有效的提示词模板，或在关联的提示词节点中指定名称。
AgentArts.103003	未为规划器指定执行推理的大语言模型（LLM），导致规划逻辑无法启动。	规划模型缺失。	请在规划器设置中关联一个已配置的模型服务。
AgentArts.103004	规划引擎在请求大模型进行任务规划时，收到错误响应或连接超时。	模型调用失败。	请检查后端模型服务的可用性、API Key有效性及网络连通性。
AgentArts.103005	系统无法根据当前的规划模式找到对应的状态管理类，通常由插件加载错误引起。	规划状态类获取失败。	请检查规划模式设置是否与当前引擎版本兼容，并确认相关扩展组件已安装。
AgentArts.103006	规划器尝试利用正则表达式从模型返回的文本中提取结构化步骤时失败，模型输出格式不符合预期。	正则解析失败。	请优化规划器的提示词，引导模型输出更加规范的格式，或调整正则表达式的匹配规则。
AgentArts.103007	规划引擎在尝试访问步骤列表或变量数组时，索引值超出了实际范围。	索引访问失败。	请排查前置逻辑是否产生了预期的列表数据，确保索引操作在合法范围内。
AgentArts.103008	规划器无法从全局配置中心读取所需的插件参数（如API密钥或Endpoint）。	插件配置获取失败。	请检查全局变量或插件中心的配置信息，确保规划器依赖的参数已正确注入。
AgentArts.103009	规划器中的条件判断表达式非法，导致无法得出布尔值（True/False）结果。	比较表达式解析失败。	请核对比较运算符（如==、>、<）的使用方式，确保比较的对象类型可兼容。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.103010	系统找不到对应规划模式的中间表示（IR）模板文件，导致无法生成可执行的执行图。	IR模板获取失败。	请确认规划模式配置正确，并检查系统资源目录中是否存在对应的.ir或配置文件。
AgentArts.103011	在将运行时变量注入规划提示词模板时发生错误，通常是由于变量占位符语法错误。	提示词格式化失败。	请检查提示词模板中的变量标签，确保其与传入的参数名称一一对应。
AgentArts.103012	规划器无法读取技能的详细配置参数。	技能配置获取失败。	请在技能管理页面检查该技能的定义，确保其配置参数已正确保存。
AgentArts.103013	规划引擎无法识别目标技能的调用类型（如HTTP、Python、Shell）。	技能类型获取失败。	请确认技能的元数据定义完整，且该类型技能在当前环境中已被支持。
AgentArts.103014	规划器在尝试调用子工作流时，无法识别该工作流的起始节点。	工作流起始组件解析失败。	请打开被调用的子工作流，确保其包含且仅包含一个合法的起始节点。
AgentArts.103015	规划器指派的工作流节点在执行过程中发生非预期中断或内部错误。	工作流组件运行失败。	请查看子工作流的执行日志，定位具体的异常节点并进行修复。
AgentArts.103016	系统尝试根据版本号加载规划模式的IR模板，但在存储库中未找到匹配的记录。可能是该版本已被移除或版本号不合法。	指定版本的IR模板获取失败。	请核对规划器配置中的版本信息，确保所引用的版本在模板库中已发布且处于可用状态。
AgentArts.103017	字段的参数类型不正确。	规划器配置类型校验失败。	请进入规划器节点的参数设置界面，根据提示的类型要求重新配置相关字段，确保数据格式合法。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.103018	规划引擎在执行流式输出（Streaming）或处理模型增量响应时发生非预期中断。	规划器流式处理异常。	请检查网络连接的稳定性，并确认模型服务端是否支持当前所选规划模式下的流式输出协议。
AgentArts.102301	在启动NL2Agent（自然语言生成Agent）流程时，传入的模型配置信息（如名称、版本、Endpoint）校验不通过。	模型信息参数非法。	请确保modelInfo中包含有效的模型标识及必要的认证信息。
AgentArts.104001	执行引擎在以流式（Stream）模式启动任务时，未能正确建立实时传输通道，或在传输过程中遭遇非预期的连接中断。	流式运行异常。	请检查前端与后端之间的WebSocket或SSE连接状态，确保中间件未拦截长连接。
AgentArts.104002	执行器在处理任务调度或线程分配时触发了底层系统异常。这通常涉及CPU调度超时或内核级指令错误。	执行器内部错误。	请尝试重启执行引擎服务，如果错误持续，请检查服务器系统日志排查资源竞争问题。
AgentArts.104003	执行器向外部服务（如数据库或第三方API）发起请求时发生未知错误，通常是由于网络抖动或请求格式不符合对方规范。	通用请求错误。	请检查网络防火墙配置，并验证执行器对外部资源的访问权限。
AgentArts.104004	执行引擎在运行特定的算子逻辑（如数据转换、数学运算节点）时抛出异常。可能是由于输入数据类型非法或算子内部算法崩溃。	算子执行失败。	请进入画布检查该节点的输入值格式，确保符合算子的业务逻辑要求。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.104005	分布式环境下的执行器无法从管理节点同步当前的Worker节点列表，导致任务无法正常分发。	从管理器获取节点列表失败。	请检查管理节点的服务状态，并确保执行器与管理器之间的RPC通信（如gRPC/HTTP）未被拦截。
AgentArts.104006	任务结束后，执行器尝试回收Session资源、删除临时文件或释放内存缓存时发生冲突。	会话清理失败。	请排查是否存在文件句柄被其他进程锁定，或增加会话清理的重试机制。
AgentArts.104007	执行器在处理用户实时注入的指令（如动态参数修改）时，由于指令校验未通过或执行环境不支持动态干预而报错。	用户输入执行失败。	请确保用户输入的参数符合预设的规范，且当前执行状态允许动态更新。
AgentArts.104008	执行引擎在解析来自模型的流式Token块或中间过程数据时发生序列化错误。	流式数据处理失败。	请检查模型输出的协议版本是否与执行引擎匹配，避免因不兼容的Stream格式导致解析中断。
AgentArts.109500	上下文引擎的基础配置设置不当	上下文配置错误。	请核对系统设置中的Context Policy，确保滑动窗口或摘要策略的参数在模型支持的范围内。
AgentArts.109501	上下文引擎在尝试初始化负责连接后端记忆存储的代理（Proxy）或客户端时失败。	记忆客户端初始化失败。	请检查后端存储服务（如Redis、Milvus或数据库）的连接地址、端口及鉴权凭证是否正确。
AgentArts.108003	在追踪调用链时，父节点下属的子节点执行顺序索引不存在。这会导致Insight无法正确还原执行拓扑图。	子节点执行索引缺失。	请检查执行引擎在下发子任务时是否正确生成了Sequence ID，并确认分布式链路追踪数据的完整性。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.108004	尝试查询特定执行记录时，系统无法找到该invoke_id。可能是数据已被清理，或者在异步写入数据库时发生了延迟。	调用ID不存在。	请确认查询的ID是否准确。如果是在任务刚结束时查询，请稍后重试以等待数据同步完成。
AgentArts.108005	Insight分析引擎预期处理的是提示词类型的调用数据。	调用类型不匹配。	请核对监控过滤器设置，确保请求的分析对象与实际节点的组件类型保持一致。
AgentArts.108006	Insight请求头中缺少必填的x-insight-project-id参数，导致无法对监控数据进行项目隔离。	项目ID缺失。	请在请求Header中补充项目ID，确保监控指标能归档到正确的业务组下。
AgentArts.999999	系统内部错误。	系统内部错误。	请联系技术支持进行定位。
AgentArts.109012	配置的策略路径在文件系统中不存在，导致动态加载器无法定位脚本文件。	策略路径不存在。	请检查路径拼写是否正确，并确认执行账户具备访问该目录的读取权限。
AgentArts.109013	提供的路径指向了一个目录而非具体的策略执行脚本（如.py或.yaml），导致解析中断。	无效的策略文件路径。	请明确指定策略脚本的完整文件名及其后缀。
AgentArts.109014	在注册中心完成initialize()过程前，尝试执行策略查找或调用。通常发生在系统启动阶段的竞态冲突。	策略注册中心未初始化。	请检查启动逻辑，确保所有的Provider和策略注册动作在主逻辑运行前已成功闭环。
AgentArts.109015	在指定的提供者下未找到规划策略。	策略未匹配。	请核对策略名称及关联的Provider标识符，确认该策略是否已在系统中注册并启用。

版权所有 © 华为云计算技术有限公司

错误码	报错信息	报错场景	排查与解决思路
AgentArts.109016	系统内置的核心策略文件缺失。这通常由于软件包安装不完整或版本更新导致的文件迁移。	内置策略加载失败。	请尝试重新部署引擎核心组件，确保builtin_strategies目录下的所有资源完整。
AgentArts.109017	策略组件在执行异步调用时触发了非预期的运行时错误。	策略异步调用异常。	排查策略内部的异步逻辑，检查是否存在未捕获的协程异常或事件循环冲突。
AgentArts.109018	智能体的最大迭代步数被设置为非正整数。这会防止模型进入死循环或无法启动任务。	Agent迭代次数配置非法。	请将max_iteration修改为大于0的整数。
AgentArts.109019	规划策略在尝试分发任务给特定插件时，发现该插件的调用URL为空。	插件地址缺失。	请检查插件管理模块中的Endpoint配置，确保该插件已正确部署并暴露了合法的API地址。
AgentArts.00071105	资产广场上架功能暂未开放。	预置模板操作权限不足。	资产广场上架功能暂未开放，如需使用，请**提交工单**申请。

12.2 更换密钥的加密方式

在AgentArts中，创建插件、MCP、连接第三方知识库或模型供应商时，需要对访问凭证进行加密存储，AgentArts提供了对访问凭证的加密方式。

服务端加密概述

如果您的业务对数据存储的安全性和合规性有较高要求，可使用AgentArts提供的服务端加密功能，对上传到AgentArts存储的数据进行加密保护。服务器端加密将对存储在AgentArts中的静态数据进行加密，即存储介质中存的是数据密文，防止存储介质（如硬盘等）被物理窃取或未授权访问时的数据泄露。在访问对象时，由AgentArts服务端对文件进行解密，然后返回解密后的数据。

AgentArts提供AgentArts加密和KMS加密方式。

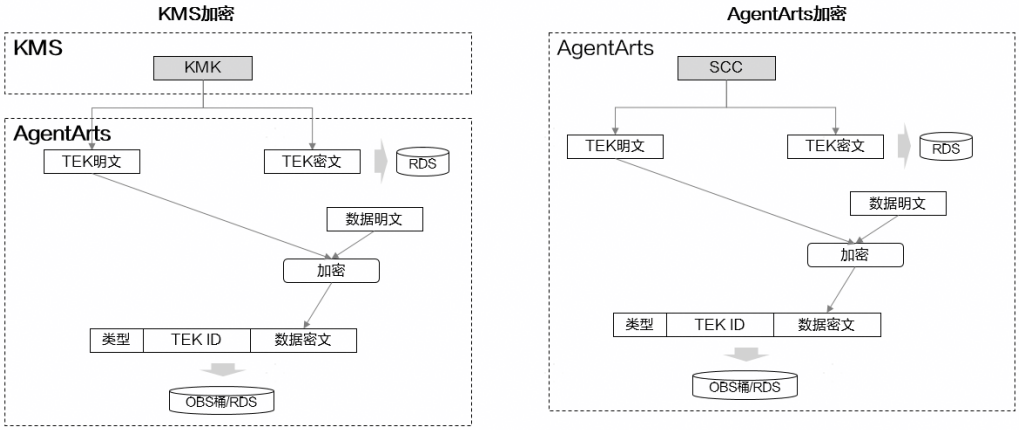
- AgentArts加密：免费，由AgentArts给每个租户生成一个租户主密钥，使用租户主密钥派生DEK。租户主密钥可以轮转，其作用类似于KMS的用户主密钥。
- KMS加密：使用用户选择的密钥派生DEK。

在AgentArts加密和KMS加密方式下，均对每个租户创建主密钥，主密钥可以版本化管理，加密时使用最新主密钥，解密时按在密文中记录的主密钥ID进行解密。可以强制下线一个主密钥，此时所有使用此主密钥加密的密钥均不能被自动解密，用于紧急情况下临时止损。

版权所有 © 华为云计算技术有限公司

- AgentArts加密密钥（TEK）：用于加密REK的密钥，可以在内存中缓存，相对数量较少，可以减少KMS使用次数。
- KMS主密钥（KMK）：用户选择的KMS主密钥，可以是默认密钥，也可以是自定义密钥。

图 12-1 加密架构图



更换密钥的加密方式

这里以更换模型供应商密钥的加密方式为例。

步骤1 执行**步骤1**~**步骤3**，在“新建模型供应商”页面，鼠标移动至“选择认证方式”右侧的 上，单击“更换加密方式”。

- **admin用户组**中的用户：可以直接修改加密方式。
- 非**admin用户组**中的用户：请联系租户修改加密方式或联系租户添加用户为admin用户组中的用户。

步骤2 在“加密方式”页面，参数配置请参考**表12-17**，单击“确定”。即可更换密钥的加密方式。

版权所有 © 华为云计算技术有限公司

图 12-2 更换加密方式



表 12-17 加密方式参数说明

参数		说明
加密方式	AgentArts加密	免费，由AgentArts给每个租户生成一个租户主密钥，使用租户主密钥派生DEK。租户主密钥可以轮转，其作用类似于KMS的用户主密钥。
	KMS加密	使用用户主密钥派生数据密钥，实现密钥自主管控、数据安全隔离，满足企业安全合规要求。
加密密钥类型	默认密钥	当“加密方式”值为“KMS加密”时，才有此参数。 默认密钥为agentarts/default，实例免费且每月提供2万次免费调用（支持缓存解密），超出部分按实际调用次数计费，详见**密码安全中心计费说明**。 默认值为kms-agentarts/default。

版权所有 © 华为云计算技术有限公司

参数		说明
	自定义密钥	当“加密方式”值为“KMS加密”时，才有此参数。 自定义密钥实例和使用均收费。 单击“查看KMS密钥”，在密码安全中心的“密钥管理”中创建密钥，创建密钥的操作请参考**创建自定义密钥**。返回“加密方式”页面，单击  刷新密钥列表，在下拉框中选择创建的密钥。

----结束

版权所有 © 华为云计算技术有限公司