盘古大模型

用户指南

文档版本 01

发布日期 2025-02-27

华为云计算技术有限公司

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 ModelArts Studio 大模型开发平台使用流程

盘古大模型服务简介

盘古大模型服务致力于深耕行业，打造多领域行业大模型和能力集。

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

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

预置模型使用流程

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

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

图 1-1 NLP 预置模型使用流程图

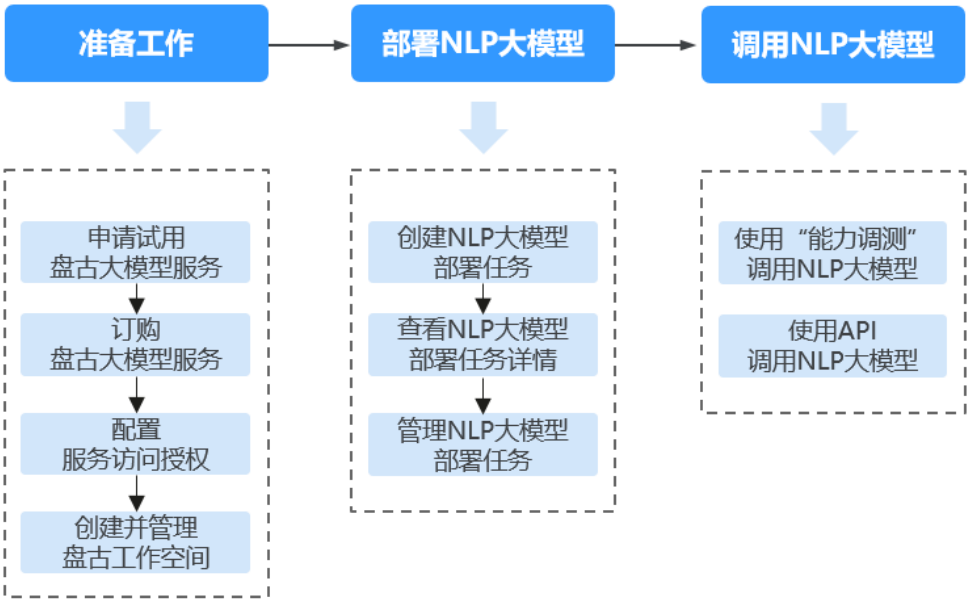


表 1-1 NLP 预置模型使用流程表

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

图 1-2 科学计算预置模型使用流程图

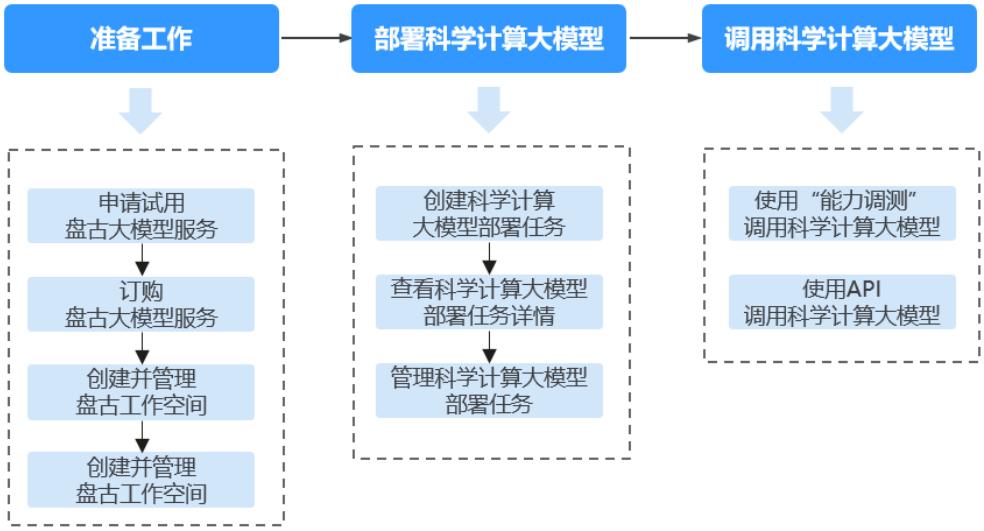


表 1-2 科学计算预置模型使用流程表

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

数据工程使用流程

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

数据工程使用流程见图1-3、表1-3。

图 1-3 数据工程使用流程图

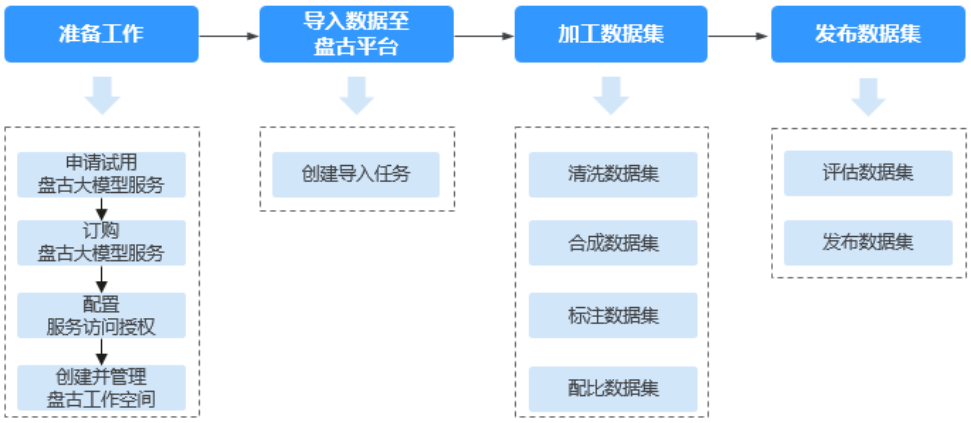


表 1-3 数据工程使用流程表

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

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

NLP 大模型开发流程

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

NLP大模型开发流程见图1-4、表1-4。

图 1-4 NLP 大模型开发流程图

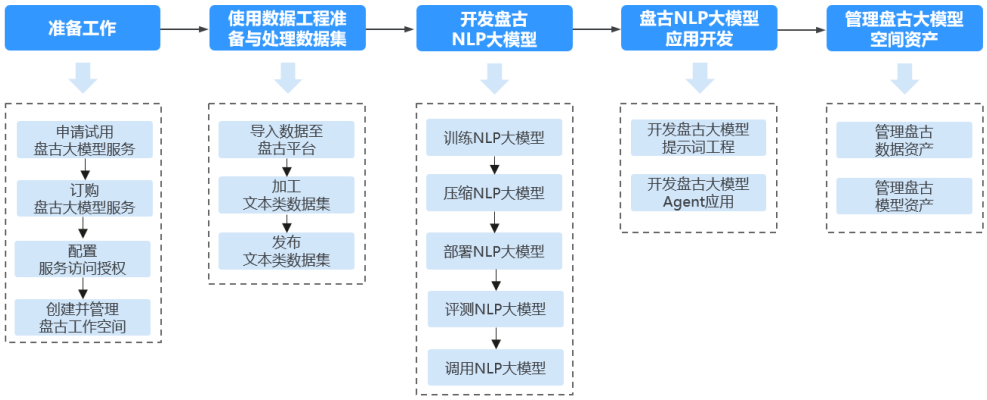


表 1-4 NLP 大模型开发流程表

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

流程	子流程	说明	操作指导
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产

科学计算大模型开发流程

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

科学计算大模型开发流程见图1-5、表1-5。

图 1-5 科学计算大模型开发流程图

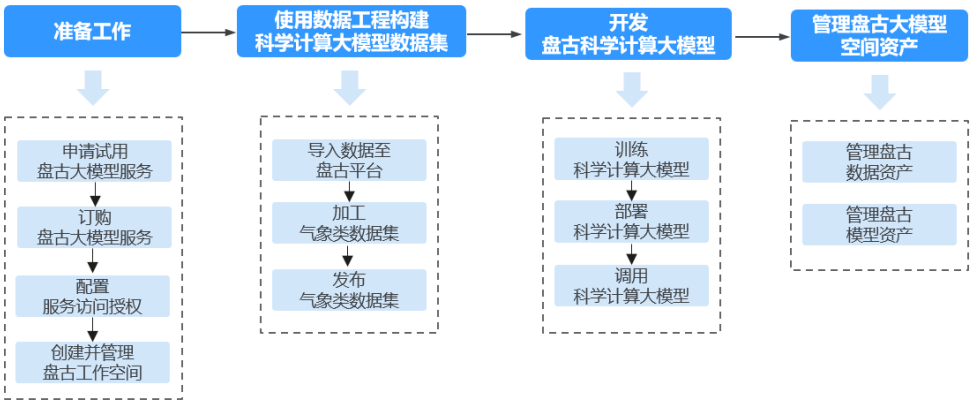


表 1-5 科学计算大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	申请试用盘古大模型服务
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	订购盘古大模型服务
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	配置服务访问授权
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	创建并管理盘古工作空间

流程	子流程	说明	操作指导
使用数据工程构建科学计算大模型数据集	导入数据至盘古平台	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
	加工气象类数据集	对气象类数据集进行加工操作。	加工气象类数据集
	发布气象类数据集	对气象类数据集进行发布操作。	发布气象类数据集
开发盘古科学计算大模型	训练科学计算大模型	进行模型的训练，如预训练、微调等训练方式。	训练科学计算大模型
	部署科学计算大模型	将模型部署用于后续模型的调用操作。	部署科学计算大模型
	调用科学计算大模型	支持“能力调测”功能与API两种方式调用大模型。	调用科学计算大模型
管理盘古大模型空间资产	管理盘古数据资产	管理从AI Gallery订阅或已发布的数据集。	管理盘古数据资产
	管理盘古模型资产	管理预置或训练后发布的模型。	管理盘古模型资产

专业大模型开发流程

ModelArts Studio大模型开发平台提供了专业大模型的部署功能。

专业大模型开发流程见图1-6、表1-6。

图 1-6 专业大模型开发流程图

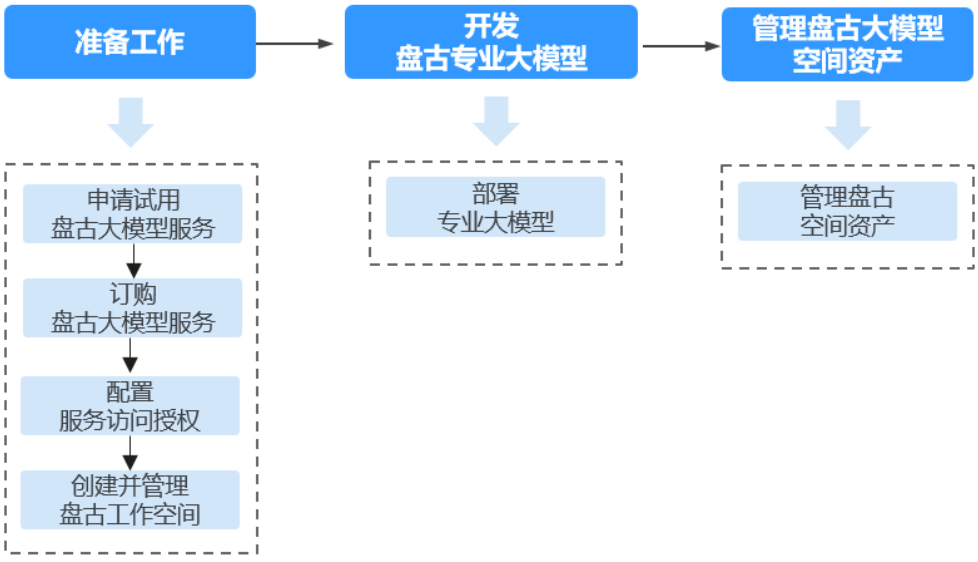


表 1-6 专业大模型开发流程表

流程	子流程	说明	操作指导
准备工作	申请试用盘古大模型服务	盘古大模型为用户提供了服务试用，用户可根据所需提交试用申请，申请通过后才可以试用盘古大模型功能。	**申请试用盘古大模型服务**
	订购盘古大模型服务	正式使用盘古大模型服务前，需要完成服务的订购操作。	**订购盘古大模型服务**
	配置服务访问授权	为了能够正常的存储数据、训练模型，需要用户配置盘古访问OBS的权限。	**配置服务访问授权**
	创建并管理盘古工作空间	平台支持用户自定义创建工作空间，并进行空间的统一管理。	**创建并管理盘古工作空间**
开发盘古专业大模型	部署专业大模型	支持专业大模型的部署操作。	**部署专业大模型**
管理盘古大模型空间资产	管理盘古模型资产	管理预置的专业大模型。	**管理盘古模型资产**

2 准备工作

2.1 申请试用盘古大模型服务

盘古大模型为用户提供了服务试用，需提交试用申请。

试用申请步骤如下：

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

图 2-1 申请试用

3. 填写姓名、联系电话等用户信息，单击“提交申请”进行表单预约。

2.2 订购盘古大模型服务

订购模型与资源

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

- 模型资产：**模型资产**可用于模型开发、应用开发等模块。当前支持订购NLP大模型、多模态大模型、CV大模型、预测大模型、科学计算大模型和专业大模型。
- 数据资源：**数据通算单元**适用于数据加工，用于正则类算子加工、**数据智算单元**适用于数据加工，用于AI类算子加工，**数据托管单元**适用于数据工程，用于存储数据集。
- 训练资源：**训练单元**可用于所有大模型的模型训练、模型压缩功能。

- 推理资源：**推理单元**可用于NLP、CV、专业大模型的模型推理功能， **模型实例**可用于预测、科学计算大模型的模型推理功能。

具体订购步骤如下：

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

图 2-2 选择开发场景

3. 在“模型资产”页面，参考**表2-1**完成模型资产的订购。

表 2-1 模型资产订购说明

模型分类	模型订阅	模型资产	计费方式
NLP大模型	盘古-NLP-N1-基模型 盘古-NLP-N1-基础功能模型	Pangu-NLP-N1-32K Pangu-NLP-N1-128K	包年/包月（1~9个月，包年为1年）
	盘古-NLP-N2-基模型 盘古-NLP-N2-基础功能模型	Pangu-NLP-N2-4K Pangu-NLP-N2-32K Pangu-NLP-N2-128K Pangu-NLP-N2-256K	包年/包月（1~9个月，包年为1年）
	盘古-NLP-N4-基模型 盘古-NLP-N4-基础功能模型	Pangu-NLP-N4-4K Pangu-NLP-N4-32K	包年/包月（1~9个月，包年为1年）
CV大模型	盘古-CV-基础模型	Pangu-CV-物体检测-N Pangu-CV-物体检测-S Pangu-CV-图像分类 Pangu-CV-语义分割	包年/包月（1~9个月，包年为1年）

模型分类	模型订阅	模型资产	计费方式
预测大模型	盘古-预测-模型	Pangu-Predict-Table-Cla Pangu-Predict-Table-Reg Pangu-Predict-Table-Anom Pangu-Predict-Table-TimSeries	包年/包月（1~9个月，包年为1年）
科学计算大模型	盘古-天气气象-基础版	Pangu-AI4S-Ocean_24h Pangu-AI4S-Weather_1h Pangu-AI4S-Weather_3h Pangu-AI4S-Weather_6h Pangu-AI4S-Weather_24h Pangu-AI4S-Weather_Precip	包年/包月（1~9个月，包年为1年）
	盘古-天气气象-专业版	Pangu-AI4S-Ocean_24h Pangu-AI4S-Weather_1h Pangu-AI4S-Weather_3h Pangu-AI4S-Weather_6h Pangu-AI4S-Weather_24h Pangu-AI4S-Weather_Precip Pangu-AI4S-Ocean_24h Pangu-AI4S-Ocean_Regional_24h Pangu-AI4S-Ocean_Ecology_24h Pangu-AI4S-Ocean_Swell_24h	包年/包月（1~9个月，包年为1年）

模型分类	模型订阅	模型资产	计费方式
专业大模型	盘古-NLP-N2-BI专业大模型	Pangu-NLP-BI-4K Pangu-NLP-BI-32K	包年/包月（1~9个月，包年为1年）

4. 参考**表2-2**，分别完成数据资源、训练资源和推理资源的订购。

表 2-2 资源订购说明

资源名称	订购项	适用场景	计费方式
数据资源	ModelArts Studio-数据托管单元	用于数据存储（包括数据集，prompt模板等）。	包年/包月
	ModelArts Studio-数据通算单元	适用于数据加工，用于正则类算子加工。 不同数据加工算子所需数据资源类型详见**数据集加工算子介绍**。	按需（时长）计费、包年/包月
	ModelArts Studio-数据智算单元	适用于数据加工，用于AI类算子加工。 不同数据加工算子所需数据资源类型详见**数据集加工算子介绍**。	按需（时长）计费、包年/包月
训练资源	ModelArts Studio-训练单元	用于所有模型的模型训练、模型压缩。	按需（时长）计费、包年/包月
推理资源	ModelArts Studio-推理单元（NLP、多模态、专业）	适用于NLP大模型、多模态大模型和BI专业大模型在基础平台和Agent开发的推理服务场景。	包年/包月
	ModelArts Studio-推理单元（CV）	适用于CV大模型在基础平台的推理服务场景。	包年/包月
	ModelArts Studio-模型实例（预测）	适用于预测大模型在基础平台的推理服务场景。	包年/包月

资源名称	订购项	适用场景	计费方式
	ModelArts Studio-模型实例（科学计算）	适用于气象大模型在基础平台的推理服务场景。	包年/包月

续订模型资产

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

续订模型资产的步骤如下：

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

退订模型资产

退订模型资产的步骤如下：

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

说明

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

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

增购模型资产

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

增购模型资产的步骤如下：

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

续订资源

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

续订资源的步骤如下：

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

说明

按需计费方式暂不支持续订。

退订资源

退订模型资产的步骤如下：

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

说明

退订后不影响已部署的服务，但退订之后将无法再选择该资源。

退订属于高危操作，请确保您已保存所有必要的数据和进度，以避免不必要的损失。

扩缩容资源

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

资源扩缩容的步骤如下：

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

说明

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

2.3 配置服务访问授权

配置 OBS 访问授权

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

配置OBS访问授权步骤如下：

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

图 2-3 配置 OBS 访问授权

- 方式2：单击首页右上角“设置”，在“授权管理”页签，单击“一键授权”。

2.4 创建并管理盘古工作空间

2.4.1 盘古工作空间介绍

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

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

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

2.4.2 创建并管理盘古工作空间

创建盘古工作空间

创建盘古工作空间步骤如下：

1. 登录**ModelArts Studio大模型开发平台**，在“我的空间”分页，单击“创建空间”。
2. 填写空间名称、描述，单击“确认”，完成空间的创建。

图 2-4 创建空间

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

   如果用户具备多个空间的访问权限，可在页面左上角单击 切换空间。

   图 2-5 切换空间

管理盘古工作空间

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

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 单击左侧导航栏的“空间管理”，在“空间设置”页签可执行如下操作：
   - 修改当前空间名称与描述。
   - 可查看当前空间的创建时间。
   - 单击右上角“删除”，可删除当前空间。

     说明

     删除空间属于高危操作，删除前请确保当前空间不再进行使用。

2.4.3 管理盘古工作空间成员

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

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

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

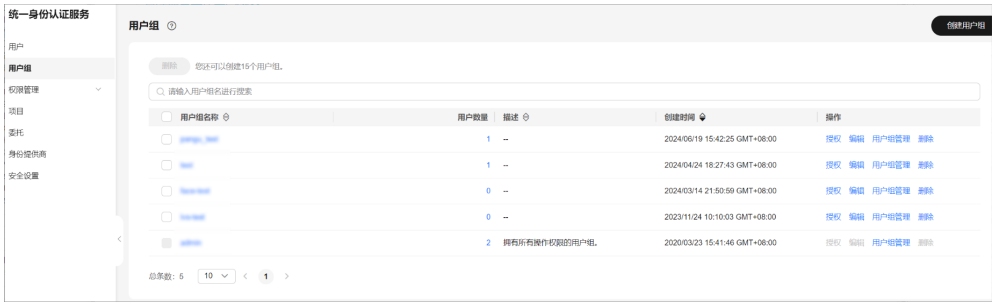
创建用户组

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

创建用户组的步骤如下：

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

图 2-6 创建用户组



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

图 2-7 用户组授权

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

表 2-3 授权项

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

图 2-8 添加用户组权限

6. 设置最小授权范围。

   根据授权项策略，系统会自动推荐授权范围方案。

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

   选择完成后，单击“确定”。

图 2-9 设置最小授权范围

7. 单击“完成”，完成用户组授权。

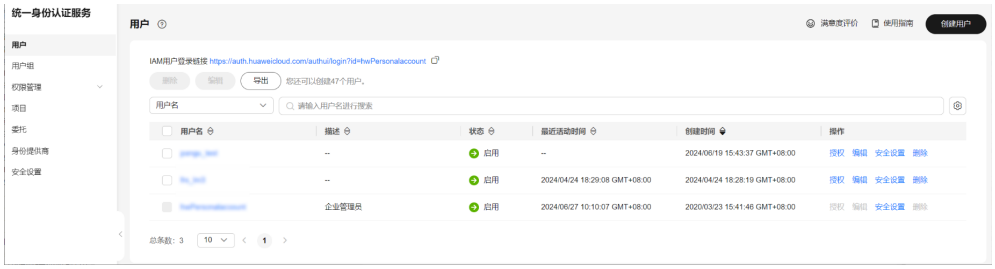
图 2-10 完成授权

创建盘古子用户

创建盘古子用户步骤如下：

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

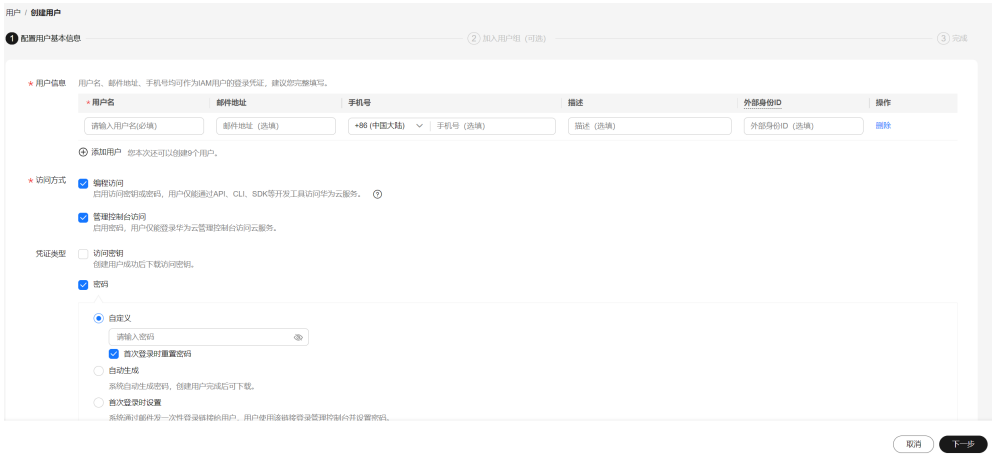
图 2-11 创建用户



3. 配置用户基本信息，单击“下一步”。

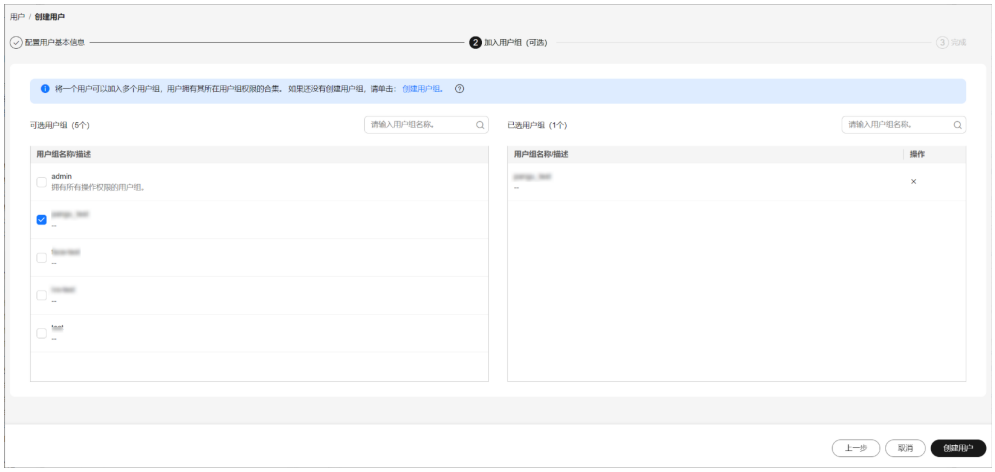
   配置用户信息时，需要勾选“编程访问”，如果未勾选此项，会导致IAM用户无法使用盘古服务API、SDK。

图 2-12 配置用户基本信息



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

图 2-13 加入用户组

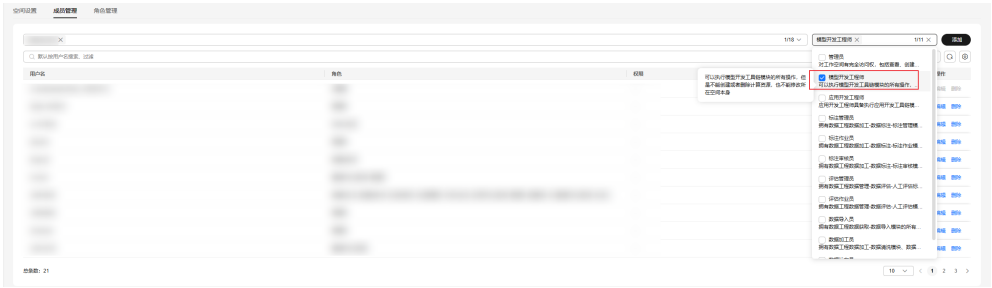


添加盘古子用户至工作空间

在添加盘古子用户至工作空间前，请先完成**创建盘古子用户**。

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

图 2-14 添加成员为“模型开发工程师”角色

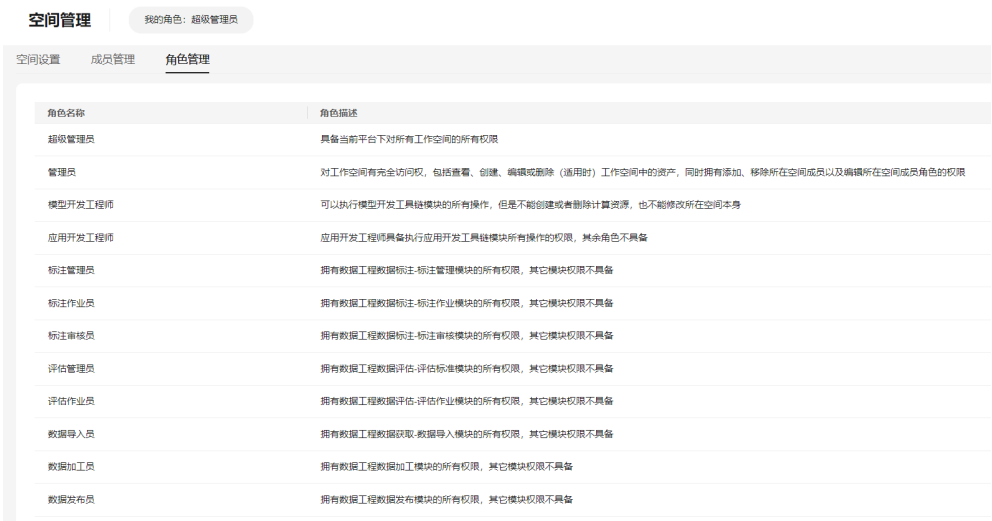


修改盘古子用户权限

当需要修改空间内某个子用户权限时可以按如下步骤操作：

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

图 2-15 角色管理



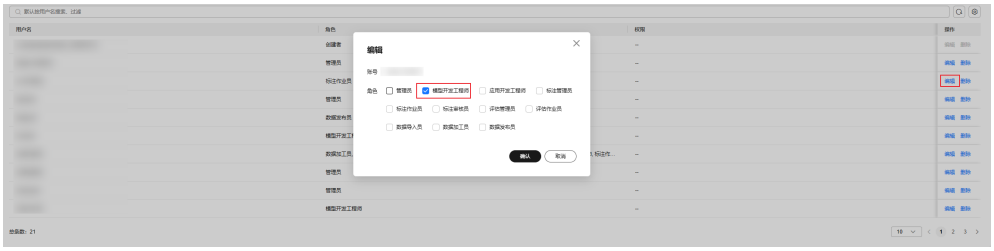
空间管理　我的角色：超级管理员

空间设置　成员管理　角色管理

角色名称	角色描述
超级管理员	具备当前平台下对所有工作空间的所有权限
管理员	对工作空间有完全访问权，包括查看、创建、编辑或删除（适用时）工作空间中的资产，同时拥有添加、移除所在空间成员以及编辑所在空间成员角色的权限
模型开发工程师	可以执行模型开发工具链模块的所有操作，但是不能创建或者删除计算资源，也不能修改所在空间本身
应用开发工程师	应用开发工程师具备执行应用开发工具链模块所有操作的权限，其余角色不具备
标注管理员	拥有数据工程数据标注-标注管理模块的所有权限，其它模块权限不具备
标注作业员	拥有数据工程数据标注-标注作业模块的所有权限，其它模块权限不具备
标注审核员	拥有数据工程数据标注-标注审核模块的所有权限，其它模块权限不具备
评估管理员	拥有数据工程数据评估-评估标准模块的所有权限，其它模块权限不具备
评估作业员	拥有数据工程数据评估-评估作业模块的所有权限，其它模块权限不具备
数据导入员	拥有数据工程数据获取-数据导入模块的所有权限，其它模块权限不具备
数据加工员	拥有数据工程数据加工模块的所有权限，其它模块权限不具备
数据发布员	拥有数据工程数据发布模块的所有权限，其它模块权限不具备

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

图 2-16 授权子用户“模型开发工程师”权限



移除盘古子用户

当需要删除空间内某个子用户时，可以按如下步骤操作：

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

图 2-17 成员管理

4. 单击“确定”进行二次确认，即可删除空间子用户。

图 2-18 删除操作二次确认

3 使用数据工程构建数据集

3.1 数据工程介绍

数据工程介绍

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

数据工程包含的具体功能如下：

- **数据获取**：数据获取是数据工程的第一步，支持将不同来源和格式的数据导入平台，并生成“原始数据集”。
  - 支持的接入方式：通过OBS服务导入数据。
  - 支持的数据类型：文本、图片、视频、气象、预测、其他。

  通过这些功能，用户可以轻松将大量数据导入平台，为后续的数据加工和模型训练等操作做好准备。
- **数据加工**：平台提供了数据加工、数据合成、数据标注、数据配比的加工操作，旨在确保原始数据能够满足各种业务需求和模型训练的标准，生成“加工数据集”。
  - 数据加工：数据加工旨在通过使用数据集加工算子对数据进行预处理操作，针对不同类型的数据集，平台设计了专用的加工算子，以确保数据符合模型训练的标准和业务需求。
  - 数据合成：数据合成利用预置或自定义的数据指令对原始数据集进行处理，并根据设定的轮数生成新的数据。
  - 数据标注：数据标注旨在为无标签的数据集添加准确的标签，标注数据的质量直接影响模型的训练效果和精度。针对不同数据集平台支持人工标注与AI预标注两种形式。

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

  通过数据加工操作，平台能够有效清理噪声数据、标准化数据格式，提升数据集的整体质量。

- **数据发布：**平台提供了数据评估、数据发布操作，旨在通过数据质量评估确保数据满足大模型训练的多样性、平衡性和代表性需求，并促进数据的高效流通与应用，生成“发布数据集”。
  - 数据评估：数据评估通过对数据集进行系统的质量检查，依据评估标准评估数据的多个维度，旨在发现潜在问题并加以解决。
  - 数据发布：将单个数据集发布为特定格式的“发布数据集”的过程，用于后续模型训练等操作。

    支持发布的数据集格式为标准格式、盘古格式（适用于训练盘古大模型时）。目前，仅文本类和图片类数据集支持发布为“盘古格式”。

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

平台支持的数据类型

ModelArts Studio大模型开发平台支持的数据类型见**表3-1**，各类型数据格式详细要求请参考**数据集格式要求**。

表 3-1 平台支持的数据类型

数据类型	数据内容	支持的文件格式
文本类	文档	txt、mobi、epub、docx、pdf
	网页	html
	预训练文本	jsonl
	单轮问答	jsonl、csv
	单轮问答（人设）	jsonl、csv
	多轮问答	jsonl
	多轮问答（人设）	jsonl
	问答排序	jsonl、csv
	偏好优化DPO	jsonl
	偏好优化DPO（人设）	jsonl
图片类	仅图片	jpg、jpeg、png、bmp、tar包
	图片+Caption	• 图片格式支持：jpg、jpeg、png、bmp，所有图片需保存为tar包。 • Caption格式支持：jsonl

数据类型	数据内容	支持的文件格式
	图片+QA对	• 图片格式支持：jpg、jpeg、png、bmp，所有图片需保存为tar包。 • QA对格式支持：jsonl
	物体检测	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：xml
	图像分类	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：txt
	异常检测	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：txt
	语义分割	jpg、png
	姿态估计	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：json
	实例分割	• 图片格式支持：jpg、jpeg、png、bmp • 标注格式支持：xml
	变化检测	• 图片格式支持：jpg、jpeg、bmp • 标注格式支持：png
	视频分类	图片格式支持：jpg、jpeg、png、bmp
视频类	视频	mp4、avi
	事件检测	• 视频格式支持：mp4、avi，每个视频时长大于128s，FPS>=10 • 标注格式支持：json
气象类	气象数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2
预测类	时序	csv
	回归分类	csv
其他类	自定义	支持构建用户自定义场景下所需的数据集类型。

各类数据支持的操作

各类型数据支持的数据工程操作见**表3-2**。

表 3-2 各类数据支持的操作

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

3.2 数据工程使用流程

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

数据工程操作流程见图3-1、表3-3。

图 3-1 数据集构建流程图

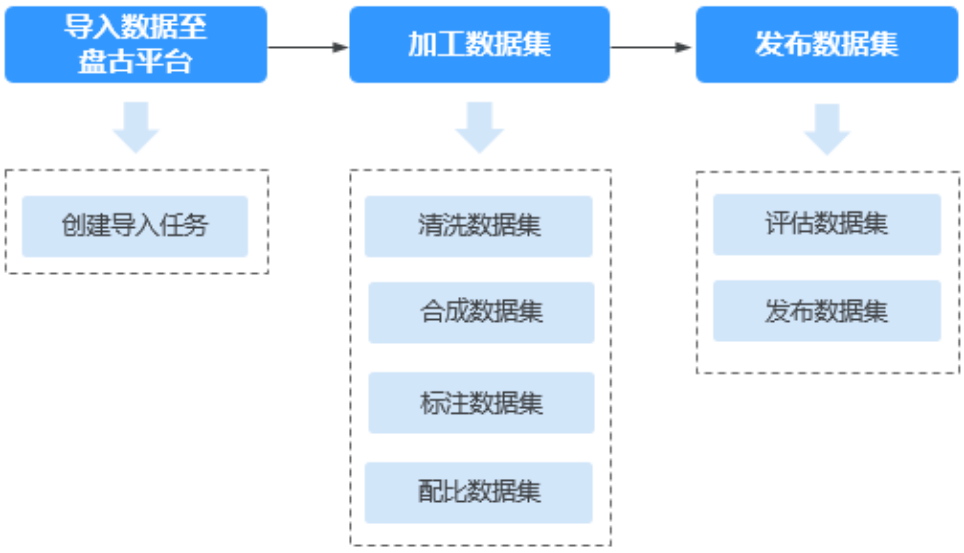


表 3-3 数据集构建流程表

流程	子流程	说明
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。

流程	子流程	说明
加工数据集	加工数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。
	合成数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。
	标注数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。
	配比数据集	数据配比是将多个数据集按特定比例组合并生成为“加工数据集”的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。
发布数据集	评估数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。
	发布数据集	数据发布是将数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**标准格式、盘古格式**。 • 标准格式：平台默认的格式。该格式的数据集可发布为资产，但不可应用于盘古大模型的开发中。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。

3.3 数据集格式要求

3.3.1 文本类数据集格式要求

ModelArts Studio大模型开发平台支持创建文本类数据集，创建时可导入多种形式的数据，具体格式要求详见**表3-4**。

表 3-4 文本类数据集格式要求

文件内容	文件格式	文件要求
文档	txt、mobi、epub、docx、pdf	单个文件大小不超过50GB，文件数量最多1000个。
网页	html	单个文件大小不超过50GB，文件数量最多1000个。

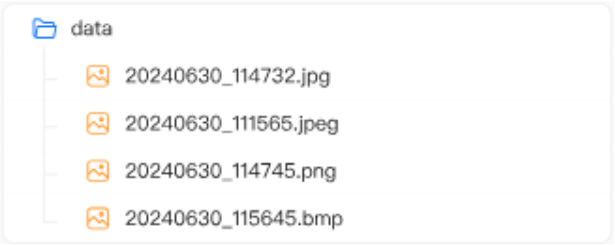
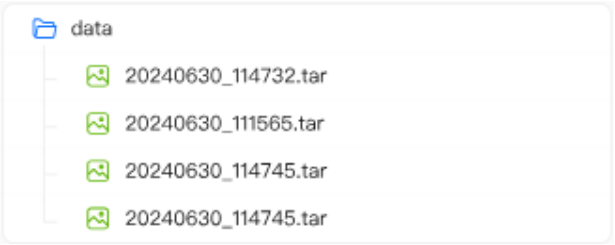
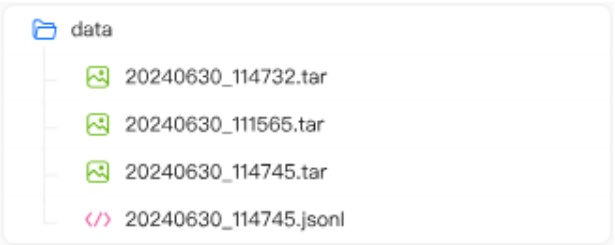
文件内容	文件格式	文件要求
预训练文本	jsonl	• jsonl格式：text表示预训练所使用的文本数据，具体格式示例如下： {"text":"盘古大模型，是华为推出的盘古系列AI大模型，包括NLP大模型、多模态大模型、CV大模型、科学计算大模型、预测大模型。"} • 单个文件大小不超过50GB，文件数量最多1000个。
单轮问答	jsonl、csv	• jsonl格式：数据由问答对构成，context、target分别表示问题、答案，具体格式示例如下： {"context": "你好，请介绍自己", "target": "我是盘古大模型"} • csv格式：csv文件的第一列对应context，第二列对应target，具体格式示例如下： "你好，请介绍自己","我是盘古大模型" • 单个文件大小不超过50GB，文件数量最多1000个。
单轮问答（人设）	jsonl、csv	• jsonl格式：system表示人设，context、target分别表示问题、答案。 {"system":"你是一个机智幽默问答助手","context":"你好，请介绍自己","target":"哈哈，你好呀，我是你的聪明助手。"} • csv格式：csv文件的第一列对应system，第二三列分别对应context、target。 "你是一个机智幽默问答助手","你好，请介绍自己","哈哈，你好呀，我是你的聪明助手。" • 单个文件大小不超过50GB，文件数量最多1000个。
多轮问答	jsonl	• jsonl格式：数组格式，至少由一组问答对构成。形式为[{"context":"context内容1","target":"target内容1"},{"context":"context内容2","target":"target内容2"}]，其中context、target分别表示问题、答案。 [{"context":"你好","target":"你好，请问有什么可以帮助你的？"},{"context":"请介绍一下华为云的产品。","target":"华为云提供包括但不限于计算、存储、网络等产品服务。"}] • 单个文件大小不超过50GB，文件数量最多1000个。
多轮问答（人设）	jsonl	• jsonl格式：数组格式，至少由一组问答对构成。system表示人设，context、target分别表示问题、答案。 [{"system":"你是一位书籍推荐专家"},{"context":"你好","target":"嗨！你好，需要点什么帮助吗？"},{"context":"能给我推荐点书吗？","target":"当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"}] • 单个文件大小不超过50GB，文件数量最多1000个。

文件内容	文件格式	文件要求
问答排序	jsonl、csv	• jsonl格式：context表示问题，targets答案1、2、3表示答案的优劣顺序，最好的答案排在最前面。 { "context":"context内容","targets":["回答1","回答2","回答3"]} • csv格式：csv文件的第一列对应context，其余列为答案。 "问题","回答1","回答2","回答3" • 单个文件大小不超过50GB，文件数量最多1000个。
偏好优化DPO	jsonl	• jsonl格式：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。 单轮问答 {"context": ["你好，请介绍自己"], "target":"我是盘古大模型", "bad_target":"我不会回答"} 多轮问答 {"context": ["你好，请介绍自己", "我是盘古大模型", "请介绍一下有哪些产品。"], "target":"提供包括但不限于计算、存储、网络等产品服务。", "bad_target":"我不会回答"} • 单个文件大小不超过50GB，文件数量最多1000个。
偏好优化DPO（人设）	jsonl	• jsonl格式：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。 带人设单轮 {"system": "你是一位机制幽默的问答助手", "context": ["你好，请介绍自己"], "target":"哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "bad_target":"我不会回答"} 带人设多轮 {"system": "你是一位机制幽默的问答助手", "context": ["你好，请介绍自己", "哈哈，你好呀，我是你的聪明助手，怎么帮到你？ ", "请介绍一下有哪些产品。"], "target":"我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！ ", "bad_target":"我不会回答"} • 单个文件大小不超过50GB，文件数量最多1000个。

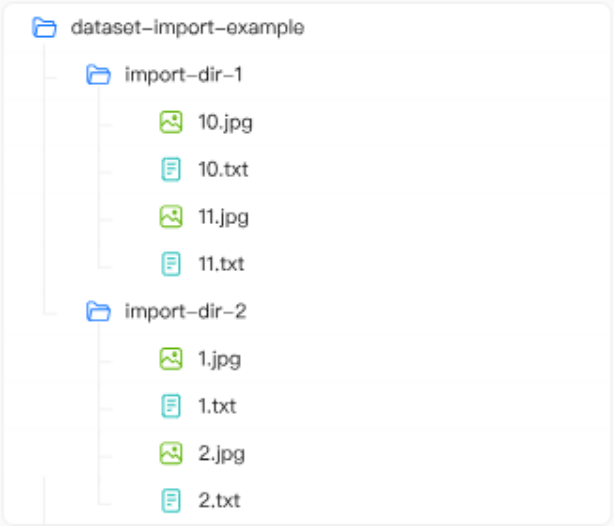
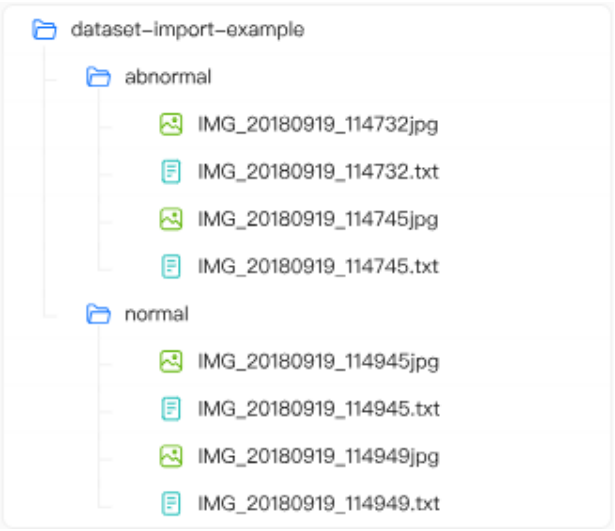
3.3.2 图片类数据集格式要求

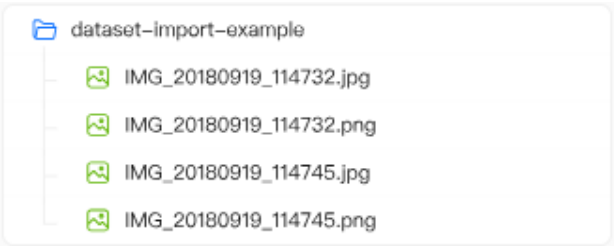
ModelArts Studio大模型开发平台支持创建图片类数据集，创建时可导入多种形式的数据，具体格式要求详见**表3-5**。

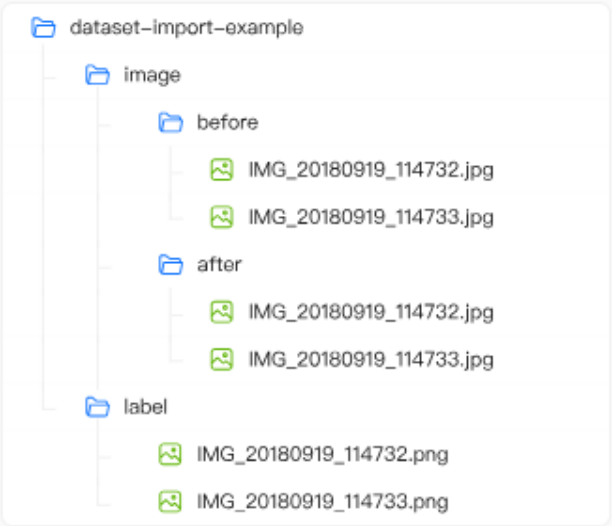
表 3-5 图片类数据集格式要求

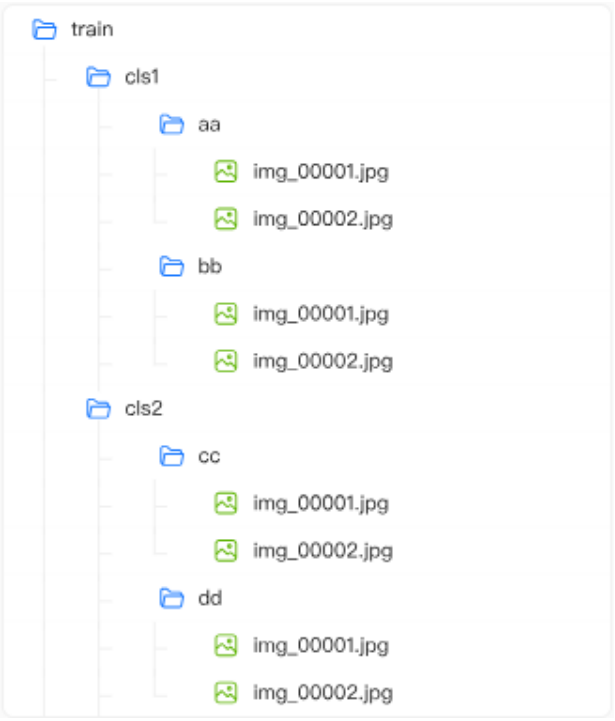
文件内容	文件格式	文件要求
仅图片	tar、图片目录	• 图片：支持jpg、jpeg、png、bmp类型。  • tar：tar包内图片支持jpg、jpeg、png、bmp图片类型。  • 单个文件大小不超过50GB，单个压缩包大小不超过50GB，文件数量最多1000个。
图片+Caption	图片支持tar，Caption支持jsonl	• 图片：图片以tar包格式存储，可以多个tar包。tar包存储原始的图片，每张图片命名要求唯一（如abc.jpg）。图片支持jpg、jpeg、png、bmp格式。 • jsonl：图片描述jsonl文件放在最外层目录，一个tar包对应一个jsonl文件，文件内容中每一行代表一段文本，形式为： `{"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）","caption":"图片对应的文本描述"}` • 单个文件大小不超过50GB，单个压缩包大小不超过50GB，文件数量最多1000个。 

文件内容	文件格式	文件要求
图片+QA对	图片支持tar，QA对支持jsonl	• 图片：图片以tar包格式存储，可以多个tar包。tar包存储原始的图片，每张图片命名要求唯一（如abc.jpg）。图片支持jpg、jpeg、png、bmp格式。 • jsonl：图片描述jsonl文件放在最外层目录，一个tar包对应一个jsonl文件，文件内容中每一行代表一段文本，形式为： {"image_name":"图片名称（abc.jpg）","tar_name":"tar包名称（1.tar）","conversations":[{"question":"问题1","answer":"回答1"},{"question":"问题2","answer","回答2"}]} • 单个文件大小不超过50GB，单个压缩包大小不超过50GB，文件数量最多1000个。 
物体检测	PASCAL VOC	• 由图片文件和对应的标注文件构成，标注文件需要满足PASCAL VOC文件格式。要求用户将标注对象和标注文件存储在同一目录，并且相互对应，如标注对象文件名为“IMG_2.jpg”，那么标注文件的文件名应为“IMG_2.xml” • 图片支持jpg、jpeg、png、bmp格式，标注文件为xml格式，标注文件说明请参见**物体检测数据集标注文件说明**。 • 单个文件大小不超过50GB，文件数量最多1000个。 

文件内容	文件格式	文件要求
图像分类	图片+txt	• 由图片文件和对应的标注文件构成，要求用户将标注对象和标注文件存储在同一目录，并且相互对应。 • 图片支持jpg、jpeg、png、bmp格式，标注文件为txt格式，标注文件说明请参见**图像分类数据集标注文件说明**。 • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示： 
异常检测	图片+txt	• 文件存放方式要求满足异常检测格式，即标注文件和图片存于同一文件夹，正常和异常分文件夹创建。 • 图片支持jpg、jpeg、png、bmp格式，标注文件为txt格式，标注文件说明请参见**异常检测数据集标注文件说明**。 • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示： 

文件内容	文件格式	文件要求
语义分割	图片+png	• 文件存放方式要求满足语义分割格式，即原图为jpg文件，标注图采用同名同尺寸的png文件。其中，标注图上的每个像素值对应原图中像素的类别，且每个类别的值需连续且从0开始，表示不同的物体或区域类别。例如，假设有一张原图为IMG_20180919_114732.jpg，对应的标注图为IMG_20180919_114732.png，其中标注图的不同像素值代表不同的类别，标注图的每个像素值直接对应原图中相应位置的类别信息，类别需连续并从0开始。 • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示： 
姿态估计	图片+json	• 由图片文件和对应的标注文件构成，图片支持jpg、jpeg、png、bmp格式，标注文件为json格式。 • 基于开源COCO人物关键点标注格式对数据集进行标注，需包含annotations，train，val文件夹，annotations文件夹下用train.json和val.json记录训练集和验证集标注，train和val文件夹下保存具体的图片。 • json标注文件的详细说明请参见**姿态估计标注json文件说明** • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示： 

文件内容	文件格式	文件要求
实例分割	图片+xml	• 文件存放方式要求满足万物分割/实例分割格式。 • 图片格式支持：jpg、jpeg、png、bmp，标注格式支持：xml • 基于PASCAL VOC矩形框格式进行标识，标注和图片同名并放在同一文件夹下。 • xml标注文件的详细说明请参见**实例分割数据集标注文件说明**。 • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示： 
变化检测	图片+png	• 文件存放方式要求满足变化检测格式。图片格式支持：jpg、jpeg、bmp，标注格式支持：png • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示：  其中，**before**文件夹：包含变化前的图片，每幅图片需与变化后的图片同名、同尺寸。 **after**文件夹：包含变化后的图片，每幅图片需与变化前的图片同名、同尺寸。 **label**文件夹：包含与变化前和变化后图片同名、同尺寸的PNG文件。每个像素值代表该位置对应的类别信息，类别应是连续的且从0开始。

文件内容	文件格式	文件要求
视频分类	图片	• 导入为目录，每个数据集目录下有train和test两个目录，目录结构一致，train目录下是多个二级目录，每个二级目录代表相应的类别，例如cls1表示类别1。 • 单个文件大小不超过50GB，文件数量最多1000个，示例如下所示：  其中，单个**cls类别目录**下的每个三级目录为一个样本，例如cls1文件的样本为aa和bb。 所有**样本文件夹（如aa）**包含的图片数量相等，例如cls1样本aa和bb、cls1样本aa和cls2的样本cc。 每个**样本文件夹（如aa）**可以视为一个视频片段，其中每张图片代表视频的一个帧，将这些帧作为一个序列来学习视频分类，有助于模型学习视频的时序特征，从而进行准确的分类。

物体检测数据集标注文件说明

该说明适用于表3-5中的物体检测标注文件格式。

物体检测数据集支持格式为ModelArts PASCAL VOC 1.0。

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

物体检测的标注文件需要满足PASCAL VOC格式，PASCAL_VOC是一个公开的图像标注数据集，它提供了一个统一的XML格式来存储标注信息。PASCAL_VOC文件格式包含

图像目录、图像文件、图像尺寸、图像中目标信息等元素，详细格式说明请参见**表3-6**。

OBS文件上传示例：

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

标注文件（.xml文件）示例：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
   <folder>NA</folder>
   <filename>bike_1_1593531469339.png</filename>
   <source>
      <database>Unknown</database>
   </source>
   <size>
      <width>554</width>
      <height>606</height>
      <depth>3</depth>
   </size>
   <segmented>0</segmented>
   <object>
      <name>Dog</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <occluded>0</occluded>
      <bndbox>
         <xmin>279</xmin>
         <ymin>52</ymin>
         <xmax>474</xmax>
         <ymax>278</ymax>
      </bndbox>
   </object>
   <object>
      <name>Cat</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <occluded>0</occluded>
      <bndbox>
         <xmin>279</xmin>
         <ymin>198</ymin>
         <xmax>456</xmax>
         <ymax>421</ymax>
      </bndbox>
   </object>
</annotation>
```

表 3-6 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示图像所在的目录名称。
filename	是	被标注文件的文件名。

字段	是否必选	说明
size	是	表示图像的像素信息。 • width：必选字段，图像的宽度。 • height：必选字段，图像的高度。 • depth：必选字段，图像的通道数。
segmented	是	表示是否用于分割，取值为0或1。0表示没有分割标注，1表示有分割标注。
object	是	目标对象信息，包括被标注物体的类别、姿态、是否被截断、是否识别困难以及边界框信息，多个物体标注会有多个object体。 • name：必选字段，标注内容的类别。 • pose：必选字段，标注内容的拍摄角度。 • truncated：必选字段，取值0或1，表示标注内容是否被截断（0表示被截断、1表示没有截断）。 • occluded：必选字段，取值0或1，表示标注内容是否被遮挡（0表示未遮挡、1表示遮挡） • difficult：必选字段，取值0或1，表示标注目标是否难以识别（0表示容易识别、1表示难易识别）。 • confidence：可选字段，标注目标的置信度，取值范围0-1之间，越接近1，表示标注越可信。 • bndbox：必选字段，标注框的类型，可选值请参见表3-7。

表 3-7 标注框类型描述

type	形状	标注信息
point	点	点的坐标。 <x>100<x> <y>100<y>
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>

type	形状	标注信息
bndbox	矩形框	左上和右下两个点坐标。 <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	圆形	圆心坐标和半径。 <cx>100<cx> <cy>100<cy> <r>50<r>

图像分类数据集标注文件说明

该说明适用于**表3-5**中的图片分类标注文件格式。

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

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

- 当目录下存在对应的txt文件时，以txt文件内容作为图像的标签。
  示例如下所示，import-dir-1和import-dir-2为导入子目录。

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

```
    2.jpg
    2.txt
```

单标签的标签文件示例，如1.txt文件内容如下所示。

```
猫
```

多标签的标签文件示例，如2.txt文件内容如下所示。

```
猫
狗
```

异常检测数据集标注文件说明

该说明适用于**表3-5**中的异常检测标注文件格式。

要求用户将标注文件和图片存于同一文件夹，正常和异常分文件夹创建。

- 当目录下存在对应的txt文件时，以txt文件内容作为正常或异常的标签。
  示例如下所示，import-dir-1和import-dir-2为导入子目录。

```
dataset-import-example
├─abnormal
│      IMG_20180919_114732.jpg
│      IMG_20180919_114732.txt
│      IMG_20180919_114745.jpg
│      IMG_20180919_114745.txt
└─normal
│      IMG_20180919_114945.jpg
│      IMG_20180919_114945.txt
│      IMG_20180919_114949.jpg
│      IMG_20180919_114949.txt
```

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

```
abnormal
```

正常标签的标签文件示例，如IMG_20180919_114945.txt文件内容如下所示。

```
normal
```

姿态估计标注 json 文件说明

该说明适用于**表3-5**中的姿态估计标注文件格式。

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

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

具体的json标注文件具体示例：

```
{
   "images": [
      {
         "license": 2,
         "file_name": "000000000139.jpg",
         "coco_url": "",
         "height": 426,
         "width": 640,
         "date_captured": "2013-11-21 01:34:01",
         "flickr_url": "",
         "id": 139
```



```
    }
  ],
  "annotations": [
    {
      "num_keypoints": 15,
      "area": 2913.1104,
      "iscrowd": 0,
      "keypoints": [
        427,
        170,
        1,
        429,
        169,
        2,
        0,
        0,
        0,
        434,
        168,
        2,
        0,
        0,
        0,
        441,
        177,
        2,
        446,
        177,
        2,
        437,
        200,
        2,
        430,
        206,
        2,
        430,
        220,
        2,
        420,
        215,
        2,
        445,
        226,
        2,
        452,
        223,
        2,
        447,
        260,
        2,
        454,
        257,
        2,
        455,
        290,
        2,
        459,
        286,
        2
      ],
      "image_id": 139,
      "bbox": [
        412.8,
        157.61,
        53.05,
        138.01
      ],
      "category_id": 1,
      "id": 230831
```



```
    },
  ],
  "categories": [
    {
      "supercategory": "person",
      "id": 1,
      "name": "person",
      "keypoints": [
        "nose",
        "left_eye",
        "right_eye",
        "left_ear",
        "right_ear",
        "left_shoulder",
        "right_shoulder",
        "left_elbow",
        "right_elbow",
        "left_wrist",
        "right_wrist",
        "left_hip",
        "right_hip",
        "left_knee",
        "right_knee",
        "left_ankle",
        "right_ankle"
      ],
      "skeleton": [
        [
          16,
          14
        ],
        [
          14,
          12
        ],
        [
          17,
          15
        ],
        [
          15,
          13
        ],
        [
          12,
          13
        ],
        [
          6,
          12
        ],
        [
          7,
          13
        ],
        [
          6,
          7
        ],
        [
          6,
          8
        ],
        [
          7,
          9
        ],
        [
          8,
```



```
            10
          ],
          [
            9,
            11
          ],
          [
            2,
            3
          ],
          [
            1,
            2
          ],
          [
            1,
            3
          ],
          [
            2,
            4
          ],
          [
            3,
            5
          ],
          [
            4,
            6
          ],
          [
            5,
            7
          ]
        ]
      }
    ]
}
```

表 3-8 COCO 格式说明

字段	是否必选	说明
images	是	图片信息。
license	否	图像的许可证标识符。
file_name	是	图像的文件名。
coco_url	否	图像在COCO官方数据集中的URL。
height	是	图像的高度，以像素为单位。
width	是	图像的宽度，以像素为单位。
date_captured	否	图像捕获的日期和时间。
flickr_url	否	图像在Flickr网站上的URL。
id	是	图像的唯一标识符。
annotations	是	标注信息。
num_keypoints	是	标注的关键点数量。

字段	是否必选	说明
area	是	边界框的面积，以像素平方为单位。
iscrowd	是	表示标注是否为复杂的群体场景（如拥挤的人群）。0表示不是拥挤场景，1表示是拥挤场景。
keypoints	是	标注的关键点坐标及其可见性，按顺序列出所有关键点，每个关键点用三个数值表示 [x, y, v]。x和y是关键点的像素坐标，v是可见性（0：不可见且不在图像中；1：不可见但在图像中；2：可见且在图像中）。
image_id	是	与该标注相关联的图像的ID，必须与images字段中的id对应。
bbox	是	目标物体的边界框，用[x, y, width, height]表示，其中，x，y是边界框左上角的坐标，width和height是边界框的宽度和高度。
category_id	是	标注类别的ID，对于人体姿态估计，通常为1（表示person）。
id	是	标注的唯一标识符。
categories	是	标注类型信息。
supercategory	是	类别的上级分类，通常为person。
id	是	类别的唯一标识符，对于人体姿态估计，通常为1。
name	是	类别的名称，通常为person。
keypoints	是	关键点的名称列表，COCO格式中通常定义了17个关键点，如nose、left_eye、right_eye、left_ear、right_ear、left_shoulder、right_shoulder、left_elbow、right_elbow、left_wrist、right_wrist、left_hip、right_hip、left_knee、right_knee、left_ankle、right_ankle。
skeleton	是	定义骨架连接的列表，用于表示关键点之间的连接关系。每个连接用一对关键点索引表示，如 [1, 2]，表示鼻子（nose）到左眼（left_eye）的连线。

实例分割数据集标注文件说明

该说明适用于**表3-5**中的实例分割标注文件格式。

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

实例分割的标注文件需要满足PASCAL VOC格式，PASCAL_VOC是一个公开的图像标注数据集，它提供了一个统一的XML格式来存储标注信息。PASCAL_VOC文件格式包含图像目录、图像文件、图像尺寸、图像中目标信息等元素，详细格式说明请参见**表 3-9**。

OBS文件上传示例：

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

标注文件（.xml文件）示例：

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

表 3-9 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示图像所在的目录名称。
filename	是	被标注文件的文件名。

字段	是否必选	说明
size	是	表示图像的像素信息。 ● width：必选字段，图像的宽度。 ● height：必选字段，图像的高度。 ● depth：必选字段，图像的通道数。
segmented	是	表示是否用于分割，取值为0或1。0表示没有分割标注，1表示有分割标注。
object	是	目标对象信息，包括被标注物体的类别、姿态、是否被截断、是否识别困难以及边界框信息，多个物体标注会有多个object体。 ● name：必选字段，标注内容的类别。 ● pose：必选字段，标注内容的拍摄角度。 ● truncated：必选字段，取值0或1，表示标注内容是否被截断（0表示被截断、1表示没有截断）。 ● occluded：必选字段，取值0或1，表示标注内容是否被遮挡（0表示未遮挡、1表示遮挡） ● difficult：必选字段，取值0或1，表示标注目标是否难以识别（0表示容易识别、1表示难易识别）。 ● confidence：可选字段，标注目标的置信度，取值范围0-1之间，越接近1，表示标注越可信。 ● bndbox：必选字段，标注框的类型，可选值请参见**表3-10**。

表 3-10 标注框类型描述

type	形状	标注信息
point	点	点的坐标。 <x>100<x> <y>100<y>
line	线	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>

type	形状	标注信息
bndbox	矩形框	左上和右下两个点坐标。 <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	圆形	圆心坐标和半径。 <cx>100<cx> <cy>100<cy> <r>50<r>

3.3.3 视频类数据集格式要求

ModelArts Studio大模型开发平台支持创建视频类数据集，创建时可导入多种形式的数据，具体格式要求详见**表3-11**。

表 3-11 视频类数据集格式要求

文件内容	文件格式	文件要求
视频	mp4或avi	• 支持mp4、avi视频格式上传，所有视频可以放在多个文件夹下，每个文件夹下可以同时包含mp4或avi格式的视频。 • 单个文件大小不超过50GB，文件数量最多1000个。

<table>
<tr><th>文件内容</th><th>文件格式</th><th>文件要求</th></tr>
<tr><td>事件检测</td><td>视频+json</td><td>数据源样本为avi、mp4格式，标注文件为json格式。必须包含两个及以上后缀名字为avi或者mp4的文件。
每个视频时长要大于128s，FPS>=10，且测试集训练集都要有视频。
支持视频的格式包括常见的mp4/avi格式文件，每个视频时长要大于128s，FPS>=10，用annotation.json对文件进行标注。
单个文件大小不超过50GB，文件数量最多1000个，示例如下所示：
 
具体的json标注文件参考：
<pre>{
  'version': 'dataset_name_v.x.x',// 数据集版本信息。
  'classes': [category1',category2', ...],// 所有类别名称的列表，每个类别对应一个 label，用于标注视频中的事件或动作。
  'database': {
    'video_name':{
      // 训练集 train 测试集 test。
      'subset': 'train',
      'duration': 1660.3, // 视频总时长 seconds。
      'fps': 30.0,// 视频帧率。
      'width': 720,// 视频宽度，单位像素。
      'height': 1280,// 视频高度，单位像素。
      'ext': 'mp4',//视频文件扩展名。
      // 标注 34.5, 42.4 分别表示起始时间和结束时间，单位为s。
      // label 表示分类，必须是classes列表中的一个元素，表示该视频片段对应的事件或动作类型。
      'annotations': [
        {'label': 'category1', 'segment': [34.5, 42.4]},
        {'label': 'category1', 'segment': [124.4, 142.9]},
        ...
      ]
    },
    'video_name':{
      'subset': xxx,//视频文件名称，不包括扩展名。
      'duration': xxx,
      'fps': xxx,
      'width': xxx,
      'height': xxx,
      'ext': xxx,
      'annotations': [
        {'label': xxx, 'segment': xxx},
        {'label': xxx, 'segment': xxx},
        ...
      ]
    },
    ...
  }
}</pre></td></tr>
</table>

3.3.4 气象类数据集格式要求

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

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

再分析数据为二进制格式，具体格式要求详见**表3-12**。

表 3-12 气象类数据集格式要求

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

文件内容	文件格式	文件样例
气象-海洋数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	● 海洋数据通常包含全球或区域性的海洋变量，如温度（T）、气压（P）、风速（U、V）等，具体格式示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SSH", "T", "P", "U", "V"], "under_sea_layers": ["0m", "6m", "10m", "20m", "30m", "50m", "70m", "100m", "125m", "150m", "200m", "250m", "300m", "400m", "500m"], "under_sea_features": ["T", "U", "V", "S"]} – geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 – time_range：数据的时间范围，时间戳格式为毫秒数。 – total_size：数据文件的总大小，单位为字节。 – surface_features：海表特征变量列表，例如海表高度（SSH）、温度（T）、风速（U、V）。 – under_sea_layers：深海层列表，例如500m、400mPa等。 – under_sea_features：高空特征变量列表，例如海盐（S）、温度（T）、海流速率（U、V）。 ● 单个文件大小不超过50GB，文件数量最多1000个。
气象-生态数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	生态数据通常包含总叶绿素浓度（Tca）、叶绿素浓度（Chl）、硅藻浓度（Dia）等生态变量。示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["Tca “, ” Chl “, ” Dia “, ” Coc “, ” Cya “, ” Irn “, ” Nit “, ” MLD “]} ● geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 ● time_range：数据的时间范围，时间戳格式为毫秒数。 ● total_size：数据文件的总大小，单位为字节。 ● surface_features：生态特征列表，例如总叶绿素浓度（Tca）、叶绿素浓度（Chl）、硅藻浓度（Dia）。 ● 单个文件大小不超过50GB，文件数量最多1000个。
气象-海浪数据	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	海浪数据通常包有效波高（SWH）。示例如下： {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SWH"]} ● geo_range：定义了数据覆盖的地理范围，纬度（lat）从-90.0到90.0，经度（lon）从0.0到360.0。 ● time_range：数据的时间范围，时间戳格式为毫秒数。 ● total_size：数据文件的总大小，单位为字节。 ● surface_features：海浪特征：有效波高（SWH）。 ● 单个文件大小不超过50GB，文件数量最多1000个。

3.3.5 预测类数据集格式要求

平台支持创建预测类数据集，创建时可导入**时序数据、回归分类数据**。

- 时序数据：时序预测数据是一种按时间顺序排列的数据序列，用于预测未来事件或趋势，过去的数据会影响未来的预测。
- 回归分类数据：回归分类数据包含多种预测因子（特征），用于预测连续变量的值，与时序数据不同，回归分类数据不要求数据具有时间顺序。

具体格式要求详见**表3-13**。

表 3-13 预测类数据集格式要求

文件内容	文件格式	文件样例
时序	csv	• 数据为结构化数据，包含列和行，每一行表示一条数据，每一列表示一个特征，并且必须包含预测目标列，预测目标列要求为连续型数据。 • 目录下只有1个数据文件时，文件无命名要求。 • 目录下有多个数据文件时，需要通过命名的方式指定数据是训练数据集、验证数据集还是测试数据集。训练数据名称需包含train字样，如train01.csv；验证数据名称需包含eval字样；测试数据名称需包含test字样。文件的命名不能同时包含train、eval和test中的两个或三个。 • 时序预测必须要包含一个时间列，时间列值的格式示例为2024-05-27 或 2024/05/27 或 2024-05-27 12:00:00 或 2024/05/27 12:00:00 。 示例如下： timestamp,feature1,feature2,target 2024-05-27 12:00:00,10.5,20.3,100 2024-05-27 12:01:00,10.6,20.5,101 2024-05-27 12:02:00,10.7,20.7,102 2024-05-27 12:03:00,10.8,20.9,103 2024-05-27 12:04:00,10.9,21.0,104 • 单个文件大小不超过50GB，文件数量最多1000个。

文件内容	文件格式	文件样例
回归 分类	csv	• 数据为结构化数据，包含列和行，每一行表示一条数据，每一列表示一个特征，并且必须包含预测目标列，预测目标列要求为连续型数据。 • 目录下只有1个数据文件时，文件无命名要求。 • 目录下有多个数据文件时，需要通过命名的方式指定数据是训练数据集、验证数据集还是测试数据集。训练数据名称需包含train字样，如train01.csv；验证数据名称需包含eval字样；测试数据名称需包含test字样。文件的命名不能同时包含train、eval和test中的两个或三个。 示例如下： feature1,feature2,target 10.5,20.3,100 10.6,20.5,101 10.7,20.7,102 10.8,20.9,103 10.9,21.0,104 • 单个文件大小不超过50GB，文件数量最多1000个。

3.3.6 其他类数据集格式要求

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

其他类数据集支持**发布其他类数据集**操作。

其他类数据集要求单个文件大小不超过50GB，单个压缩包大小不超过50GB，文件数量最多1000个。

3.4 导入数据至盘古平台

数据集是一组用于处理和分析的相关数据样本。

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

创建导入任务

创建导入任务前，请先按照**数据集格式要求**提前准备数据。

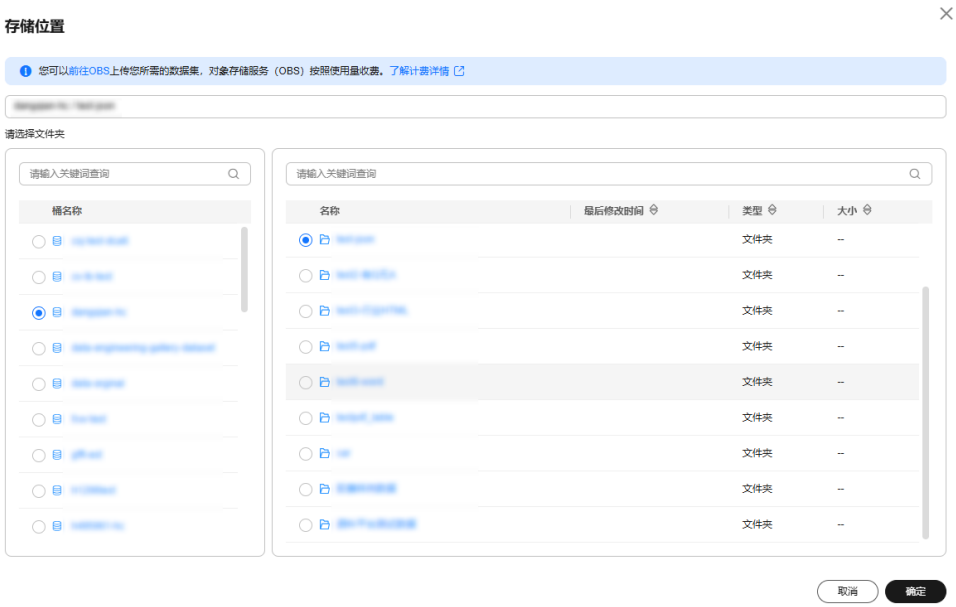
平台支持使用OBS服务导入数据，请详见**通过控制台快速使用OBS**。

创建导入任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”，单击空间名称进入操作空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取 > 导入任务”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面，选择“数据集类型”、“文件格式”和“导入来源”。

4. 单击“  ”，在“存储位置”弹窗中选择需导入的数据，单击“确定”。

图 3-2 选择导入的数据



5. 填写“数据集名称”和“描述”，可选择填写“扩展信息”。

   扩展信息包括“数据集属性”与“数据集版权”：

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

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

说明

如果任务状态为“运行失败”，可能由以下原因导致：

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

管理原始数据集

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

1. 登录ModelArts Studio大模型开发平台，单击进入操作空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取 > 原始数据集”，单击需要查看的数据集名称。

- 查看数据集基本信息。在“基本信息”页签，可以查看数据详情、数据来源以及扩展信息。
- 下载原始数据集。在“数据预览”页签，可以查看数据内容，单击右上角“下载”即可下载原始数据集。
- 查看数据血缘。在“数据血缘”页签，可以查看当前数据集所经历的完整操作，如加工、标注等。
- 查看操作记录。在“操作记录”页签，可以查看当前数据集的操作记录，如创建该数据集的时间、状态、操作人员等。

3. 删除原始数据集。单击操作列的“删除”，并在弹窗中单击“确定”。

说明

删除“原始数据集”属于高危操作，删除前，请确保该数据集不再使用。

3.5 加工数据集

3.5.1 数据集加工场景介绍

数据加工介绍

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

- **数据加工**

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

- **数据合成**

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

- **数据标注**

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

- **数据配比**

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

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

数据加工意义

数据加工在大模型开发中具有至关重要的作用，具体体现在以下几个方面：

- **提高数据质量**

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

- **扩展数据集的多样性和泛化能力**

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

- **增强模型训练的有效性**

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

- **确保业务需求对接**

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

- **提升数据处理效率**

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

- **确保数据质量和适配性**

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

- **提高数据的多样性和代表性**

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

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

支持数据加工的数据集类型

当前支持数据加工操作的数据集类型见**表3-14**。

表 3-14 支持数据加工操作的数据集类型

数据类型	数据加工	数据合成	数据标注	数据配比
文本类	√	√	√	√
图片类	√	-	√	√
视频类	√	-	√	-
气象类	√	-	-	-

数据类型	数据加工	数据合成	数据标注	数据配比
预测类	-	-	-	-
其他类	√（仅可使用自定义算子）	-	-	-

3.5.2 数据集加工算子介绍

3.5.2.1 文本类加工算子介绍

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

平台支持文本类数据集的加工操作，分为数据提取、数据转换、数据过滤、数据打标四类，文本类加工算子能力清单见**表3-15**。

表 3-15 文本类加工算子能力清单

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

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

算子分类	算子名称	算子描述
数据过滤	异常字符过滤	查找数据集每一条数据中携带的异常字符，并将异常字符替换为空值，数据条目不变。 ● 不可见字符，比如U+0000-U+001F。 ● 表情符□□。 ● 网页标签符号<style></style>。 ● 特殊符号，比如● ■ ◆ 。 ● 乱码和无意义的字符□□□□□。 ● 特殊空格：[\u2000-\u2009]
	自定义正则过滤	删除符合自定义正则表达式的数据。
	自定义关键词过滤	剔除包含关键词的数据。
	敏感词过滤	对文本中涉及黄色、暴力、政治等敏感数据进行自动检测和过滤。
	文本长度过滤	按照设置的文本长度，保留长度范围内的数据进行。
	冗余信息过滤	按照段落粒度，删除文本中的冗余信息，不改变数据条目。 例如图注表注和参考文献。
	N-gram特征过滤	用于判断文档重复度，根据特征N值计算文档内词语按N值组合后的重复此时，可通过以下两种算法比较结果是否大于特征阈值，大于特征阈值的文档删除。 ● top-gram过滤：计算**重复最多**的gram占总长度的比例，大于特征阈值则删除。 ● gram重复率过滤：计算**所有重复**的gram占总长度的比例，大于特征阈值则删除。
	段落特征过滤	根据如下特征过滤： ● 段落重复率。 ● 重复段落长度占比。 ● 非中文字符占比。
	句子特征过滤	该算子将文档中的标点符号作为句子分隔符，统计每句字符长度，若文档平均字符长度大于设置字符，则保留，反之则删除整篇文档。根据如下特征过滤： ● 待保留的平均句长。

算子分类	算子名称	算子描述
	词语特征过滤	词个数表示按照系统词库，对文档进行分词，分词后统计词的总个数，平均词长度为所有词的长度总和除以词总个数，两者都满足则保留当前文档。根据如下特征过滤： ● 待保留的词个数。 ● 待保留的平均词长度。
	段落结尾不完整句子过滤	按照句子的过滤粒度，自动识别段落结尾处的内容是否完整，如果不完整，则过滤。
	广告数据过滤	按照句子的过滤粒度，删除文本中包含广告数据的句子。
	QA对过滤	过滤包含以下情况的QA对： ● 问题不是string格式。 ● 回答为空。 ● 回答无意义。
	语种过滤	通过语种识别模型得到文档的语言类型，筛选所需语种的文档。
	全局文本去重	检测并去除数据中重复或高度相似的文本，防止模型过拟合或泛化性降低。
数据打标	预训练文本分类	针对预训练文本进行内容分类，例如新闻、教育、健康等类别，支持分析语种包括：中文、英文。
	通用质量评估	针对文本进行通用质量的评估，例如流畅度、清晰度、丰富度等。
	问题时效性评估	判断问题是否具有时效性，并给出判断原因。
	回答质量评分	针对微调数据集的回答进行质量评分，例如逻辑连贯性、事实正确性等。
	语法质量评估	针对文本进行语法质量的评估，例如相关性、规范性等。

说明

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

3.5.2.2 视频类加工算子介绍

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

平台支持视频类数据集的加工操作，分为数据提取、数据过滤、数据打标三类，视频类加工算子能力清单见**表3-16**。

表 3-16 视频类加工算子能力清单

算子分类	算子名称	算子描述
数据提取	镜头拆分	根据视频中的镜头场景变化将长视频拆分为短视频片段，如果某个镜头片段的长度超过设定的时间阈值，该镜头片段将按时长进行进一步拆分。
数据过滤	视频裁剪	裁剪视频中字幕/Logo/水印/黑框等无用信息，生成新视频。
	视频元数据过滤	基于视频元数据进行过滤，包括帧率、分辨率和视频时长。注：电影标准帧率为24或30FPS。
	宽高比过滤	根据视频的宽高比进行过滤。
数据打标	视频鉴黄评分	对视频的涉黄程度进行评分，分数越高越危险。评分范围(0, 100)，评分≥50分的视频可视为涉黄视频。
	视频暴恐评分	对视频的暴恐程度进行评分，分数越高越危险。评分范围(0, 100)，评分≥50分的视频可视为暴恐视频。
	视频涉政评分	对视频的涉政程度进行评分，分数越高越危险。评分范围(0, 100)，评分≥90分的视频可视为涉政视频。
	运动幅度评分	通过计算每个像素在每一帧中的移动范围进行评分，识别运动幅度过快（如＞100光流）或过慢（如≤2光流）的视频，数值越大表示运动过快。
	质量基础评分	对视频的基础质量（清晰度、亮度、模糊、画面抖动重影、低光过曝、花屏等）进行评分。分值范围(0, 1)，数值越高质量越好，评分>0.05可认为是视频基础质量较高的视频。
	美学评分	从内容（吸引人，清晰度）、构图（目标物位置良好）、颜色（有活力，令人愉悦）、光线（光线明显有对比度）、轨迹（连续、稳定）等维度评价视频美感得分。分值范围(0, 1)，数值越高美感越好，评分>0.95可视为视频基础质量较高的视频。
	水印识别	识别视频中是否包含水印。
	字幕识别	识别视频中是否包含字幕。
	Logo识别	识别视频中是否包含Logo。
	视频黑边识别	识别视频中是否包含黑边。
	密集文字识别	识别视频中是否包含密集文字，达到密集文字面积占比的视频则为含密集文字视频，一般裁剪面积占比≥7%为密集文字视频。

3.5.2.3 图片类加工算子介绍

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

平台提供了图文类、图片类加工算子，算子能力清单见表3-17。

表 3-17 图片类加工算子能力清单

算子分类	算子名称	算子描述
数据提取	图文提取	提取图文压缩包中的JSON文本和图片，并对图片进行结构化解析（BASE64编码）。
数据过滤	图片元数据过滤	基于图片存储大小、宽高比属性进行图片/图文数据加工。
	图文文本长度过滤	过滤文本长度不在“文本长度范围”内的图文对。一个中文汉字或一个英文字母，文本长度均计数为1。
	图文文本语言过滤	通过语种识别模型得到图文对的文本语种类型，“待保留语种”之外的图文对数据将被过滤。
	图文去重	• 基于结构化图片去重 • 判断相同文本对应不同的图片数据是否超过阈值，如果超过则去重。
	图片去重	通过把图片结构化处理后，过滤重复的图片/图文对数据。
数据打标	图片鉴黄评分	对图片的涉黄程度进行评分，分数越高越危险。默认评分不小于50分的视频可视为涉黄视频。
数据转换	图文异常字符过滤	将文本数据中携带的异常字符替换为空值，数据条目不变。 • 不可见字符，例如U+0000-U+001F • 表情符□□ • 网页标签符号<p> • 特殊符号，比如● ◆ • 乱码和无意义的字符□□□□□

3.5.2.4 气象类加工算子介绍

平台支持气象类数据集的加工操作，气象类加工算子能力清单见表3-18。

表 3-18 气象类加工算子能力清单

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

3.5.3 加工文本类数据集

3.5.3.1 加工文本类数据集

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

创建文本类数据集加工任务步骤如下：

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

   图 3-3 算子编排

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

   若选择使用加工模板，将删除当前已编排的加工步骤。

图 3-4 选择加工模板

5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。
   当数据加工任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成加工。

说明

在完成数据加工后，如果无需使用数据标注、数据合成功能，可直接在“加工任务”页面单击操作列“生成”，生成加工数据集。

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

3.5.3.2 合成文本类数据集

当前，数据合成功能支持合成单轮问答、单轮问答（人设）类型的数据。

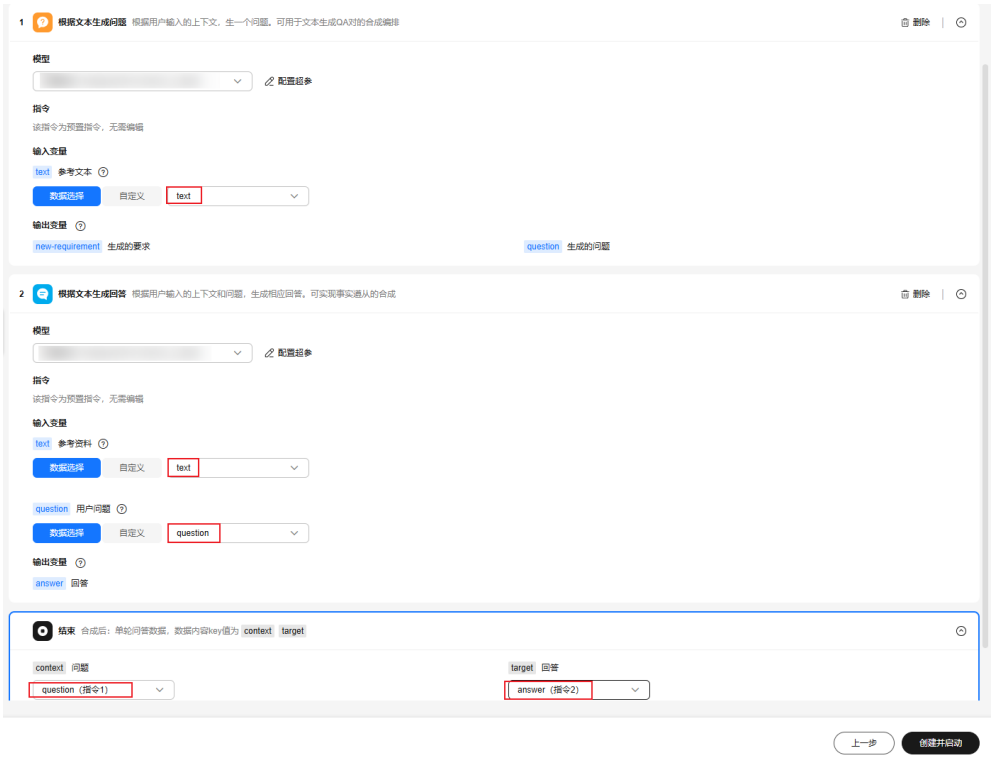
创建文本类数据集合成任务

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

创建文本类数据集合成任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 合成任务”，单击界面右上角“创建合成任务”。
3. 在“创建合成任务”页面，选择需要合成的数据集，选择合成内容与合成轮数。
4. 如果合成前的数据集与合成后的数据集结构相同，可选择开启“将源数据集整合至合成后数据”，在所有合成轮数运行完成后，将生成的数据与原始数据集合并，单击“下一步”。
5. 进入“指令编排”页面，在左侧“添加指令”页面可选择预置指令或自定义指令。
   - 预置指令。平台为用户提供了多种预置指令，便于用户执行合成任务，请详见**预置数据指令介绍**。
   - 自定义指令。平台支持编排用户自定义指令。自定义指令的创建详见**创建自定义数据合成指令**。
6. 指令选择完成后，单击“确定”，并配置指令参数。
   如**图3-5**，展示了预训练文本类数据集的合成指令参数配置示例，该合成任务实现利用预训练文本生成问答对。

图 3-5 预训练文本类数据集合成指令参数配置示例



7. 指令编排完成后，单击右上角“启用调测”，可以对当前编排的指令效果进行预览。
8. 指令调测完成后，单击“创建并启动”，平台将启动合成任务。
9. 当数据合成任务运行成功后，状态将从“运行中”变为“运行成功”，表示数据已经完成合成操作。

说明

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

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

预置数据指令介绍

ModelArts Studio平台的数据合成功能为用户提供了预置指令，用户可以在“数据工程 > 数据管理 > 数据指令 > 系统预置”查看指令详情，如图3-6，单击“调测”可查看调测指南，如图3-7，帮助用户更好地使用该指令。

预置的数据指令清单详见表3-19。

图 3-6 指令详情

图 3-7 调测指南

表 3-19 预置数据指令清单

指令分类	指令名称	指令描述
生成问题	问题改写为更低难度	该指令可以通过用户输入的问题，使大模型按要求生成一个难度更低、更为简单的问题。
	问题改写为更高难度	该指令通过用户输入的问题，使大模型按要求生成一个难度更高、更为复杂的问题。
	基于提问生成作答要求	该指令根据输入的问题，使大模型泛化一个相应问题的作答要求，该要求与原问题内容不直接相关。该指令可与根据作答要求回答问题的指令进行编排，实现风格多样回答的合成。

指令分类	指令名称	指令描述
	根据样例生成相似问题_few-shot	该指令通过用户输入的多个问题样例，生成一个或多个与样例风格相匹配的新问题。
	根据文本生成问题	根据用户输入的上下文，生一个问题。可用于文本生成QA对的合成编排
	问题改写	改写问题，生成更复杂的问题，可用于指令泛化
生成回答	回答改写	根据用户指定人设，改写回答的风格，不改变回答内容。可与人设泛化指令编排，实现问答对泛化
	根据文本生成回答_遵循要求	根据用户指定的指令要求和问题，根据输入的上下文生成相应回答。可与指令泛化进行编排，实现事实遵循类问答对泛化
	问题生成回答	根据提问，生成回答
	根据文本生成回答_扮演指定人设	根据用户指定的人设和问题，根据输入的上下文生成相应回答。可与人设泛化指令编排，实现事实遵循类问答对泛化。
	问题生成回答_扮演指定人设	根据用户指定的人设和问题，生成相应回答。可与人设泛化指令编排，实现问答对泛化。
	根据文本生成回答	根据用户输入的上下文和问题，生成相应回答。可实现事实遵从的合成
生成问答对	文本生成问答对_判断题	该指令能够从用户提供的参考文本中构建出一个判断题，同时给出其正确回答。
	文本生成问答对_填空题	该指令能够从用户提供的参考文本中构建出一个填空题，同时给出其正确回答。
	文本生成问答对_单选题	该指令能够从用户提供的参考文本中构建出一个包含四个选项的单选题，同时给出其正确回答
	文本生成问答对_多选题	该指令能够从用户提供的参考文本中构建出一个包含四个选项的多选题，同时给出其正确回答。
	文本生成问答对_问答题	该指令能够从用户提供的参考文本中构建出一个问题，同时给出其相应回答。
	根据文本抽取问答对_金融场景	根据用户输入的金融类文档进行问答对的抽取。
生成人设	根据问题生成人设	根据用户输入的问题生成一个人物设定。
其他	BadCase问题泛化	该指令通过用户提供的badcase问题和回答，利用大模型生成在类似情景下可能犯错的攻击性问题。用户可指定生成的攻击性问题个数，个数不超过10。

指令分类	指令名称	指令描述
	根据答案推导解题思路	指令通过用户输入的问题和回答，利用大模型生成包含相应解题思路的回答。
	指令泛化	根据用户指定风格，进行指令泛化。可与指定要求类的问答对生成相关指令编排，实现问答对泛化

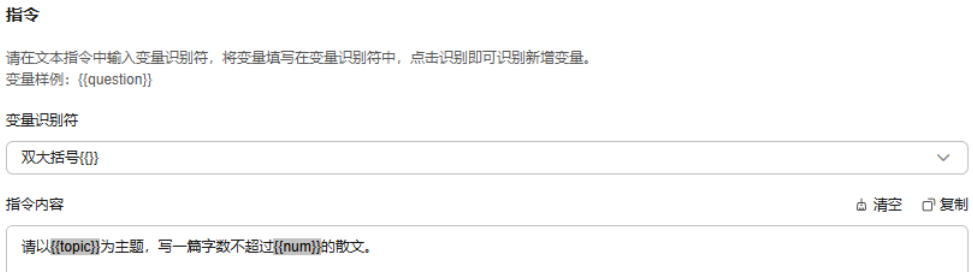
创建自定义数据合成指令

平台支持用户创建自定义数据合成指令。

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

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据指令”，在“自定义”页签，单击“创建指令”。
3. 在“创建指令”弹窗中，输入名称与描述，单击“确定”，进入配置合成指令页面。
4. 选择变量标识符为“双大括号{{}}”，输入指令为“请以{{topic}}为主题，写一篇字数不超过{{num}}的散文。”
   单击“识别”，再单击“确定”。

   图 3-8 配置指令

5. 按照表3-20进行变量配置。

   表 3-20 数据指令变量配置

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

其中，输出变量的“变量描述”字段为大模型理解的内容，需仔细填写。

图 3-9 配置变量

6. 调测数据指令。
   - 在“调试 > 模型”中，选择指令所需的模型，单击“配置超参”可自定义设定超参数值。
   - 在“调试 > 输入”中，可通过给变量赋值来查看效果。

   图 3-10 指令调测

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

3.5.3.3 标注文本类数据集

创建文本类数据集标注任务

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

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表3-21**。

表 3-21 不同角色支持的数据标注任务权限清单

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

创建文本类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。
3. 在“创建标注任务”页面选择需要标注的文本类数据集，并选择标注项。选择标注项时，不同类型的数据文件对应的标注项有所差异，可基于页面提示进行选择。

   其中，“单轮问答”标注项支持“AI辅助标注”功能，若开启该功能，需要选择已部署的NLP服务作为AI辅助标注模型。
4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表3-22**配置标注分配与审核。

   表 3-22 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
标注审核	是否审核	• 否，标注后不进行审核操作。 • 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。

参数类型	参数名称	参数说明
	审核要求	• 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 • 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

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

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

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

   如图3-11，以标注单轮问答数据为例，需要逐一确认问题（Q）及答案（A）是否正确，如果问题或答案不正确，可以对其进行二次编辑。

   图 3-11 文本类数据集标注示例

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

说明

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

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

审核标注后的文本类数据集

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

该操作需要具备**标注审核员**角色权限。

审核文本类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

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

说明

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

管理标注后的文本类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

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

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

3.5.3.4 配比文本类数据集

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

如果单个数据集已满足您的需求，**可跳过此章节至发布文本类数据集**。

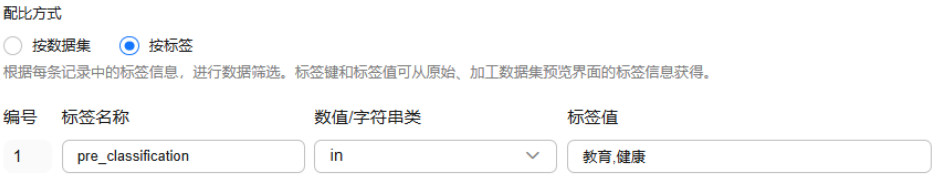
创建文本类数据集配比任务

创建文本类数据集配比任务步骤如下：

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

     填写示例如**图3-12**所示。

     图 3-12 “按标签”配比方式填写示例

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

3.5.4 加工图片类数据集

3.5.4.1 加工图片类数据集

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

创建图片类数据集加工任务

创建图片类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的图片类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于图片类数据集，可选择的加工算子请参见**表3-17**。
   a. 在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

      图 3-13 算子编排

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

图 3-14 选择加工模板

5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。

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

说明

在完成数据加工后，如果无需使用数据标注、数据合成功能，可直接在“加工任务”页面单击操作列“生成”，生成加工数据集。

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

3.5.4.2 标注图片类数据集

创建图片类数据集标注任务

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

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表3-23**。

表 3-23 不同角色支持的数据标注任务权限清单

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

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

创建图片类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。
3. 在“创建标注任务”页面选择需要标注的图片类数据集与标注项。

   如果选择“图片Caption”或“物体检测”标注项，则可开启“AI预标注”功能。AI预标注将自动生成标注内容，不会覆盖原始数据集，供标注人员参考，以提高标注效率。
4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表3-24**配置标注分配与审核。

表 3-24 标注分配与审核配置

参数类型	参数名称	参数说明
标注分配	标注员	添加标注人员与数量。
	标注要求	选择标注项为“图片Caption”且开启AI预标注功能时，可设置以下两种方式的“标注要求”： ● 选择“全部标注”：要求标注人员需要对全部的数据进行人工标注后才可提交标注结果。 ● 选择“可部分标注”：允许标注人员在确认AI预标注满足要求后，直接使用AI预标注功能完成数据集的标注并提交标注结果。
标注审核	是否审核	● 否，标注后不进行审核操作。 ● 是，审核员会检查标注员的标注内容，若发现问题，审核员可注明原因并驳回标注数据，标注员需重新标注。
	审核员	添加审核人员与数量。

参数类型	参数名称	参数说明
	审核要求	• 全部审核：要求审核员对全部数据，逐条进行人工审核，才能完成审核任务。 • 可部分审核：审核员在审核一部分数据后，发现标注质量均很高，则可以一键提交剩余待审核数据，默认审核通过，即可完成审核任务。

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

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

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

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

说明

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

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

审核标注后的图片类数据集

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

该操作需要具备**标注审核员**角色权限。

审核视频类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

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

说明

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

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

管理标注后的图片类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

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

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

3.5.4.3 配比图片类数据集

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

如果单个数据集已满足您的需求，**可跳过此章节至发布图片类数据集**。

创建图片类数据集配比任务

创建图片类数据集配比任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 配比任务”，单击界面右上角“创建配比任务”。
3. 在“数据集选择”页签选择需要配比的文本类数据集（至少选择两个），单击“下一步”。
4. 在“数据配比”页面，可以设置不同数据集的配比数量，单击“确定”。
5. 页面将返回至“数据配比”页面，配比任务运行成功后，状态将显示为“运行成功”。
6. 单击操作列“生成”，将生成“发布数据集”。

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

3.5.5 加工视频类数据集

3.5.5.1 加工视频类数据集

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

创建视频类数据集加工任务

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。

3. 在“创建加工任务”页面，选择需要加工的视频类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于视频类数据集，可选择的加工算子请参见**表3-16**。
   a. 在左侧“添加算子”分页勾选所需算子。
   b. 在右侧“加工步骤编排”页面配置各算子参数，可拖动右侧“ ⠿ ”以调整算子执行顺序。

      图 3-15 算子编排

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

      若选择使用加工模板，将删除当前已编排的加工步骤。

      图 3-16 选择加工模板

5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。

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

   说明

   在完成数据加工后，如果无需使用数据标注功能，可直接在“加工任务”页面单击操作列“生成”，生成加工数据集。

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

3.5.5.2 标注视频类数据集

创建视频类数据集标注任务

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

数据标注功能支持创建标注任务、标注数据集（标注作业）、审核标注后的数据集（审核作业）与管理标注任务（任务管理）。其中，不同角色权限支持的功能及展示的前端界面略有差异，详见**表3-25**。

表 3-25 不同角色支持的数据标注任务权限清单

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

创建视频类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”，单击页面右上角“创建标注任务”。
3. 在“创建标注任务”页面选择需要标注的视频类数据集与标注项，单击“下一步”。

   如果选择“视频Caption”标注项，则可开启“AI预标注”功能。AI预标注将自动生成标注内容，不会覆盖原始数据集，供标注人员参考，以提高标注效率。
4. 可选择开启“多人作业”功能，开启后，可选择多人协同完成作业，并增加审核功能可供选择。参考**表3-26**配置标注分配与审核。

表 3-26 标注分配与审核配置

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

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

   其中，对于“标注作业员”角色，可单击“标注”执行标注作业任务。

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

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

说明

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

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

审核标注后的视频类数据集

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

该操作需要具备**标注审核员**角色权限。

审核视频类数据集标注结果的步骤如下：

1. 登录ModelArts Studio平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 标注任务”。
3. 单击“审核”可进入审核页面审核数据。

   如果需要将该审核任务移交给其他人员，可以单击“移交”，并设置移交人员以及移交数量，单击“确定”。

4. 进入审核页面后，可通过单击“通过”或“不通过”逐一对数据进行审核，直至所有数据审核完成。

说明

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

管理标注后的视频类数据集

平台支持超级管理员、管理员、标注管理员对标注的数据集进行如下操作：

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

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

3.5.6 加工气象类数据集

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

创建气象类数据集加工任务

创建气象类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据加工 > 加工任务”，单击界面右上角“创建加工任务”。
3. 在“创建加工任务”页面，选择需要加工的气象类数据集，单击“下一步”。
4. 进入“加工步骤编排”页面。对于气象类数据集，可选择的加工算子请参见**表3-18**。

   a. 在左侧“添加算子”分页勾选所需算子。

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

图 3-17 算子编排

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

若选择使用加工模板，将删除当前已编排的加工步骤。

图 3-18 选择加工模板

5. 加工步骤编排完成后，单击“启动加工”，将启动加工任务。

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

6. 在完成数据加工后，在“加工任务”页面单击操作列“生成”，生成加工数据集。

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

3.5.7 管理加工后的数据集

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

平台支持对加工数据集查看基本信息、数据血缘等管理操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据集 > 加工数据集”。
3. 单击数据集名称查看加工数据集的基本信息、数据预览、数据血缘以及操作记录。
   - 在“基本信息”页签可查看数据集的详细信息。
   - 在“数据预览”页签可查看加工后的数据内容。
   - 在“数据血缘”页签查看该数据集所经历的操作，如导入、合成等操作。
   - 在“操作记录”页签可以查看数据集所经历的操作及状态等信息。
4. 单击操作列的“删除”，可删除不需要的数据集。
   - 如果需要恢复删除的数据集，可单击右上角“显示已删除数据”，被删除的数据集将在列表显示，可将数据集恢复。
   - 如果需要彻底删除数据集，可单击数据集名称进入详情页，确认数据集内容后彻底删除该数据集。

说明

删除“加工数据集”属于高危操作，删除前，请确保该数据集不再使用。

3.6 发布数据集

3.6.1 数据集发布场景介绍

数据发布介绍

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

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

- **数据评估**

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

- **数据发布**

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

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

数据发布意义

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

- **多格式支持**

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

- **提高训练效率**

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

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

支持数据发布的数据集类型

支持数据发布的数据集类型见**表3-27**。

表 3-27 支持数据发布的数据集类型

数据类型	数据评估	数据发布
文本类	√	√
图片类	√	√
视频类	√	√
气象类	-	√
预测类	-	√
其他类	-	√

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

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

除文本类、图片类数据集外，其余类型的数据集当前仅支持发布为标准格式。

3.6.2 发布文本类数据集

3.6.2.1 评估文本类数据集

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

如果无需使用数据评估操作，**可跳过此章节至发布文本类数据集。**

创建文本类数据集评估标准

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

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

创建文本类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，在“人工评估标准”页签，平台预置了文本类数据集评估标准“NLP数据质量标准 V1.0”，单击评估标准名称，可以查看具体的评估项。

   图 3-19 预置文本类数据集评估标准

3. 在“人工评估标准”页面，单击“创建标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 单击“下一步”，编辑评估项。

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

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

创建文本类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

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

创建文本类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，单击界面右上角“创建评估任务”。

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

   如图3-20，对于文本类数据集而言，可选中问题内容后，右键标记数据问题。

   图 3-20 标记数据集问题

8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。

   单击操作列“报告”，可查看数据集质量评估报告。

3.6.2.2 发布文本类数据集

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

文本类数据集支持发布的格式为：

- **标准格式：**数据工程功能支持的原始格式。

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

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

- **盘古格式：**训练盘古大模型时，需要将数据集格式发布为“盘古格式”。

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

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

创建文本类数据集发布任务

创建文本类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“文本 > 预训练文本”类型的数据集。

图 3-21 选择数据集模态

4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。

   由于数据工程需要支持对接盘古大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

   当前支持标准格式、盘古格式：

   - **标准格式**：数据工程功能支持的原始格式。该格式的数据集可发布到资产中，但下游模型开发不可见。
   - **盘古格式**：使用盘古大模型训练时所需要使用的数据格式，该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。

   说明

   如果使用该数据集训练盘古大模型，请将选择格式配置为**盘古格式**。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.3 发布图片类数据集

3.6.3.1 评估图片类数据集

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

如果无需使用数据评估操作，**可跳过此章节至发布图片类数据集。**

创建图片类数据集评估标准

ModelArts Studio大模型开发平台针对图片类数据集预设了一套基础评估标准，涵盖了图像清晰度、分辨率、标签准确性、图像一致性等多个质量维度，用户可以直接使用该标准或在该标准的基础上创建评估标准。

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

创建图片类数据集评估标准步骤如下：

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

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

图 3-22 预置图片类数据集评估标准

3. 在“人工评估标准”页面，单击“创建标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 单击“下一步”，编辑评估项。

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

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

创建图片类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

创建图片类数据集评估任务前，请参考**加工图片类数据集**，生成一个“加工数据集”。

创建图片类数据集评估任务步骤如下：

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

8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。
   单击操作列“报告”，可查看数据集质量评估报告。

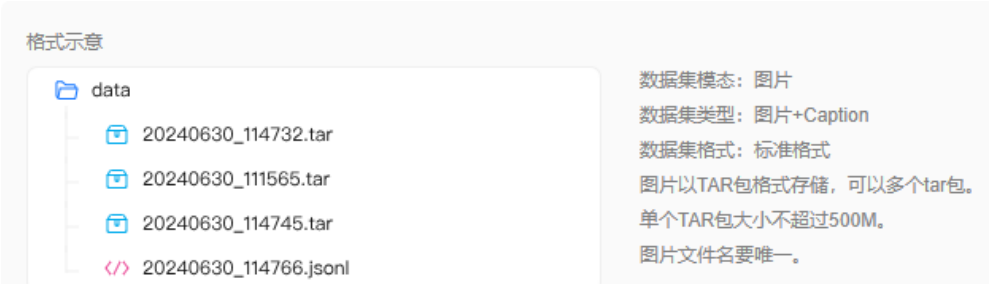
3.6.3.2 发布图片类数据集

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

图片类数据集支持发布的格式为：

- **标准格式：**如图3-23，平台默认的格式。该格式的数据集可发布到资产中，但下游模型开发不可见。

  图 3-23 图片类数据集标准格式示例

- **盘古格式：**如图3-24，训练盘古大模型时，需要将数据集格式发布为“盘古格式”，该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。

  图 3-24 图片类数据集盘古格式示例

创建图片类数据集发布任务

创建图片类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“图片 > 图片+Caption”类型的数据集。

图 3-25 选择数据集模态

4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。

   由于数据工程需要支持对接盘古大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

   当前支持标准格式、盘古格式：

   - **标准格式**：数据工程功能支持的原始格式。该格式的数据集可发布到资产中，但下游模型开发不可见。
   - **盘古格式**：使用盘古大模型训练时所需要使用的数据格式，该数据集将被用于ModelArts Studio大模型开发平台的模型开发中使用。

   说明

   如果使用该数据集训练盘古大模型，请将发布格式配置为**盘古格式**。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.4 发布视频类数据集

3.6.4.1 评估视频类数据集

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

如果无需使用数据评估操作，**可跳过此章节至发布视频类数据集。**

创建视频类数据集评估标准

ModelArts Studio大模型开发平台针对视频类数据集预设了一套基础评估标准，涵盖了视频的清晰度、帧率、完整性、标签准确性等多个质量维度，用户可以直接使用该标准或在该标准的基础上创建评估标准。

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

创建视频类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，在“人工评估标准”页签，平台预置了视频类数据集评估标准“视频数据质量标准 V1.0”，单击评估标准名称，可以查看具体的评估项。

图 3-26 预置视频类数据集评估标准

3. 在“人工评估标准”页面，单击“创建标准”，选择预置标准作为参考项，并填写“评估标准名称”和“描述”。
4. 单击“下一步”，编辑评估项。

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

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

创建视频类数据集评估任务

平台仅支持对“加工数据集”执行评估操作。

创建视频类数据集评估任务前，请参考**加工视频类数据集**，生成一个“加工数据集”。

创建视频类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据评估”，单击界面右上角“创建评估任务”。
3. 选择需要评估的加工数据集，并设置抽样样本的数量。
4. 单击“下一步”，选择评估标准。单击“下一步”设置评估人员，单击“下一步”填写任务名称。
5. 单击“完成创建”，将返回至“数据评估”页面，评估任务创建成功后状态将显示为“已创建”。
6. 单击操作列的“评估”，进入评估页面。
7. 在评估页面，可参考评估项对当前数据的问题进行标注，且满足则单击“通过”，不满足则单击“不通过”。
8. 全部数据评估完成后，在“人工评估”页面可查看评估进展为“100%”。

   单击操作列“报告”，可查看数据集质量评估报告。

3.6.4.2 发布视频类数据集

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

视频类数据集当前仅支持发布为“标准格式”。

创建视频类数据集发布任务

创建视频类数据集发布任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“视频 > 仅视频”类型的数据集。

图 3-27 选择数据集模态

4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前视频类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

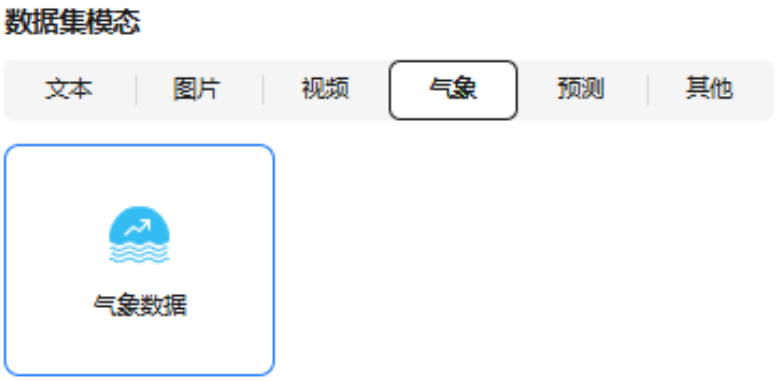
   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.5 发布气象类数据集

气象类数据集当前仅支持发布为“标准格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“气象 > 气象数据”类型的数据集。

图 3-28 选择数据集模态



4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前气象类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.6 发布预测类数据集

预测类数据集当前仅支持发布为“标准格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“预测 > 时序”类型的数据集。

图 3-29 选择数据集模态

4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前预测类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

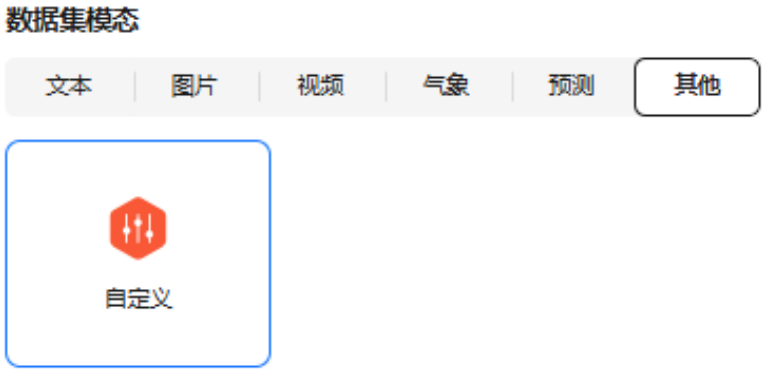
   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.7 发布其他类数据集

其他类数据集当前仅支持发布为“标准格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据发布 > 发布任务”，单击界面右上角“创建发布任务”。
3. 在“创建发布任务”页面，选择数据集模态，如“其他 > 自定义”类型的数据集。

   图 3-30 选择数据集模态

4. 选择数据集，单击“下一步”。
5. 在“基本配置”中选择数据用途、数据集可见性、适用场景。当前其他类数据集仅支持发布标准格式。
6. 填写数据集名称、描述，设置扩展信息后，单击“确定”执行数据集发布操作。

   当任务状态显示为“运行成功”时，说明数据发布任务执行成功，生成的“发布数据集”可在“数据工程 > 数据管理 > 数据集 > 发布数据集”中查看。

3.6.8 管理发布后的数据集

完成数据配比、或数据流通任务的数据集，在对应任务列表执行“生成”操作，将生成“发布数据集”被平台统一管理，并用于后续的发布任务。

平台支持对发布数据集查看基本信息、数据血缘等管理操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据管理 > 数据集 > 发布数据集”。
3. 单击数据集名称查看发布数据集的基本信息、数据预览、数据血缘以及操作记录。
   - 在“基本信息”页签可查看数据集的详细信息。
   - 在“数据预览”页签可查看发布后的数据内容。
   - 在“数据血缘”页签查看该数据集所经历的操作，如导入、合成、训练等操作。
   - 在“操作记录”页签可以查看数据集所经历的操作及状态等信息。
4. 单击操作列的“删除”，可删除不需要的数据集。

- 如果需要恢复删除的数据集，可单击右上角“显示已删除数据”，被删除的数据集将在列表显示，可将数据集恢复。
- 如果需要彻底删除数据集，可单击数据集名称进入详情页，确认数据集内容后彻底删除该数据集。

说明

删除“发布数据集”属于高危操作，删除前，请确保该数据集不再使用。

3.7 数据工程常见报错与解决方案

数据工程常见报错及解决方案请详见**表3-28**。

表 3-28 数据工程常见报错与解决方案

功能模块	常见报错	解决方案
数据获取	File format mismatch, require [{0}].	请检查创建数据集时使用的数据，与平台要求的文件内容格式是否一致。
	Verification failed. Please check the content format is consistent with the template requirements.	请检查创建数据集时使用的数据，与平台要求的文件内容格式是否一致。
	content type [%s] not support, only [%s] support.	数据集中的内容不支持，请保证上传的数据格式与平台要求的一致。
	get obs bucket folders error.	请检查OBS服务是否正常，是否可以访问OBS桶数据。
数据加工	dataset is not online.	数据加工使用的数据集未上线，请先执行上线操作。
	invalid obs path.	请检查数据集对应的OBS路径是否有效，是否可正常访问。
数据标注	annotate data not exist.	请检查标注数据集是否存在，是否被删除。
	obs url invalid.	请检查数据集对应的OBS路径是否有效，是否可正常访问。
	data management query dataset data invalid.	请检查标注数据集是否存在，是否被删除。
	dataset obs file empty.	检查数据集文件是否还存在于原先的OBS桶中。
	download obs file failed.	请检查网络是否正常，是否可以访问OBS桶中的数据。

功能模块	常见报错	解决方案
数据评估	annotate type is invalid.	请检查上传的数据中，使用的数据标注类型、数据标注要求与平台要求的是否一致。
	annotate data not exist.	待评测数据不存在，请检查数据是否导入成功，OBS桶是否为空。
	obs url invalid.	请检查数据集对应的OBS路径是否有效，是否可正常访问。
	standard item not exist.	请检查评估标准是否存在，是否被删除。
	the corresponding data has been marked as deleted.	请检查评估数据是否被删除。
	the current data not exist.	请检查评估数据是否存在，是否被删除。
	dataset file type does not match standard file type.	请检查上传的数据集文件类型与平台要求的标准文件类型是否一致。
	data management query dataset data invalid.	请检查数据集中是否有异常格式的数据。
	dataset obs file empty.	检查数据集文件是否还存在于原先的OBS桶中。
数据流通	Dataset [%s] is not found.	请检查数据集是否存在。
	Dataset [%s] status is invalid.	请检查数据集状态是否正常。

4 开发盘古 NLP 大模型

4.1 使用数据工程构建 NLP 大模型数据集

NLP 大模型支持接入的数据集类型

盘古NLP大模型仅支持接入文本类数据集，数据集文件内容包括：预训练文本、单轮问答、多轮问答、带人设单轮问答、带人设多轮问答等，不同训练方式所需要使用的数据见表4-1，该数据集格式要求请参见文本类数据集格式要求。

表 4-1 训练 NLP 大模型数据集类型要求

基模型	训练场景	数据集类型	数据集内容	文件格式
NLP	预训练	文本	预训练文本	jsonl
	微调	文本	单轮问答	jsonl、csv
		文本	多轮问答	jsonl
		文本	单轮问答（人设）	jsonl、csv
		文本	多轮问答（人设）	jsonl

训练 NLP 大模型所需数据量

使用数据工程构建盘古NLP大模型数据集进行模型训练时，所需数据量见表4-2。

表 4-2 构建 NLP 大模型所需数据量

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N1	微调	-	1000条/每场景	≥ 1万条/每场景	32K

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N2	微调	-	1000条/每场景	≥ 1万条/每场景	32K
N4	微调	-	1000条/每场景	≥ 1万条/每场景	4K版本：4096 32K版本：32768

评测 NLP 大模型所需数据量

要求所有文本大小最大不超过100MB，目录下文件数量最多不超过100个。数据条数范围为：3-1000条。

构建 NLP 大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程构建盘古NLP大模型数据集流程见**表4-3**。

表 4-3 盘古 NLP 大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	**导入数据至盘古平台**
加工文本类数据集	加工文本类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	**加工文本类数据集**
	合成文本类数据集	利用预置或自定义的数据指令对原始数据进行处理，并根据设定的轮数生成新数据。该过程能够在一定程度上扩展数据集，增强训练模型的多样性和泛化能力。	**合成文本类数据集**
	标注文本类数据集	为无标签数据集添加准确的标签，确保模型训练所需的高质量数据。平台支持人工标注和AI预标注两种方式，用户可根据需求选择合适的标注方式。数据标注的质量直接影响模型的训练效果和精度。	**标注文本类数据集**

流程	子流程	说明	操作指导
	配比文本类数据集	数据配比是将多个数据集按特定比例组合的过程。通过合理的配比，确保数据集的多样性、平衡性和代表性，避免因数据分布不均而引发的问题。	配比文本类数据集
发布文本类数据集	评估文本类数据集	平台预置了多种数据类型的基础评估标准，包括NLP、视频和图片数据，用户可根据需求选择预置标准或自定义评估标准，从而精确优化数据质量，确保数据满足高标准，提升模型性能。	评估文本类数据集
	发布文本类数据集	发布流通是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。 平台支持发布的数据集格式为**默认格式、盘古格式**。 • 默认格式：平台默认的格式。 • 盘古格式：训练盘古大模型时，需要发布为该格式。当前仅文本类、图片类数据集支持发布为盘古格式。	发布文本类数据集

4.2 训练 NLP 大模型

4.2.1 NLP 大模型训练流程与选择建议

NLP 大模型训练流程介绍

NLP大模型的训练分为两个关键阶段：预训练和微调。

- **预训练阶段：**在这一阶段，模型通过学习大规模通用数据集来掌握语言的基本模式和语义。这一过程为模型提供了处理各种语言任务的基础，如阅读理解、文本生成和情感分析，但它还未能针对特定任务进行优化。

  针对**预训练阶段**，还可以继续进行训练，这一过程称为**增量预训练**。增量预训练是在已经完成的预训练的基础上继续训练模型。增量预训练旨在使模型能够适应新的领域或数据需求，保持其长期的有效性和准确性。
- **微调阶段：**基于预训练的成果，微调阶段通过在特定领域的数据集上进一步训练，使模型能够更有效地应对具体的任务需求。这一阶段使模型能够精确执行如文案生成、代码生成和专业问答等特定场景中的任务。在微调过程中，通过设定训练指标来监控模型的表现，确保其达到预期的效果。完成微调后，将对用户模型进行评估并进行最终优化，以确保满足业务需求，然后将其部署和调用，用于实际应用。

NLP 大模型选择建议

选择合适的NLP大模型类型有助于提升训练任务的准确程度。您可以根据模型**可处理最大Token长度**，选择合适的模型，从而提高模型的整体效果，详见表4-4。

此外，不同类型的NLP大模型在训练过程中，读取中文、英文内容时，字符长度转换为Token长度的转换比有所不同，详见表4-5。

表 4-4 不同系列 NLP 大模型对处理文本的长度差异

模型名称	可处理最大上下文长度	可处理最大输出长度	说明
Pangu-NLP-N1-Chat-32K-20241130	32K	4K	2024年11月发布的版本，支持8K序列长度训练，4K/32K序列长度推理。全量微调、LoRA微调8个训练单元起训，1个推理单元即可部署。
Pangu-NLP-N1-Chat-128K-20241130	128K	4K	2024年11月发布的版本，仅支持128K序列长度推理。
Pangu-NLP-N1-32K-3.1.34	32K	4K	2024年11月发布的版本，支持8K序列长度训练，4K/32K序列长度推理。全量微调、LoRA微调8个训练单元起训，1个推理单元即可部署，4K支持256并发，32K支持256并发。
Pangu-NLP-N1-32K-3.2.36	**32K**	**4K**	**2025年1月发布的版本，支持32K序列长度训练，4K/32K序列长度推理。全量微调、LoRA微调8个训练单元起训，1个推理单元即可部署，4K支持256并发，32K支持256并发。**
Pangu-NLP-N1-128K-3.1.34	128K	4K	2024年11月发布的版本，仅支持128K序列长度推理，4卡2并发。
Pangu-NLP-N1-128K-3.2.36	**128K**	**4K**	**2025年1月发布的版本，仅支持128K序列长度推理，4个推理单元8并发。**
Pangu-NLP-N2-Base-20241030	-	4K	2024年11月发布的版本，仅支持模型增量预训练。32个训练单元起训，预训练后的模型版本需要通过微调之后，才可支持推理部署。
Pangu-NLP-N2-Chat-32K-20241030	32K	4K	2024年10月发布版本，支持8K序列长度训练，4K/32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署。此模型版本差异化支持预训练特性、INT8量化特性。

模型名称	可处理最大上下文长度	可处理最大输出长度	说明
Pangu-NLP-N2-4K-3.2.35	**4K**	**4K**	**2025年1月发布的版本，支持4K序列长度训练，4K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持192并发。此模型版本差异化支持预训练特性、INT8量化特性。**
Pangu-NLP-N2-32K-3.1.35	32K	4K	2024年12月发布版本，支持8K序列长度训练，4K/32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，4K支持64并发，32K支持64并发。此模型版本差异化支持预训练特性、INT8量化特性。
Pangu-NLP-N2-32K-3.1.35	**32K**	**4K**	**2025年1月发布的版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。此模型版本差异化支持预训练特性、INT8量化特性。**
Pangu-NLP-N2-128K-3.1.35	128K	4K	2024年12月发布的版本，仅支持128K序列长度推理部署，8个推理单元64并发。
Pangu-NLP-N2-256K-3.1.35	256K	4K	2024年12月发布的版本，仅支持256K序列长度推理部署，8个推理单元64并发。
Pangu-NLP-N4-Chat-4K-20241130	32K	4K	2024年11月发布的版本，支持4K序列长度训练，4K序列长度推理。全量微调64个训练单元起训，LoRA微调32个训练单元起训，8个训练单元即可部署。此模型版本差异化支持预训练、INT8/INT4量化特性。
Pangu-NLP-N4-Chat-32K-20241130	32K	4K	2024年11月发布的版本，仅支持32K序列长度推理部署。
Pangu-NLP-N4-4K-2.5.35	**4K**	**4K**	**2025年1月发布的版本，支持4K序列长度训练，4K序列长度推理。全量微调64个训练单元起训，LoRA微调32个训练单元起训，8个推理单元即可部署，支持128并发。此模型版本差异化支持预训练、INT8/INT4量化特性。**

模型名称	可处理最大上下文长度	可处理最大输出长度	说明
Pangu-NLP-N4-4K-2.5.32	4K	4K	2024年11月发布的版本，支持4K序列长度训练，4K序列长度推理。全量微调64个训练单元起训，LoRA微调32个训练单元起训，8个推理单元即可部署，支持64并发。此模型版本差异化支持预训练、INT8/INT4量化特性。
Pangu-NLP-N4-32K-2.5.32	32K	4K	2024年11月发布的版本，仅支持32K序列长度推理部署，8个推理单元64并发。
Pangu-NLP-N4-32K-2.5.35	**32K**	**4K**	**2025年1月发布的版本，仅支持32K序列长度推理部署，8个推理单元128并发。**

表 4-5 Token 转换比

模型规格	Token比（Token/英文单词）	Token比（Token/汉字）
N1	0.75	1.5
N2	0.88	1.24
N4	0.75	1.5

说明

针对Token转换比，平台提供了**Token计算器**功能，可以根据您输入的文本计算Token数量，您可以通过以下方式使用该功能：

- 在左侧导航栏选择“能力调测”，单击右下角“Token计算器”使用该功能。
- 使用API调用Token计算器，详见《API参考》“API > Token计算器”。

NLP 大模型训练类型选择建议

平台针对NLP大模型提供了两种训练类型，包括预训练、微调，二者区别详见表4-6。

表 4-6 预训练、微调训练类型区别

训练方式	训练目的	训练数据	模型效果	应用场景举例
预训练	**关注通用性**：预训练旨在让模型学习广泛的通用知识，建立词汇、句法和语义的基础理解。通过大规模的通用数据训练，模型可以掌握丰富的语言模式，如语言结构、词义关系和常见的句型。	**使用大规模通用数据**：通常使用海量的无监督数据（如文本语料库、百科文章），这些数据覆盖广泛的领域和语言表达方式，帮助模型掌握广泛的知识。	**适合广泛应用**：经过预训练后，模型可以理解自然语言并具备通用任务的基础能力，但还没有针对特定的业务场景进行优化。预训练后的模型主要用于多个任务的底层支持。	通过使用海量的互联网文本语料对模型进行预训练，使模型理解人类语言的基本结构。
微调	**关注专业性**：微调是对预训练模型的参数进行调整，使其在特定任务中达到更高的精度和效果。微调的核心在于利用少量的特定任务数据，使模型的表现从通用性向具体任务需求过渡。	**使用小规模的特定任务数据**：微调通常需要小规模但高质量的标注数据，直接与目标任务相关。通过这些数据，模型可以学习到任务特定的特征和模式。	**在特定任务上具有更高的准确性**：微调后的模型在具体任务中表现更优。相较于预训练阶段的通用能力，微调能使模型更好地解决细分任务的需求。	在一个客户服务问答系统中，可以用特定领域（如电商、保险）的对话数据对预训练模型进行微调，使其更好地理解和回答与该领域相关的问题。

此外，针对微调训练任务，平台提供了两种微调方式：

- **全量微调**：适合有充足数据并关注特定任务性能的场景。在全量微调中，模型的所有参数都会调整，以适应特定任务的需求。这种方式适合样本量较大、对推理效果要求较高的任务。例如，在特定领域（如金融、医疗）中，若拥有大量标注数据，且需要更高的特定任务推理精度，则全量微调是优先选择。
- **LoRA微调**：适用于数据量较小、侧重通用任务的情境。LoRA（Low-Rank Adaptation）微调方法通过调整模型的少量参数，以低资源实现较优结果，适合聚焦于领域通用任务或小样本数据情境。例如，在针对通用客服问答的场景中，样本量少且任务场景广泛，选择LoRA微调既能节省资源，又能获得较好的效果。

微调方式选择建议：

- 若项目中数据量有限或任务场景较为广泛，可以选择**LoRA微调**以快速部署并保持较高适用性。
- 若拥有充足数据且关注特定任务效果，选择**全量微调**有助于大幅提升在特定任务上的模型精度。

4.2.2 创建 NLP 大模型训练任务

创建 NLP 大模型预训练任务

创建NLP大模型预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“创建训练任务”页面，参考**表4-7**完成训练参数设置。

表 4-7 NLP 大模型预训练参数说明

参数分类	训练参数	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“NLP大模型”。
	训练类型	选择“预训练”。
	基础模型	选择预训练所需的基础模型，可从“已发布模型”或“未发布模型”中进行选择。
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	数据批量大小	数据集进行分批读取训练，设定每个批次数据的大小。 通常情况下，较大的数据批量可以使梯度更加稳定，从而有利于模型的收敛。然而，较大的数据批量也会占用更多的显存资源，这可能导致显存不足，并且会延长每次训练的时长。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。

参数分类	训练参数	参数说明
	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
	模型保存步数	每训练一定数量的步骤（或批次），模型的状态将会被保存。可以通过以下公式预估已训练的数据量： token_num = step * batch_size * sequence ● token_num：已训练的数据量（以Token为单位）。 ● step：已完成的训练步数。 ● batch_size：每个训练步骤中使用的样本数量。 ● sequence：每个数据样本中的Token数量。
	数据预处理并发个数	定义了在预处理数据时，能够同时处理文件的并行进程数量。设定这个参数的主要目的是通过并发处理来加速数据预处理，从而提升训练效率。
数据配置	训练数据	选择训练模型所需的数据集。
资源配置	训练单元	创建当前训练任务所需的训练单元数量。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。

5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建 NLP 大模型增量预训练任务

在模型完成**创建NLP大模型预训练任务**预训练后，可以对训练后的模型继续训练，该过程称为“增量预训练”。

创建NLP大模型增量预训练任务前，请确保有已完成预训练的NLP大模型。

创建NLP大模型增量预训练任务的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“创建训练任务”页面，选择“盘古大模型 > NLP大模型 > 预训练”。
4. 选择基础模型，可选“从资产选模型”、“从任务选模型”，在弹窗中支持从“本空间”或“其他空间”选择预训练好的NLP大模型，单击“确定”。
5. 其余参数配置等步骤同**创建NLP大模型预训练任务**。

创建 NLP 大模型全量微调任务

创建NLP大模型全量微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“创建训练任务”页面，参考**表4-8**完成训练参数设置。

表 4-8 NLP 大模型全量微调参数说明

参数分类	训练参数	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“NLP大模型”。
	训练类型	选择“微调”。
	训练目标	选择“全量微调”。 • 全量微调：在模型进行有监督微调时，对大模型的所有参数进行更新。这种方法通常能够实现最佳的模型性能，但需要消耗大量计算资源和时间，计算开销较大。
	基础模型	选择全量微调所用的基础模型， 可从“已发布模型”或“未发布模型”中进行选择。

参数分类	训练参数	参数说明
	高级设置	checkpoints：在模型训练过程中，用于保存模型权重和状态的机制。 ● 关闭：关闭后不保存checkpoints，无法基于checkpoints执行续训操作。 ● 自动：自动保存训练过程中的所有checkpoints。 ● 自定义：根据设置保存指定数量的checkpoints。
训练参数	热身比例	热身比例是指在模型训练初期逐渐增加学习率的过程。 由于训练初期模型的权重通常是随机初始化的，预测能力较弱，若直接使用较大的学习率，可能导致更新过快，进而影响收敛。为解决这一问题，通常在训练初期使用较小的学习率，并逐步增加，直到达到预设的最大学习率。通过这种方式，热身比例能够避免初期更新过快，从而帮助模型更好地收敛。
	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	单步迭代时处理的数据批量大小	指定每次迭代时处理的数据批量大小。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。

参数分类	训练参数	参数说明
	模型保存步数	每训练一定数量的步骤（或批次），模型的状态将会被保存。可以通过以下公式预估已训练的数据量： token_num = step * batch_size * sequence ● token_num：已训练的数据量（以Token为单位）。 ● step：已完成的训练步数。 ● batch_size：每个训练步骤中使用的样本数量。 ● sequence：每个数据样本中的Token数量。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 ● adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
数据配置	训练数据	选择训练模型所需的数据集。
	验证数据	● 若选择“从训练数据拆分”，则需进一步配置数据拆分比例。 ● 若选择“从已有数据导入”，则需选择导入的数据集。
资源配置	训练单元	创建当前训练任务所需的训练单元数量。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同模型训练参数默认值存在一定差异，请以前端页面展示的默认值为准。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建 NLP 大模型 LoRA 微调任务

创建NLP大模型LoRA微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。

3. 在“创建训练任务”页面，参考表4-9完成训练参数设置。

表 4-9 NLP 大模型 LoRA 微调参数说明

参数分类	训练参数	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“NLP大模型”。
	训练类型	选择“微调”。
	训练目标	选择“LoRA微调”。 ● LoRA微调：在模型微调过程中，只对特定的层或模块的参数进行更新，而其余参数保持冻结状态。这种方法可以显著减少计算资源和时间消耗，同时在很多情况下，依然能够保持较好的模型性能。
	基础模型	选择全量微调训练所用的基础模型， 可从“已发布模型”或“未发布模型”中进行选择。
训练参数	数据批量大小	数据批量是指训练过程中将数据集分成小批次进行读取，并设定每个批次的数据大小。 通常，较大的批量能够使梯度更加稳定，有助于模型的收敛。然而，较大的批量也会占用更多显存，可能导致显存不足，并延长每次训练时间。
	学习率衰减比率	用于控制训练过程中学习率下降的幅度。 计算公式为：最低学习率 = 初始学习率 × 学习率衰减比率。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。
	Lora矩阵的轶	较高的取值意味着更多的参数被更新，模型具有更大的灵活性，但也需要更多的计算资源和内存。较低的取值则意味着更少的参数更新，资源消耗更少，但模型的表达能力可能受到限制。
	Agent微调	在训练Agent所需的NLP大模型时，可以开启此参数。通过调整训练数据中的Prompt，引导模型在特定领域或任务上生成更符合预期的回答。 在使用此参数前，请先联系盘古客服，调整Prompt和训练数据。

参数分类	训练参数	参数说明
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂。
	优化器	优化器参数用于更新模型的权重，常见包括adamw。 • adamw是一种改进的Adam优化器，增加了权重衰减机制，有效防止过拟合。
数据配置	训练数据	选择训练模型所需的数据集。
	验证数据	• 若选择“从训练数据拆分”，则需进一步配置数据拆分比例。 • 若选择“从已有数据导入”，则需选择导入的数据集。
资源配置	训练单元	创建当前训练任务所需的训练单元数量。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

说明

不同模型训练参数默认值存在一定差异，请以前端页面展示的默认值为准。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

4.2.3 查看 NLP 大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表4-10。

表 4-10 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。

训练状态	训练状态含义
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于训练状态为“已完成”的任务，单击任务名称，可在“训练结果”页面查看训练指标，模型的训练指标介绍请参见**表4-11**。

图 4-1 查看训练指标

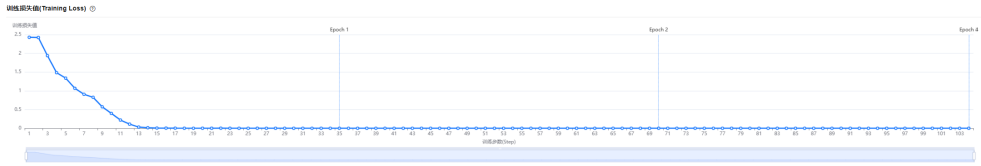


表 4-11 训练指标说明

模型	训练指标	指标说明
NLP大模型	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	验证损失值	模型在验证集上的损失值。值越小，意味着模型对验证集数据的泛化能力越好。

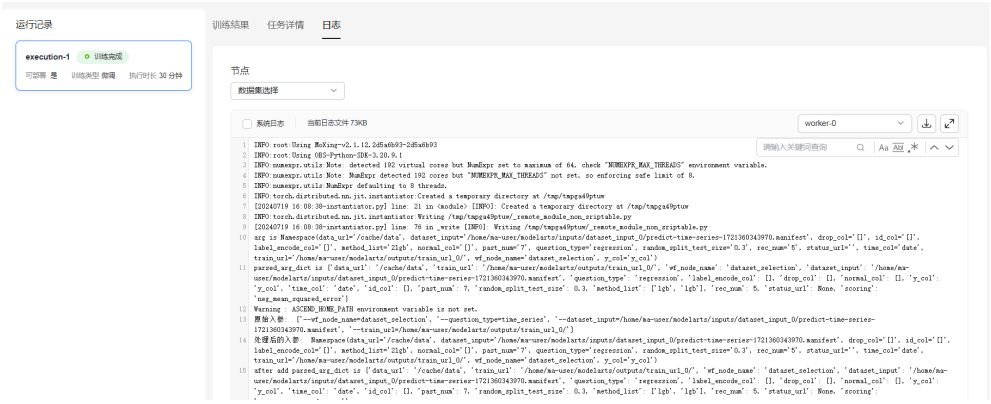
获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。

对于训练异常或失败的任务可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**NLP大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 4-2 获取训练日志



4.2.4 发布训练后的 NLP 大模型

NLP大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 4-3 训练结果页面

4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

4.2.5 管理 NLP 大模型训练任务

在训练任务列表中，任务创建者可以对任务进行编辑、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，可进行如下操作：
   - 编辑。单击操作列的“编辑”，可以修改模型的checkpoints、训练参数、训练数据以及基本信息等。
   - 克隆。单击操作列的“更多 > 克隆”，参照**创建NLP大模型训练任务**填写参数，可以复制当前训练任务。

- 停止。单击操作列的“更多 > 停止”，可以停止处于“排队中”或“运行中”状态的任务。
- 重试。单击操作列的“更多 > 重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
- 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的训练任务。

说明

删除属于高危操作，删除前请确保当前任务不再需要。

4.2.6 NLP 大模型训练常见报错与解决方案

NLP大模型训练常见报错及解决方案请详见表4-12。

表 4-12 NLP 大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空。	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。
训练日志提示“root: XXX valid number is 0”	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。
训练日志提示“ValueError: label_map not match”	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

4.3 压缩 NLP 大模型

模型在部署前，通过模型压缩可以降低推理显存占用，节省推理资源提高推理性能。

平台当前仅可对NLP大模型进行压缩，**支持压缩的模型清单请详见**《产品介绍》> “模型能力与规格 > 盘古NLP大模型能力与规格”。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型压缩”，单击界面右上角“创建压缩任务”。
3. 在“创建压缩任务”页面，选择需要压缩的基础模型，支持选择已发布模型或未发布模型。
4. 选择压缩策略。除INT8压缩策略外，部分模型支持INT4压缩策略，可在选择模型后，根据页面展示的策略进行选择。
   - INT8：该压缩策略将模型参数压缩至8位字节，可以有效降低推理显存占用。
   - INT4：该压缩策略与INT8相比，可以进一步减少模型的存储空间和计算复杂度。
5. 配置资源。选择计费模式并设置训练单元。
6. 可选择开启订阅提醒。开启后，系统将在本次压缩任务状态变更时，向用户发送短信/邮件提醒。
7. 填写基本信息，包括任务名称、压缩后模型名称与描述，单击“立即创建”。

   当压缩任务状态为“已完成”时，表示模型已完成压缩操作。

4.4 部署 NLP 大模型

4.4.1 创建 NLP 大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，参考**表4-13**完成部署参数设置。

表 4-13 NlP 大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	模型来源	选择“盘古大模型”。
	模型类型	选择“NLP大模型”。
	部署模型	选择需要进行部署的模型。
	最大TOKEN长度	模型可最大请求的上下文TOKEN数。

参数分类	部署参数	参数说明
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。 – 负载均衡：创建负载均衡步骤请详见**步骤5：创建负载均衡**。 – 实例数：设置部署模型时所需的实例数。
安全护栏	选择模式	安全护栏保障模型调用安全。
	计费模式	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置（选择云上部署时）	计费模式	限时免费。
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

4. 参数填写完成后，单击“立即部署”。

4.4.2 查看 NLP 大模型部署任务详情

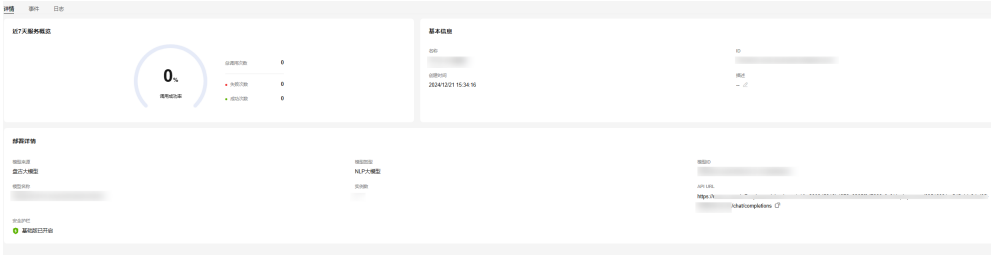
部署任务创建成功后，可以查看大模型部署任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建NLP大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。

3. 可单击模型名称可进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

图 4-4 部署详情



4.4.3 管理 NLP 大模型部署任务

模型更新

完成**创建NLP大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 部署模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 4-5 升级模式

说明

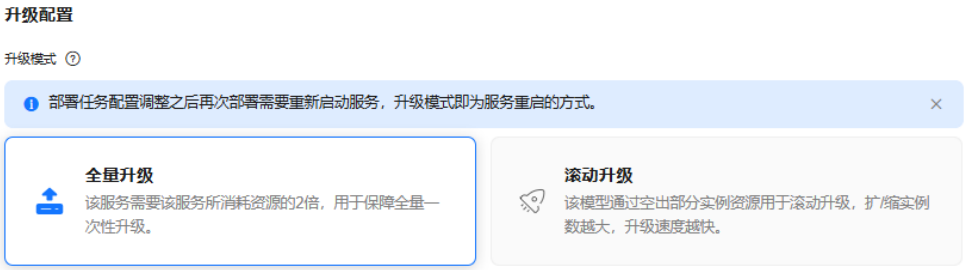
升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建NLP大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 4-6 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成**创建NLP大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确定”。

4.5 评测 NLP 大模型

4.5.1 创建 NLP 大模型评测数据集

NLP大模型支持人工评测与自动评测，在执行模型评测任务前，需创建评测数据集。

评测数据集的创建步骤与训练数据集一致，本章节仅做简单介绍，详细步骤请参见**使用数据工程构建NLP大模型数据集**。

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“数据工程 > 数据获取”，单击界面右上角“创建导入任务”。
3. 在“创建导入任务”页面选择所需要的“文件内容”、“文件格式”、“导入来源”，并单击“选择路径”上传数据文件。

   NLP大模型评测数据集支持的格式见**表4-14**。

   表 4-14 评测数据集格式

模型类型	评测数据集格式
NLP大模型	文本-单轮问答-jsonl格式

4. 上传数据文件后，填写“数据集名称”与“描述”，单击“立即创建”。
5. 在左侧导航栏中选择“数据工程 > 数据发布 > 数据流通”，单击界面右上角“创建流通任务”。
6. 在“创建流通任务”页面选择数据集模态并选择数据集文件。
7. 单击“下一步”，选择发布格式，填写名称，选择数据集可见性，单击“下一步”。

   说明

   如果评测盘古大模型， 需要在流通数据集时，将数据集格式发布为“盘古格式”。
8. 选择“资源配置”，并单击“确定”。待任务状态为“运行成功”后，单击“启动”，生成“发布数据集”。

4.5.2 创建 NLP 大模型评测任务

创建NLP大模型评测任务前，请确保已完成**创建NLP大模型评测数据集**操作。

预训练的NLP大模型不支持评测。

创建 NLP 大模型自动评测任务

创建NLP大模型自动评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型评测 > 任务管理”，单击界面右上角“创建评测任务”。
3. 在“创建评测任务”页面，参考**表4-15**完成部署参数设置。

表 4-15 NLP 大模型自动评测任务参数说明

参数分类	参数名称	参数说明
选择服务	模型来源	选择“NLP大模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openai、tgi、自定义三种格式。openai格式即是由OpenAI公司开发并标准化的一种大模型请求格式；tgi格式即是Hugging Face团队推出的一种大模型请求格式。 – 接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。
评测配置	评测类型	选择“自动评测”。
	评测规则	选择“基于规则”。
	评测数据集	● 评测模板：使用预置的专业数据集进行评测。 ● 单个评测集：由用户指定评测指标（F1分数、准去率、BLEU、Rouge）并上传评测数据集进行评测。 选择“单个评测集”时需要上传待评测数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“模型评测 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”查看模型评测结果，包括模型详细的得分以及评测明细。

创建 NLP 大模型人工评测任务

创建NLP大模型人工评测任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型评测 > 任务管理”，单击界面右上角“创建评测任务”。
3. 在“创建评测任务”页面，参考**表4-16**完成部署参数设置。

表 4-16 NLP 大模型人工评测任务参数说明

参数分类	参数名称	参数说明
选择服务	模型来源	选择“NLP大模型”。
	服务来源	支持已部署服务、外部服务两种选项。单次最多可评测10个模型。 ● 已部署服务：选择部署至ModelArts Studio平台的模型进行评测。 ● 外部服务：通过API的方式接入外部模型进行评测。选择外部服务时，需要填写外部模型的接口名称、接口地址、请求体、响应体等信息。 – 请求体支持openai、tgi、自定义三种格式。openai格式即是由OpenAI公司开发并标准化的一种大模型请求格式；tgi格式即是Hugging Face团队推出的一种大模型请求格式。 – 接口的响应体需要按照jsonpath语法要求进行填写，jsonpath语法的作用是从响应体的json字段中提取出所需的数据。
评测配置	评测类型	选择“人工评测”。
	评测指标	由用户自定义评测指标并填写评测标准。
	评测数据集	待评测的数据集。
	评测结果存储位置	模型评测结果的存储位置。
基本信息	评测任务名称	填写评测任务名称。
	描述	填写评测任务描述。

4. 参数填写完成后，单击“立即创建”，回退至“模型评测 > 人工评测”页面。
5. 当状态为“待评测”时，可以单击操作列“在线评测”进入评测页面。

6. 依据页面提示对评估效果区域进行评测打分，全部数据评测完成后单击“提交”。

图 4-7 人工评测示例

7. 在“人工测评”页面，评测任务的状态将显示为“已完成”，单击操作列“评测报告”查看模型评测结果。

4.5.3 查看 NLP 大模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型评测 > 任务管理”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**NLP大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 评测明细”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“下载记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

NLP 大模型评测指标说明

NLP大模型支持自动评测与人工评测，各指标说明如**表4-17**、**表4-18**、**表4-19**。

表 4-17 NLP 大模型自动评测指标说明-不使用评测模板

评测指标（自动评测-不使用评测模板）	指标说明
F1_SCORE	精准率和召回率的调和平均数，数值越高，表明模型性能越好。
BLEU-1	模型生成句子与实际句子在单字层面的匹配度，数值越高，表明模型性能越好。
BLEU-2	模型生成句子与实际句子在词组层面的匹配度，数值越高，表明模型性能越好。
BLEU-4	模型生成结果和实际句子的加权平均精确率，数值越高，表明模型性能越好。

评测指标（自动评测-不使用评测模板）	指标说明
ROUGE-1	模型生成句子与实际句子在单个词的相似度，数值越高，表明模型性能越好。
ROUGE-2	模型生成句子与实际句子在两个词的相似度，数值越高，表明模型性能越好。
ROUGE-L	模型生成句子与实际句子在最长公共子序列的相似度，数值越高，表明模型性能越好。
PRECISION	问答匹配的精确度，模型生成句子与实际句子相比的精确程度，数值越高，表明模型性能越好。

表 4-18 NLP 大模型自动评测指标说明-使用评测模板

评测指标（自动评测-使用评测模板）	指标说明
评测得分	每个数据集上的得分为模型在当前数据集上的通过率；评测能力项中若有多个数据集则按照数据量的大小计算通过率的加权平均数。
综合能力	综合能力是计算所有数据集通过率的加权平均数。

表 4-19 NLP 大模型人工评测指标说明

评测指标（人工评测）	指标说明
准确性	模型生成答案正确且无事实性错误。
average	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
goodcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
badcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分以下的占比。
用户自定义的指标	由用户定义的指标，如有用性、逻辑性、安全性等。

4.5.4 管理 NLP 大模型评测任务

管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型评测 > 任务管理”，可进行如下操作：
   - 克隆。单击操作列的“ 克隆”，可以复制当前评测任务。
   - 启动。单击操作列的“启动”，可以重启运行失败的评测任务。
   - 删除。单击操作列的“删除”，可以删除当前不需要的评测任务。

   说明

   删除属于高危操作，删除前请确保当前任务不再需要。

4.6 调用 NLP 大模型

4.6.1 使用“能力调测”调用 NLP 大模型

能力调测功能支持用户调用预置或训练后的NLP大模型。使用该功能前，请完成模型的部署操作，步骤详见**创建NLP大模型部署任务**。

使用“能力调测”调用NLP大模型可实现文本对话能力，即在输入框中输入问题，模型将基于问题输出相应的回答，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“能力调测”，单击“文本对话”页签。
3. 选择需要调用的服务。可从“预置服务”或“我的服务”中选择。
4. 填写系统人设。如“你是一个AI助手”，若不填写，将使用系统默认人设。
5. 在页面右侧配置参数，具体参数说明见**表4-20**。

表 4-20 NLP 大模型能力调测参数说明

参数	说明
搜索增强	搜索增强通过结合大语言模型与传统搜索引擎技术，提升了搜索结果的相关性、准确性和智能化。 例如，当用户提出复杂查询时，传统搜索引擎可能仅返回一系列相关链接，而大模型则能够理解问题的上下文，结合多个搜索结果生成简洁的答案，或提供更详细的解释，从而进一步改善用户的搜索体验。
温度	用于控制生成文本的多样性和创造力。调高温度会使得模型的输出更多样性和创新性。 默认值：0
核采样	控制生成文本多样性和质量。调高核采样可以使输出结果更加多样化。 默认值：1.0
最大口令限制	用于控制聊天回复的长度和质量。 默认值：2048

参数	说明
话题重复度控制	用于控制生成文本中的重复程度。调高参数模型会更频繁地切换话题，从而避免生成重复内容。 默认值：0
词汇重复度控制	用于调整模型对频繁出现的词汇的处理方式。调高参数会使模型减少相同词汇的重复使用，促使模型使用更多样化的词汇进行表达。 默认值：0
历史对话保留轮数	选择“文本对话”功能时具备此参数。表示系统能够记忆的历史对话数。 默认值：10

6. 如图4-8，输入对话，单击“生成”，模型将输出相应的回答。

图 4-8 调测 NLP 大模型

4.6.2 使用 API 调用 NLP 大模型

预置模型或训练后的模型部署成功后，可以使用“文本对话”API实现模型调用。

表 4-21 NLP 大模型 API 清单

API分类	API访问路径（URI）
文本对话	/v1/{project_id}/deployments/{deployment_id}/chat/completions

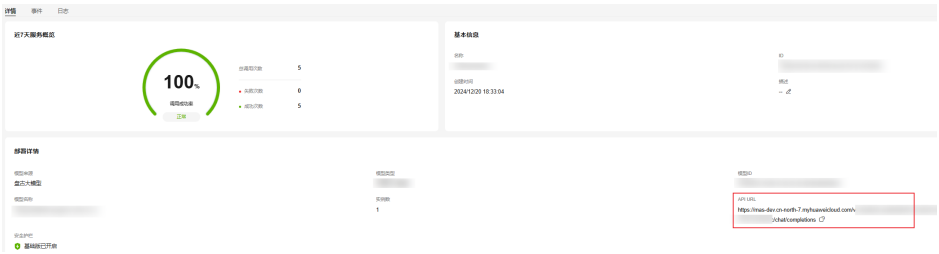
获取调用路径

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

   在左侧导航栏中选择“模型开发 > 模型部署”。

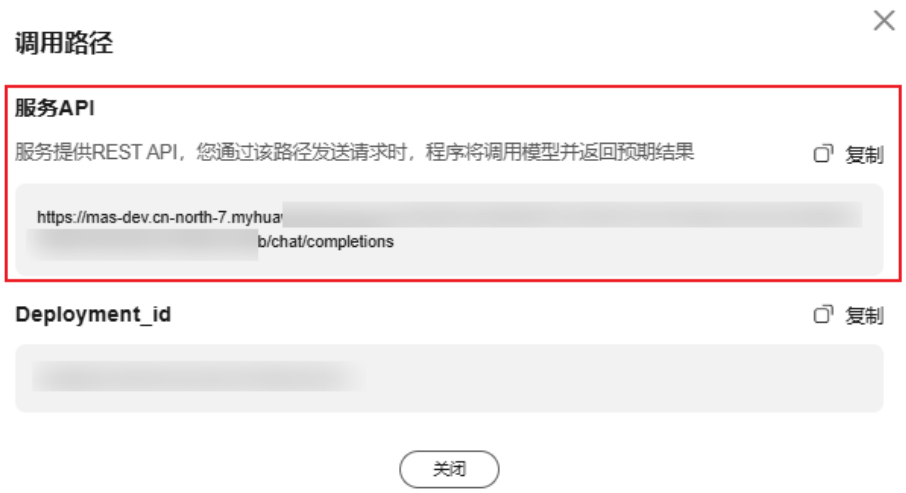
   - 获取已部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“详情”页签，可获取模型调用路径，如图4-9。

     图 4-9 获取已部署模型的调用路径

   - 获取预置服务的调用路径。在“预置服务”页签中，选择所需调用的NLP大模型，单击“调用路径”，在“调用路径”弹窗可获取模型调用路径，如图4-10。

     图 4-10 获取预置服务的调用路径

使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见获取调用路径。
2. 调用API有两种认证方式，包括Token认证和AppCode认证。其中，AppCode认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过AppCode认证调用请求。

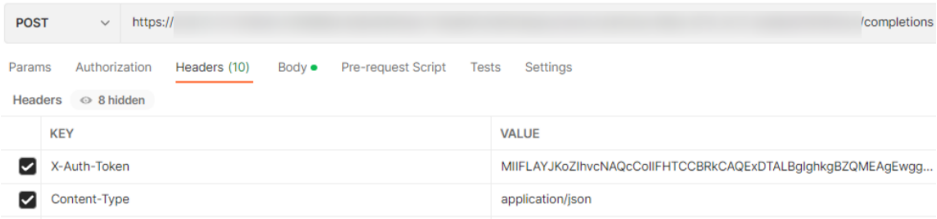
   参考表4-22填写请求Header参数。

表 4-22 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
AppCode认证	Content-Type	application/json
	X-Apig-AppCode	AppCode值，获取AppCode步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“模型开发 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“APP Code”操作列中可获取APPCode值。

如图4-11，为Token认证方式的请求Header参数填写示例。

图 4-11 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。

```
{
   "messages": [
     {
        "content": "介绍下长江，以及长江中典型的鱼类"
     }
    ],
   "temperature": 0.9,
   "max_tokens": 600
}
```

4. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示NLP大模型API调用成功。

4.6.3 统计 NLP 大模型调用信息

针对调用的大模型，平台提供了统一的管理功能。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 单击左侧导航栏“调用统计”，选择“NLP”页签。
3. 选择当前调用的NLP大模型，可以按照不同时间跨度查看当前模型的调用总数、调用失败的次数、调用的总Tokens数、以及输入输出的Tokens数等基本信息。

   此外，该功能还提供了可视化界面，可额外查看响应时长以及安全护栏拦截次数。

5 开发盘古科学计算大模型

5.1 使用数据工程构建科学计算大模型数据集

科学计算大模型支持接入的数据集类型

盘古科学计算大模型仅支持接入气象类数据集，该数据集格式要求请参见**气象类数据集格式要求**。

训练科学计算大模型训练数据要求所需数据量

构建科学计算大模型进行训练的数据要求见**表5-1**。

表 5-1 科学计算大模型训练数据要求

模型类别	特征要求	水平分辨率要求	区域范围要求	时间要求	数据获取方式
气象/降水模型	需包含4个表面层特征（10m u风、10m v风、2米温度、海平面气压），13高空层次（1000、925、850、700、600、500、400、300、250、200、150、100、50hPa）的5个高空层特征（重力位势、u风、v风、比湿、温度）。	25km*25km。	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	训练数据一般可通过公开数据集获取，例如ERA5。ERA5是由欧洲中期天气预报中心（ECMWF）提供的全球气候的第五代大气再分析数据集，它覆盖从1940年1月至今的时间段，提供每小时的大气、陆地和海洋气候变量的估计值。 • ERA5数据下载官方指导：**https://confluence.ecmwf.int/display/CKB/How+to+download+ERA5** • 高空变量数据下载链接：**https://cds.climate.copernicus.eu/datasets**，查找名称中包含ERA5和pressure levels的数据集。 • 表面变量数据下载链接：**https://cds.climate.copernicus.eu/datasets**，查找名称中包含ERA5和single levels的数据集。

模型类别	特征要求	水平分辨率要求	区域范围要求	时间要求	数据获取方式
海洋模型	需包含5个表面层特征（10m u风、10m v风、2米温度、海平面气压、海表面气压），15个深海层次（"0m", "6m", "10m", "20m", "30m", "50m", "70m", "100m", "125m", "150m", "200m", "250m", "300m", "400m", "500m"）的4个深海层特征（海盐、海洋流速u、海洋流速v、温度）。	-	全球范围，纬度90N~-90S，经度0W~360E。	训练集和验证集均推荐使用>1个月的历史数据。	海洋模型数据获取方式：**https://data.hycom.org/datasets/GLBv0.08/expt_53.X/data/**

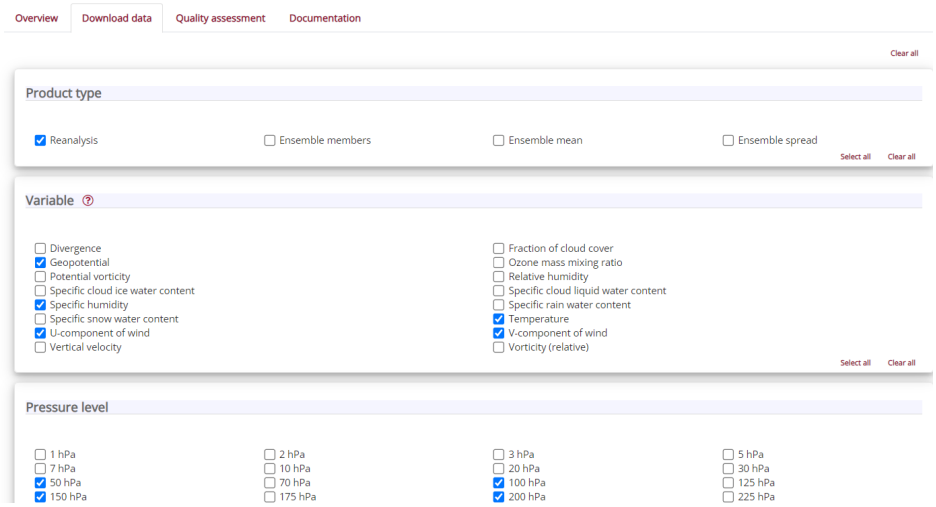
气象/降水模型获取方式示例：

1. 示例一：以下载2021年7月16日高空变量数据为例，下载内容为高空变量（重力位势、u风、v风、比湿、温度，1000、925、850、700、600、500、400、300、250、200、150、100、50hPa高空层次）0点、6点、12点、18点时刻的数据文件，下载步骤示例如下：
   a. 注册并登录数据下载平台，在高空变量数据下载链接中：
      - Product type选择Reanalysis。
      - Variable新选择Geopotential、Specific humidity、Temperature、U-component of wind、V-component of wind。
      - Pressure level选择1000hPa、925hPa、850hPa、700hPa、600hPa、500hPa、400hPa、300hPa、250hPa、200hPa、150hPa、100hPa、50hPa。
      - Year选择2021，Month选择July，Day选择16。

- Time选择00:00、06:00、12:00、18:00。
- Geographical area选择Whole available region。
- Format选择NetCDF(experimental)。

b. 数据准备好后，单击“Submit Form”，基于页面提示单击“Download”下载数据。

图 5-1 下载高空变量数据



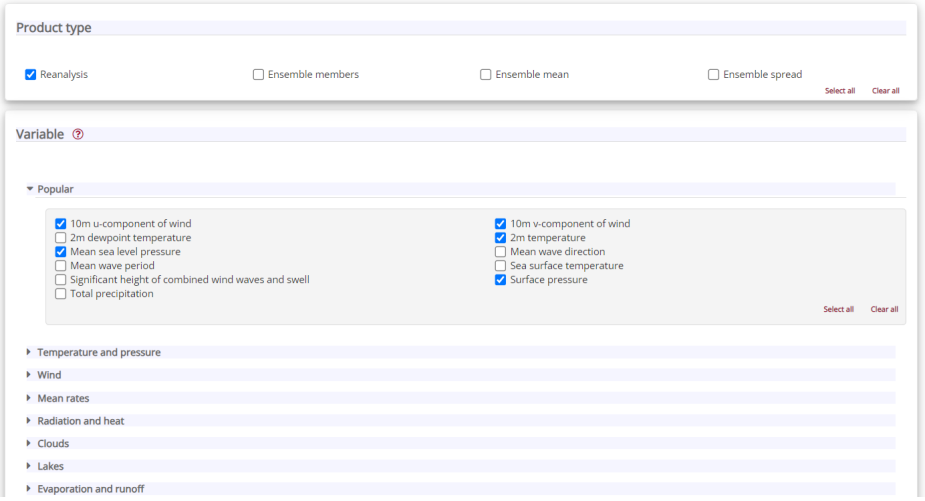
2. 示例二：以下载2021年7月16日表面变量数据为例，下载内容为表面变量（10m u风、10m v风、2米温度、海平面气压）0点、6点、12点、18点时刻的数据文件，下载步骤示例如下：

a. 注册并登录数据下载平台，在表面变量数据下载链接中：

- Product type选择Reanalysis。
- Popular选择10m u-component of wind、10m v-component of wind、2m temperature、Mean sea level pressure，Surface pressure。
- Year选择2021，Month选择July，Day选择16。
- Time选择00:00、06:00、12:00、18:00。
- Geographical area选择Whole available region。
- Format选择NetCDF(experimental)。

b. 数据准备好后，单击“Submit Form”，基于页面提示单击“Download”下载数据。

图 5-2 下载表面变量数据



构建科学计算大模型数据集流程

在ModelArts Studio大模型开发平台中，使用数据工程创建盘古科学计算大模型数据集流程见表5-2。

表 5-2 盘古科学计算大模型数据集构建流程

流程	子流程	说明	操作指导
导入数据至盘古平台	创建导入任务	将存储在OBS服务中的数据导入至平台统一管理，用于后续加工或发布操作。	导入数据至盘古平台
加工气象类数据集	加工气象类数据集	通过专用的加工算子对数据进行预处理，确保数据符合模型训练的标准和业务需求。不同类型的数据集使用专门设计的算子，例如去除噪声、冗余信息等，提升数据质量。	加工气象类数据集
发布气象类数据集	发布气象类数据集	数据发布是将单个数据集发布为特定格式的“发布数据集”，用于后续模型训练等操作。	发布气象类数据集

5.2 训练科学计算大模型

5.2.1 科学计算大模型训练流程与选择建议

科学计算大模型训练流程介绍

科学计算大模型的训练主要分为两个阶段：预训练与微调。

- **预训练阶段**：预训练是模型学习基础知识的过程，基于大规模通用数据集进行。例如，在区域海洋要素预测中，可以重新定义深海变量、海表变量，调整深度

层、时间分辨率、水平分辨率以及区域范围，以适配自定义区域的模型场景。此阶段需预先准备区域的高精度数据。

- **微调阶段**：在预训练模型的基础上，微调利用特定领域的数据进一步优化模型，使其更好地满足实际任务需求。例如，区域海洋要素预测的微调是在已有模型上添加最新数据，不改变模型结构参数或引入新要素，以适应数据更新需求。

在实际流程中，通过设定训练指标对模型进行监控，以确保效果符合预期。在微调后，评估用户模型，并进行最终优化，确认其满足业务需求后，进行部署和调用，以便实际应用。

科学计算大模型选择建议

科学计算大模型支持训练的模型类型有：全球中期天气要素模型、降水模型、区域中期海洋智能预测模型。

- **全球中期天气要素预测模型、降水模型选择建议**：

  科学计算大模型的全球中期天气要素预测模型、降水模型，可以对未来一段时间的天气和降水进行预测，具备以下优势：

  - 高时间精度：全球中期天气要素预测模型可以预测未来1、3、6、24小时的天气情况，降水模型可预测未来6小时的降水情况。高时间精度对于农业、交通、能源等领域的决策和规划非常重要。
  - 全球覆盖：全球中期天气要素预测模型和降水模型能够在全球范围内进行预测，不仅仅局限于某个地区。它的分辨率相当于赤道附近每个点约25公里*25公里的空间。
  - 数据驱动：全球中期天气要素预测模型和降水模型使用历史天气数据来训练模型，从而提高预测的准确性。这意味着它可以直接利用过去的观测数据，而不仅仅依赖于数值模型。

  全球中期天气要素预测模型、降水模型信息见**表5-3**。

表 5-3 全球中期天气要素预测模型、降水模型信息表

模型	预报层次	预报高空变量	预报表面变量	降水	时间分辨率	水平分辨率	区域范围
全球中期天气要素预测模型	13层（1000hpa, 925hpa, 850hpa, 700hpa, 600hpa, 500hpa, 400hpa, 300hpa, 250hpa, 200hpa, 150hpa, 100hpa, 50hpa）	T：温度 Q：比湿 Z：重力位势 U：U风 V：V风	MLSP：海平面气压 U10：10米U风，经度方向 V10：10米V风，纬度方向 T2M：2米温度	-	1、3、6、24小时	0.25°*0.25°	全球

模型	预报层次	预报高空变量	预报表面变量	降水	时间分辨率	水平分辨率	区域范围
降水基模型	13层（1000hpa,925hpa,850hpa,700hpa,600hpa,500hpa,400hpa,300hpa,250hpa,200hpa,150hpa,100hpa,50hpa）	T：温度 Q：比湿 Z：重力位势 U：U风 V：V风	MLSP：海平面气压 U10：10米U风，经度方向 V10：10米V风，纬度方向 T2M：2米温度	PRECIP6：过去6h累计降水 PRECIP24：过去24h累计降水	1、3、6、24小时	0.25°*0.25°	全球

支持训练的模型清单见表5-4，您可根据具体使用场景选择合适的模型。例如天气基础要素预测，需要时间分辨率为1小时的场景下，您可以选择**Pangu-AI4S-Weather_1h-3.0.0**模型。

表 5-4 中期天气要素预测模型、降水模型的类型

模型名称	说明
Pangu-AI4S-Weather_Precip-20241030	2024年10月发布的版本，用于降水预测，支持1个实例部署推理。
Pangu-AI4S-Weather-Precip_6h-3.0.0	2024年12月发布的版本，相较于10月发布的版本模型运行速度有提升，用于降水预测，支持1个实例部署推理。
Pangu-AI4S-Weather-Precip_6h-3.1.0	2025年1月发布的版本，用于降水预测，支持1个实例部署推理。

模型名称	说明
Pangu-AI4S-Weather_1h-20241030	2024年10月发布的版本，用于天气基础要素预测，时间分辨率为1小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_1h-3.0.0	2024年12月发布的版本，相较于10月发布的版本模型运行速度有提升，用于天气基础要素预测，时间分辨率为1小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_1h-3.1.0	2025年1月发布的版本，用于天气基础要素预测，时间分辨率为1小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_3h-20241030	2024年10月发布的版本，用于天气基础要素预测，时间分辨率为3小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_3h-3.0.0	2024年12月发布的版本，相较于10月发布的版本模型运行速度有提升，用于天气基础要素预测，时间分辨率为3小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_3h-3.1.0	2025年1月发布的版本，用于天气基础要素预测，时间分辨率为3小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_6h-20241030	2024年10月发布的版本，用于天气基础要素预测，时间分辨率为6小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_6h-3.0.0	2024年12月发布的版本，用于天气基础要素预测，时间分辨率为6小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_6h-3.1.0	2025年1月发布的版本，用于天气基础要素预测，时间分辨率为6小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_6h-3.1.1	2025年1月发布的版本，用于天气基础要素预测，时间分辨率为6小时，相较于3.1.0版本预报准确度更高，1个实例部署。
Pangu-AI4S-Weather_24h-20241030	2024年10月发布的版本，用于天气基础要素预测，时间分辨率为24小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_24h-3.0.0	2024年12月发布的版本，相较于10月发布的版本运行速度有提升，用于天气基础要素预测，时间分辨率为24小时，1个训练单元起训及1个实例部署。
Pangu-AI4S-Weather_24h-3.1.0	2025年1月发布的版本，用于天气基础要素预测，时间分辨率为24小时，1个训练单元起训及1个实例部署。

- **中期海洋智能预测模型选择建议**：

  科学计算大模型的中期海洋智能预测模型，可以对未来一段时间海洋要素进行预测。可为海上防灾减灾，指导合理开发和保护渔业等方面有着重要作用。中期海洋智能预报主要分全球海洋要素模型、区域海洋要素模型、全球海洋生态模型、全球海浪模型，信息见表5-5。

  表 5-5 中期海洋智能预测模型信息

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
全球海洋要素模型	0m, 6m, 10m, 20m, 30m, 50m, 70m, 100m, 125m, 150m, 200m, 250m, 300m, 400m, 500m	T：海温(℃) S：海盐(PSU) U：海流经向速率(ms-1) V：海流纬向速率(ms-1)	SSH：海表高度(m)	24h	0.25°*0.25°	在60°S至65°N，180°W至180°E覆盖全球海洋主要海域（以下简称“全球海域”）
区域海洋要素模型	0m, 6m, 10m, 20m, 30m, 50m, 70m, 100m, 125m, 150m, 200m, 250m, 300m, 400m, 500m	T：海温(℃) S：海盐(PSU) U：海流经向速率(ms-1) V：海流纬向速率(ms-1)	SSH：海表高度(m)	24h	1/12°	特定区域

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
全球海洋生态模型	0m	/	Tca：总叶绿素浓度(mg/m3) Chl：叶绿素浓度(mg/m3) Dia ：硅藻浓度(mg/m3) Coc：颗石藻浓度(mg/m3) Cya：蓝澡浓度(mg/m3) Irn：铁浓度 (nano mole/L) Nit：硝酸盐浓度(micro mole/L) MLD：混合层深度(m)	24h	1°	在60°S至65°N，180°W至180°E覆盖全球海洋主要海域（以下简称“全球海域”）
全球海浪模型	0m	/	SWH有效波高 (m)	24h	0.5°	在60°S至65°N，180°W至180°E覆盖全球海洋主要海域（以下简称“全球海域”）

支持训练的模型清单见**表5-6**，您可根据具体使用场景选择合适的模型。例如区域海洋基础要素预测场景下，您可以选择**Pangu-AI4S-Ocean_Regional_24h-20241030**模型。

表 5-6 区域中期海洋智能预测模型的类型

模型名称	说明
Pangu-AI4S-Ocean_24h-20241130	2024年11月发布的版本，用于海洋基础要素预测，可支持1个实例部署推理。
Pangu-AI4S-Ocean_24h-3.1.0	2025年1月发布的版本，用于海洋基础要素预测，可支持1个实例部署推理。
Pangu-AI4S-Ocean_Regional_24h-20241130	2024年11月发布的版本，用于区域海洋基础要素预测，1个训练单元起训及1个实例部署。
Pangu-AI4S-Ocean-Regional_24h-3.1.0	2025年1月发布的版本，用于区域海洋基础要素预测，1个训练单元起训及1个实例部署。
Pangu-AI4S-Ocean_Ecology_24h-20241130	2024年11月发布的版本，用于海洋生态要素预测，可支持1个实例部署推理。
Pangu-AI4S-Ocean-Ecology_24h-3.1.0	2025年1月发布的版本，用于海洋生态要素预测，可支持1个实例部署推理。
Pangu-AI4S-Ocean_Swell_24h-20241130	2024年11月发布的版本，用于海浪预测，可支持1个实例部署推理。
Pangu-AI4S-Ocean-Swell_24h-3.1.0	2025年1月发布的版本，用于海浪预测，可支持1个实例部署推理。

科学计算大模型训练类型选择建议

目前，全球中期天气要素模型提供训练功能和推理功能，降水模型仅提供推理功能。

- **全球中期天气要素预测模型的训练类型选择建议：**

  全球中期天气要素预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的中期天气要素预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

  - 预训练：训练用于添加新的高空层次、高空变量或表面变量。如果您需要在现有模型中引入新要素，需要使用训练（重新训练模型）。在重训配置参数时，您可以选择新要素进行训练。请注意，所选的数据集必须包含您想要添加的新要素。此外，您还可以通过训练更改所有的模型参数，以优化模型性能。
  - 微调：微调是将新数据应用于已有模型的过程。它适用于不改变模型结构参数和引入新要素的情况。如果您有新的观测数据，可以使用微调来更新模型的权重，以适应新数据。

- **中期海洋智能预测模型的训练类型选择建议**：

  中期海洋智能预测模型的训练支持预训练、微调两种操作，如果直接使用平台预置的区域中期海洋智能预测模型不满足您的使用要求时，可以进行预训练或微调。预训练、微调操作的适用场景如下：

- 预训练：可以在重新指定深海变量、海表变量、以及深海层深、时间分辨率、水平分辨率以及区域范围，适用于想自定义自己的区域模型的场景，需预先准备好区域高精度数据。
- 微调：在已有模型的基础上添加新数据，它适用于不改变模型结构参数和引入新要素的情况，添加最新数据的场景。

5.2.2 创建科学计算大模型训练任务

创建科学计算大模型中期天气要素预测微调任务

创建科学计算大模型中期天气要素预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“创建训练任务”页面，参考表5-7完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 5-7 科学计算大模型中期天气要素预测微调训练参数说明

参数分类	参数名称	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“科学计算大模型”。
	场景	选择“中期天气要素预测”。
	训练类型	选择“微调”。
	基础模型	选择所需微调的基础模型，可从“已发布模型”或“未发布模型”中进行选择。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	层次	设置训练数据的层次信息。在“预训练”场景中，可以添加或去除高空层次，训练任务将根据配置的层次信息重新训练模型。
	高空变量	设置训练数据的高空变量信息。在“预训练”场景中，可以添加或去除新的高空变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的高空变量重新训练模型。

参数分类	参数名称	参数说明
	表面变量	设置训练数据的表面变量信息。在“预训练”场景中，可以添加或去除新的表面变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的表面变量重新训练模型。
	表面静态量	表面静态量默认包括地形高度、LAND_MASK和SOIL_TYPE，用于初始化模型状态并提供地表特性信息。当前不支持添加或去除这些静态量。 ● LAND_MASK：一个二维数组，表示模型网格中每个单元格是否是陆地。 ● SOIL_TYPE：表示地表土壤分类，影响土壤的物理和化学特性，如水分保持能力、热容量和导热性。
模型输出控制参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。取值范围：[1-1000]。
	损失类型	用来衡量模型预测结果与真实结果之间的差距的函数，提供MAE（平均绝对误差）、MSE（均方误差）两种损失函数。 ● MSE对于异常值非常敏感，因为它会放大较大的误差。因此，如果您数据中没有异常值，或者希望模型对大的误差给予更大的惩罚，可选择MSE。 ● 如果数据中存在异常值，或者希望模型对所有的误差都一视同仁，可选择MAE。
	表面变量相对高空变量的权重	指在模型训练过程中对表面变量相对于深海层变量赋予的权重，总Loss=高空Loss+surface_loss_weight*表面Loss。取值范围：(0.05, 10)。
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输入数据加噪音的尺度	定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。

参数分类	参数名称	参数说明
	给输出数据加噪音的概率	定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器种类	优化器种类	优化器是用于更新模型参数的算法，目前支持ADAM优化器。
	第一个动量矩阵的指数衰减率(beta1)	用于定义ADAM优化器中的一阶矩估计的指数衰减率。一阶矩估计相当于动量，可以加速梯度在相关方向的下降并抑制震荡。取值范围：(0,1)。
	第二个动量矩阵的指数衰减率(beta_2)	用于定义ADAM优化器中的二阶矩估计的指数衰减率。二阶矩估计相当于RMSProp，可以调整学习率。取值范围：(0,1)。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂，取值需≥0。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	学习率调整策略	用于选择学习率调度器的类型。学习率调度器可以在训练过程中动态地调整学习率，以改善模型的训练效果。目前支持CosineDecayLR调度器。
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型中期天气要素预测预训练任务

创建科学计算大模型中期天气要素预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“创建训练任务”页面，参考**表5-8**完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 5-8 科学计算大模型中期天气要素预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“科学计算大模型”。
	场景	选择“中期天气要素预测”。
	训练类型	选择“预训练”。
	基础模型	选择所需训练的基础模型，可从“已发布模型”或“未发布模型”中进行选择。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业，加工时需选择气象预处理算子。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	层次	设置训练数据的层次信息。在“预训练”场景中，可以添加或去除高空层次，训练任务将根据配置的层次信息重新训练模型。
	高空变量	设置训练数据的高空变量信息。在“预训练”场景中，可以添加或去除新的高空变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的高空变量重新训练模型。
	表面变量	设置训练数据的表面变量信息。在“预训练”场景中，可以添加或去除新的表面变量，选中后会在变量权重中增加或移除该变量，训练任务将根据配置的表面变量重新训练模型。

参数分类	参数名称	参数说明
	表面静态量	表面静态量默认包括地形高度、LAND_MASK 和 SOIL_TYPE，用于初始化模型状态并提供地表特性信息。当前不支持添加或去除这些静态量。 ● LAND_MASK：一个二维数组，表示模型网格中每个单元格是否是陆地。 ● SOIL_TYPE：表示地表土壤分类，影响土壤的物理和化学特性，如水分保持能力、热容量和导热性。
模型输出控制参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。取值范围：[1-1000]。
	损失类型	用来衡量模型预测结果与真实结果之间的差距的函数，提供MAE（平均绝对误差）、MSE（均方误差）两种损失函数。 ● MSE对于异常值非常敏感，因为它会放大较大的误差。因此，如果您数据中没有异常值，或者希望模型对大的误差给予更大的惩罚，可选择MSE。 ● 如果数据中存在异常值，或者希望模型对所有的误差都一视同仁，可选择MAE。
	表面变量相对高空变量的权重	指在模型训练过程中对表面变量相对于深海层变量赋予的权重，总Loss=高空Loss+surface_loss_weight*表面Loss。取值范围：(0.05, 10)。
模型结构参数	深度	用于定义深度学习网络的层数。数值越大，模型复杂性越高。模型参数量会增加。然而，这也会导致模型的结果文件变大，可能会占用大量的显存。在设置深度时，需要权衡模型的复杂性和显存的使用情况。推荐设置为[2, 6]。
	补丁尺度	用于将气象场划分为多个小块的大小，每个小块都会被模型单独处理。较大的patch_size意味着模型主干部分的一个网格代表更大范围的区域，但局部的细节信息可能会被忽略，较小的patch_size则相反。需要注意： ● 数据格式为[int,int,int]，第一个值需要大于0小于等于4，第二、三个参数都需要大于1小于等于20。 ● 在高方向patch_size[0]*window_size[0]需小于高空层次个数。 ● 在东西方向patch_size[2]*window_size[2]需能整除1440。

参数分类	参数名称	参数说明
	窗口尺度	用于定义模型主干网格中计算自注意力的区域大小。需注意： • 数据格式为[int,int,int]，第一个值需要大于0小于等于4，第二、三个参数需要大于1小于等于20。 • 在高方向patch_size[0]*window_size[0]需小于高空层次个数。 • 在东西方向patch_size[2]*window_size[2]需能整除1440。
	多头注意力头数	用于定义多头注意力机制中的头数。在设置这个参数时，需要注意init_channels要能够整除num_heads里的两个数。取值需大于1。
	第一层的通道数量	用于定义卷积神经网络中第一层卷积核的数量。在设置这个参数时，需要注意init_channels要能够整除num_heads里的两个数。调大此参数，模型会变大，可能会导致内存不足的问题。取值需大于0。注意此值调大可能会引起内存不足的场景，导致训练作业失败。
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输入数据加噪音的尺度	定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	给输出数据加噪音的概率	定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器种类	优化器种类	优化器是用于更新模型参数的算法，目前支持ADAM优化器。

参数分类	参数名称	参数说明
	第一个动量矩阵的指数衰减率(beta1)	用于定义ADAM优化器中的一阶矩估计的指数衰减率。一阶矩估计相当于动量，可以加速梯度在相关方向的下降并抑制震荡。取值范围：(0,1)。
	第二个动量矩阵的指数衰减率(beta_2)	用于定义ADAM优化器中的二阶矩估计的指数衰减率。二阶矩估计相当于RMSProp，可以调整学习率。取值范围：(0,1)。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂，取值需≥0。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。 预训练时，默认值为：0.00001，范围为[0, 0.001]
	学习率调整策略	用于选择学习率调度器的类型。学习率调度器可以在训练过程中动态地调整学习率，以改善模型的训练效果。目前支持CosineDecayLR调度器。
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域中期海洋智能预测微调任务

创建科学计算大模型区域中期海洋智能预测微调任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。

3. 在“创建训练任务”页面，参考表5-9完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 5-9 科学计算大模型区域中期海洋智能预测微调参数说明

参数分类	参数名称	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“科学计算大模型”。
	场景	选择“区域中期海洋智能预测”。
	训练类型	选择“微调”。
	基础模型	选择所需微调的基础模型，可从“已发布模型”或“未发布模型”中进行选择。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
	深海层深	海深层深是指海洋模型将整个水柱（从海面到海底）按一定深度间隔划分成多个层次，每个深度值代表模型在这个深度层进行计算和模拟。例如，"0m"代表海平面，"6m"代表在海平面以下6米处的一层，以此类推。范围包括：0m、6m、10m、20m、30m、50m、70m、100m、125m、150m、200m、250m、300m、400m、500m。
	深海变量	深海变量是用于模拟和描述海洋状态的关键物理量。 T：15层：海温(℃) S：15层：海盐(PSU) U：15层：海流经向速率 (ms-1) V：15层：海流纬向速率 (ms-1)

参数分类	参数名称	参数说明
	海表变量	海表变量用于描述海洋表层和其上方大气的状态的关键物理量。它们主要用于模拟和分析海洋表面的风速、温度、和气压等特征。 U10：1层：海表面10m经向风速(ms-1) V10：1层：海表面10m纬向风速(ms-1) T2m：1层：海表面2m温度 (℃) MSL：1层：平均海平面气压 (Pa) SP：1层：海表面气压 (Pa)
	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
模型输出控制参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。取值范围：[1-1000]。
	损失类型	用来衡量模型预测结果与真实结果之间的差距的函数，提供MAE（平均绝对误差）、MSE（均方误差）两种损失函数。 ● MSE对于异常值非常敏感，因为它会放大较大的误差。因此，如果您数据中没有异常值，或者希望模型对大的误差给予更大的惩罚，可选择MSE。 ● 如果数据中存在异常值，或者希望模型对所有的误差都一视同仁，可选择MAE。
	海表变量相对深海变量的权重	指在模型训练过程中对海表变量相对于深海层变量赋予的权重，总Loss=深海层Loss+surface_loss_weight*海表Loss。取值范围：(0.05, 10)。
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率，定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。

参数分类	参数名称	参数说明
	给输入数据加噪音的尺度	给输入数据加噪音的尺度，定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	给输出数据加噪音的概率	给输出数据加噪音的概率，定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	给输出数据加噪音的尺度，定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器参数	优化器种类	优化器种类。优化器是用于更新模型参数的算法，目前支持ADAM优化器。
	第一个动量矩阵的指数衰减率(beta1)	数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	第二个动量矩阵的指数衰减率(beta_2)	用于定义ADAM优化器中的二阶矩估计的指数衰减率。二阶矩估计相当于RMSProp，可以调整学习率。取值范围：(0,1)。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂，取值需≥0。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。
	学习率调整策略	用于选择学习率调度器的类型。学习率调度器可以在训练过程中动态地调整学习率，以改善模型的训练效果。目前支持CosineDecayLR调度器。
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。

参数分类	参数名称	参数说明
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

创建科学计算大模型区域中期海洋智能预测预训练任务

创建科学计算大模型区域中期海洋智能预测预训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击右上角“创建训练任务”。
3. 在“创建训练任务”页面，参考**表5-10**完成训练参数设置。

   其中，“数据配置”展示了各训练数据涉及到的全部参数，请根据具体前端页面展示的参数进行设置。

表 5-10 科学计算大模型区域中期海洋智能预测预训练参数说明

参数分类	参数名称	参数说明
训练配置	模型来源	选择“盘古大模型”。
	模型类型	选择“科学计算大模型”。
	场景	选择“区域中期海洋智能预测”。
	训练类型	选择“预训练”。
	基础模型	选择所需微调的基础模型，可从“已发布模型”或“未发布模型”中进行选择。
	模型水平分辨率	模型网格在水平方向上的精细程度，通常用来表示模拟或预测中空间网格的大小。根据训练数据和业务需求，自行定义模型水平分辨率，取值>0。
数据配置	训练数据	选择数据集中已发布的数据集，这里数据集需为再分析类型数据，同时需要完成加工作业。
	训练集	选择训练数据中的部分时间数据，训练数据集尽可能多一些。
	验证集	选择验证集中的部分时间数据，验证集数据不能跟训练集数据重合。
模型数据配置	深海层深	海深层深是指海洋模型将整个水柱（从海面到海底）按一定深度间隔划分成多个层次，每个深度值代表模型在这个深度层进行计算和模拟。例如，"0m"代表海平面，"6m"代表在海平面以下6米处的一层，以此类推。范围包括：0m、6m、10m、20m、30m、50m、70m、100m、125m、150m、200m、250m、300m、400m、500m。

参数分类	参数名称	参数说明
	深海变量	深海变量是用于模拟和描述海洋状态的关键物理量。 T：15层：海温(℃) S：15层：海盐(PSU) U：15层：海流经向速率 (ms-1) V：15层：海流纬向速率 (ms-1)
	海表变量	海表变量用于描述海洋表层和其上方大气的状态的关键物理量。它们主要用于模拟和分析海洋表面的风速、温度、和气压等特征。 U10：1层：海表面10m经向风速(ms-1) V10：1层：海表面10m纬向风速(ms-1) T2m：1层：海表面2m温度 (℃) MSL：1层：平均海平面气压 (Pa) SP：1层：海表面气压 (Pa)
	表面静态量	表面静态量默认支持地形高度、LAND_MASK、SOIL_TYPE，用于初始化模型状态和在模型运行过程中提供必要的地表特性信息，暂时不支持添加和去除。 其中，LAND_MASK是一个二维数组，通常用于表示模型网格中每个单元格是否是陆地。SOIL_TYPE是指地表土壤的分类，它影响土壤的物理和化学特性，如土壤的水分保持能力、热容量和导热性。
模型输出控制参数	训练轮数	表示完成全部训练数据集训练的次数。每个轮次都会遍历整个数据集一次。取值范围：[1-1000]。
	损失类型	用来衡量模型预测结果与真实结果之间的差距的函数，提供MAE（平均绝对误差）、MSE（均方误差）两种损失函数。 ● MSE对于异常值非常敏感，因为它会放大较大的误差。因此，如果您数据中没有异常值，或者希望模型对大的误差给予更大的惩罚，可选择MSE。 ● 如果数据中存在异常值，或者希望模型对所有的误差都一视同仁，可选择MAE。
	海表变量相对深海变量的权重	指在模型训练过程中对海表变量相对于深海层变量赋予的权重，总Loss=深海层Loss+surface_loss_weight*海表Loss。取值范围：(0.05, 10)。
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。

参数分类	参数名称	参数说明
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率，定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输入数据加噪音的尺度	给输入数据加噪音的尺度，定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	给输出数据加噪音的概率	给输出数据加噪音的概率，定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	给输出数据加噪音的尺度，定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器参数	优化器种类	优化器种类。优化器是用于更新模型参数的算法，目前支持ADAM优化器。
	第一个动量矩阵的指数衰减率(beta1)	数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	第二个动量矩阵的指数衰减率(beta_2)	用于定义ADAM优化器中的二阶矩估计的指数衰减率。二阶矩估计相当于RMSProp，可以调整学习率。取值范围：(0,1)。
	权重衰减系数	通过在损失函数中加入与模型权重大小相关的惩罚项，鼓励模型保持较小的权重，防止过拟合或模型过于复杂，取值需≥0。
	学习率	学习率决定每次训练中模型参数更新的幅度。 选择合适的学习率至关重要： ● 如果学习率过大，模型可能无法收敛。 ● 如果学习率过小，模型的收敛速度将变得非常慢。 预训练时，默认值为：0.00001，范围为[0, 0.001]。

参数分类	参数名称	参数说明
	学习率调整策略	用于选择学习率调度器的类型。学习率调度器可以在训练过程中动态地调整学习率，以改善模型的训练效果。目前支持CosineDecayLR调度器。
变量权重	变量权重	训练数据设置完成后，会显示出各变量以及默认的权重。您可以基于变量的重要情况调整权重。
资源配置	训练单元	选择训练模型所需的训练单元。 当前展示的完成本次训练所需要的最低训练单元要求。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 参数填写完成后，单击“立即创建”。
5. 创建好训练任务后，页面将返回“模型训练”页面，可随时查看当前任务的状态。

5.2.3 查看科学计算大模型训练状态与指标

模型启动训练后，可以在模型训练列表中查看训练任务的状态，单击任务名称可以进入详情页查看训练结果、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表5-11。

表 5-11 训练状态说明

训练状态	训练状态含义
初始化	模型训练任务正在进行初始化配置，准备开始训练。
排队中	模型训练任务正在排队，请稍等。
运行中	模型正在训练中，训练过程尚未结束。
停止中	模型训练正在停止中。
已停止	模型训练已被用户手动停止。
失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已完成	模型训练已完成。

查看训练指标

对于已完成训练，训练状态是“训练完成”状态的任务，单击任务名称，可在“训练结果”页面查看训练指标，不同模型的训练指标介绍请参见表5-12。

图 5-3 查看训练指标

表 5-12 训练指标说明

模型	训练指标	指标说明
科学计算大模型	Loss	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。这里代表高空Loss（深海Loss）和表面Loss（海表Loss）的综合Loss。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	高空Loss（深海Loss）	高空Loss（深海Loss）是衡量模型在高空层次变量或在深海变量预测结果与真实结果之间差距的指标。该值越小，表示模型在高空（深海）变量的预测精度越高。
	表面Loss（海表Loss）	表面Loss（海表Loss）是衡量模型在表面层次变量或在海表变量预测结果与真实结果之间差距的指标。该值越小，表示模型在表面（海表）变量的预测精度越高。
	RMSE	均方根误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的平方误差的平均值的平方根。该值越小，代表模型性能越好。
	MAE	平均绝对误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的绝对误差的平均值。该值越小，代表模型性能越好。

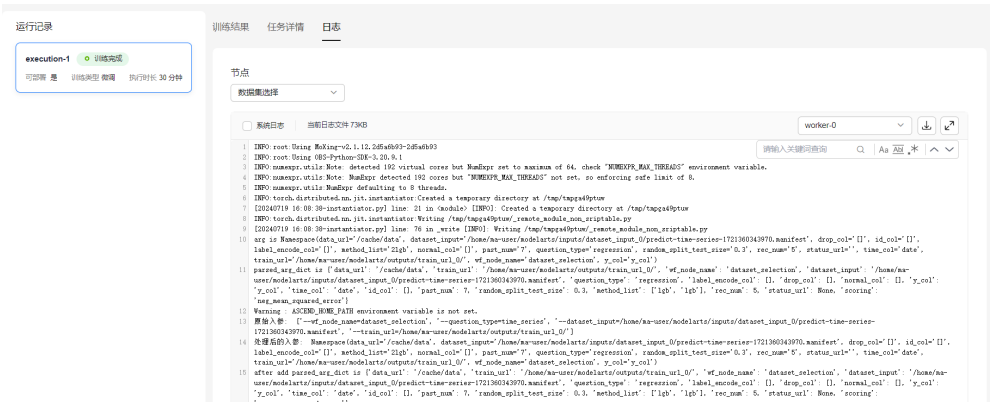
模型	训练指标	指标说明
	ACC	ACC（异常相关系数，距平相关系数，Anomaly Correlation Coefficient）是一个重要的统计指标，用于衡量预报系统的质量。它通过计算预报值与观测值之间的相关性来评估预报的准确性。ACC的计算涉及到预报值、观测值和气候平均值的差异，其值范围从-1到+1，值越接近+1表示预报与观测的一致性越好，值为0表示没有相关性，而负值则表示反向相关。
	RQE	衡量预测值与真实值之间差距的指标。它是所有单个观测的相对误差的平方和。该值越小，代表模型性能越好。

获取训练日志

单击训练任务名称，可以在“日志”页面查看训练过程中产生的日志。对于训练异常或失败的任务也可以通过训练日志定位训练失败的原因。典型训练报错和解决方案请参见**科学计算大模型训练常见报错与解决方案**。

训练日志可以按照不同的节点（训练阶段）进行筛选查看。分布式训练时，任务被分配到多个工作节点上进行并行处理，每个工作节点负责处理一部分数据或执行特定的计算任务。日志也可以按照不同的工作节点（如worker-0表示第一个工作节点）进行筛选查看。

图 5-4 获取训练日志



5.2.4 发布训练后的科学计算大模型

科学计算大模型训练完成后，需要执行发布操作，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击模型名称进入任务详情页。
3. 单击进入“训练结果”页签，单击“发布”。

图 5-5 训练结果页面

4. 填写资产名称、描述，选择对应的可见性，单击“确定”发布模型。
   发布后的模型会作为模型资产同步显示在“空间资产 > 模型”列表中。

5.2.5 管理科学计算大模型训练任务

在训练任务列表中，任务创建者可以对创建好的任务进行编辑、启动、克隆（复制训练任务）、重试（重新训练任务）和删除操作。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型训练”，进入模型训练页面，可进行如下操作：
   - 编辑。单击操作列的“编辑”，可以修改模型的训练参数、训练数据以及基本信息等。
   - 克隆。单击操作列的“更多 > 克隆”，参照**创建科学计算大模型训练任务**填写参数，可以复制当前训练任务。
   - 停止。单击操作列的“更多 > 停止”，可以停止处于“排队中”或“运行中”状态的任务。
   - 重试。单击操作列的“更多 > 重试”，可以重试处于“失败”状态的节点，重试该节点的训练。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前不需要的训练任务。

5.2.6 科学计算大模型训练常见报错与解决方案

科学计算大模型训练常见报错及解决方案请详见**表5-13**。

表 5-13 科学计算大模型训练常见报错与解决方案

常见报错	问题现象	原因分析	解决方案
创建训练任务时，数据集列表为空	创建训练任务时，数据集选择框中显示为空，无可用的训练数据集。	数据集未发布。	请提前创建与大模型对应的训练数据集，并完成数据集发布操作。

常见报错	问题现象	原因分析	解决方案
训练日志提示“root: XXX valid number is 0”报错	日志提示“root: XXX valid number is 0”，表示训练集/验证集的有效样本量为0，例如： INFO: root: Train valid number is 0.	该日志表示数据集中的有效样本量为0，可能有如下原因： • 数据未标注。 • 标注的数据不符合规格。	请检查数据是否已标注或标注是否符合算法要求。

5.3 部署科学计算大模型

5.3.1 创建科学计算大模型部署任务

平台支持部署训练后的模型或预置模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，参考表5-14完成部署参数设置。

表 5-14 科学计算大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	模型来源	选择“盘古大模型”。
	模型类型	选择“科学计算大模型”。

参数分类	部署参数	参数说明
	场景	选择模型场景，分为“全球中期天气要素预测”、“全球中期降水预测”、“全球中期海洋智能预测”、“区域中期海洋智能预测”、“全球中期海洋生态智能预测”、“全球中期海浪智能预测”。 ● 全球中期天气要素预测模型可以选择1个或者多个模型进行部署。
	部署模型	在“从资产选模型”选择所需模型。
	部署方式	支持“云上部署”和“边缘部署”，其中，云上部署指算法部署至平台提供的资源池中。边缘部署指算法部署至客户的边缘设备中（仅支持边缘部署的模型可配置边缘部署）。 ● 部分模型资产支持边缘部署方式，若选择“边缘部署”： – 本地挂载路径（选填）：在容器内部将卷挂载的本地路径。挂载后，容器中的应用程序可以通过这个路径访问宿主机上的数据。 – 资源池：选择部署模型所需的边缘资源池，创建边缘资源池步骤请详见**创建边缘资源池**。 – CPU：部署需要使用的最小CPU值（物理核）。 – 内存：部署需要使用的最小内存值。 – Ascend：部署使用的NPU数量。 – 负载均衡：创建负载均衡步骤请详见**步骤5：创建负载均衡**。 – 实例数：设置部署模型时所需的实例数。
	作业输入方式	选择 “OBS”表示从OBS中读取数据。
	作业输出方式	选择 “OBS”表示将输出结果存储在OBS中。
	作业配置参数	设置模型部署参数信息，平台已给出默认值。
安全护栏	选择模式	安全护栏保障模型调用安全。
	选择类型	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置	计费模式	包年包月计费模式。
	实例数	设置部署模型时所需的实例数。

参数分类	部署参数	参数说明
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

4. 参数填写完成后，单击“立即部署”。

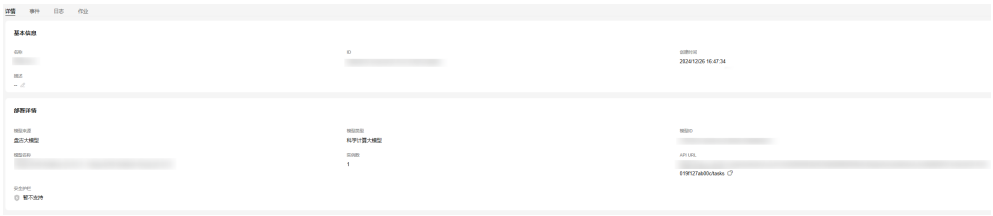
5.3.2 查看科学计算大模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建科学计算大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。
3. 可单击模型名称可进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

图 5-6 部署详情



5.3.3 管理科学计算大模型部署任务

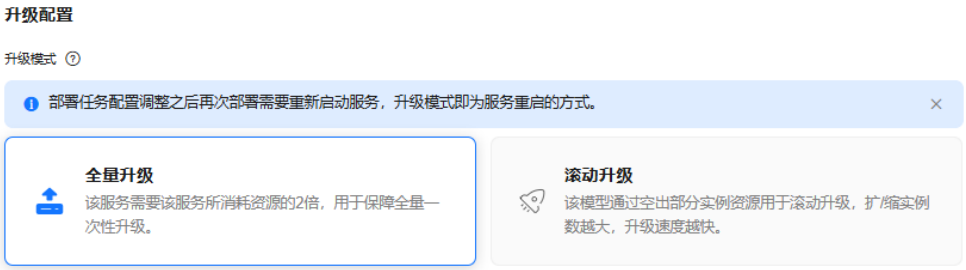
模型更新

完成**创建科学计算大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 部署模型”中，可选择模型以替换当前已部署的模型。

5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 5-7 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成[创建科学计算大模型部署任务]后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 5-8 升级模式

> **说明**
>
> 升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成**创建科学计算大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确定”。

5.4 调用科学计算大模型

5.4.1 使用“能力调测”调用科学计算大模型

能力调测功能支持用户调用预置或训练后的科学计算大模型。使用该功能前，请完成模型的部署操作，步骤详见**创建科学计算大模型部署任务**。

使用“能力调测”调用科学计算大模型可实现包括全球中期天气要素预测、全球中期降水预测、全球海洋要素、区域海洋要素、全球海洋生态、全球海浪高度场景的预测能力。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“能力调测”，单击“科学计算”页签。
3. 根据不同场景完成页面参数配置。
   - **天气/降水预测场景**的参数配置，请参考**表5-15**。

表 5-15 科学计算大模型能力调测参数说明（天气/降水预测）

参数	说明
场景	支持选择全球中期天气要素预测、全球中期降水预测。 • 全球中期天气要素预测：通过该模型可以对未来一段时间的天气进行预测。 • 全球中期降水预测：通过该模型可以对未来一段时间的降水情况进行预测。
模型服务	支持选择用于启动推理作业的模型。 • 中期天气要素模型包括1h分辨率、3h分辨率、6h分辨率、24小时分辨率模型，即以起报时刻开始，分别可以逐1h、3h、6h、24h往后进行天气要素的预测。 • 中期天气要素模型包括6h分辨率模型，即以起报时刻开始，可以逐6h往后进行降水情况的预测。
结果存储路径	用于存放模型推理结果的OBS路径。

参数	说明
输入数据	支持选择用于存放作为初始场数据的文件路径。
预报天数	支持选择以起报时间点为开始，对天气要素或降水进行预报的天数，范围为1~14天。
起报时间	支持选择多个起报时间作为推理作业的开始时间，每个起报时间需为输入数据中存在的时间点。
表面变量	支持选择推理结果输出的表面变量，包括10m u风、10m v风、2米温度、海平面气压，没有选择的变量推理结果将不输出。
高空变量	设置高空变量参数，包括：4个表面层特征（10m u风、10m v风、2米温度、海平面气压），13高空层次（1000、925、850、700、600、500、400、300、250、200、150、100、50hPa）的5个高空层特征（重力位势、u风、v风、比湿、温度），分辨率为25km*25km的网格数据。
集合预报	用于选择是否开启集合预报。 在气象预报中，集合预报是指对初始场加入一定程序的扰动，使其生成一组由不同初始场预报的天气预报结果，从而提供对未来天气状态的概率信息。这种方法可以更好地表达预报的不确定性，从而提高预报的准确性和可靠性。
集合成员数	用于选择生成预报的不同初始场的数量，取值为2~10。
扰动类型	用于选择生成集合预报初始场的扰动类型，包括perlin加噪和CNOP加噪两种方式。 ● Peilin噪音通过对输入数据（比如空间坐标）进行随机扰动，让模拟出的天气接近真实世界中的变化。 ● CNOP噪音通过在初始场中引入特定的扰动来研究天气系统的可预报性，会对扰动本身做一定的评判，能够挑选出预报结果与真实情况偏差最大的一类初始扰动。这些扰动不仅可以用来识别最可能导致特定天气或气候事件的初始条件，还可以用来评估预报结果的不确定性。
初始扰动数量	用于选择集合预报的CNOP初始扰动数量。 ● 在CNOP的加噪方式中，会先对初始场进行一定数量的加噪得到一组加噪后的初始场，然后从这组初始场中选择能量变化最大的初始场作为集合预报的初始场，启动推理作业。
ensemble_noise_perlin_scale	用于选择集合预报的Perlin加噪强度。
ensemble_noise_perlin_x	用于选择集合预报的Perlin加噪x经度方向的尺度。

参数	说明
ensemble_noise_perlin_octave	用于选择集合预报的Perlin加噪octave。Perlin噪音的octave指的是噪音的频率，在生成Perlin噪音时，可以将多个不同频率的噪音叠加在一起，以增加噪音的复杂度和细节。每个频率的噪音称为一个octave，而叠加的octave数越多，噪音的复杂度也就越高。
ensemble_noise_perlin_y	用于选择集合预报的Perlin加噪y纬度方向的尺度。
输出设置	用于选择是否输出图片结果。

天气/降水预测场景的参数配置示例如下：

图 5-9 调测科学计算大模型示例 1（天气/降水预测）

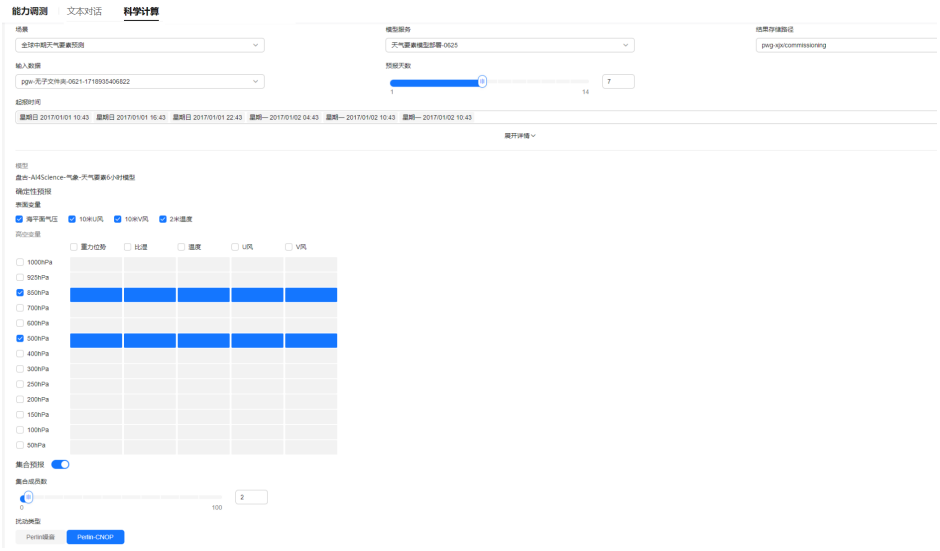
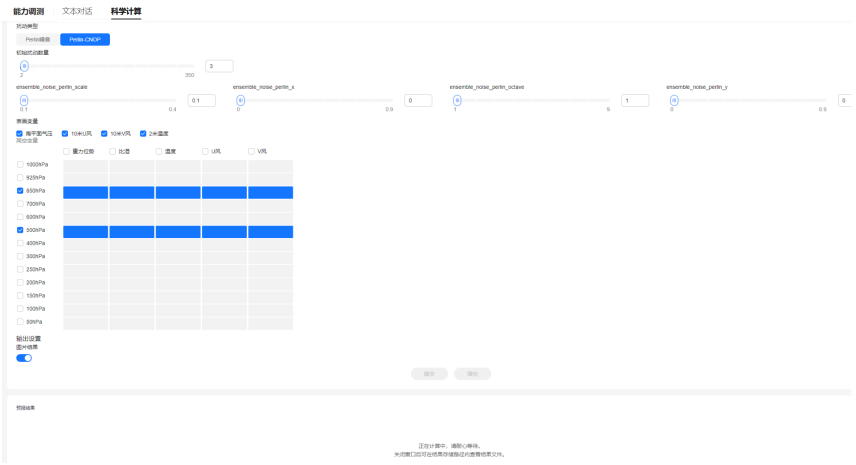


图 5-10 调测科学计算大模型示例 2（天气/降水预测）



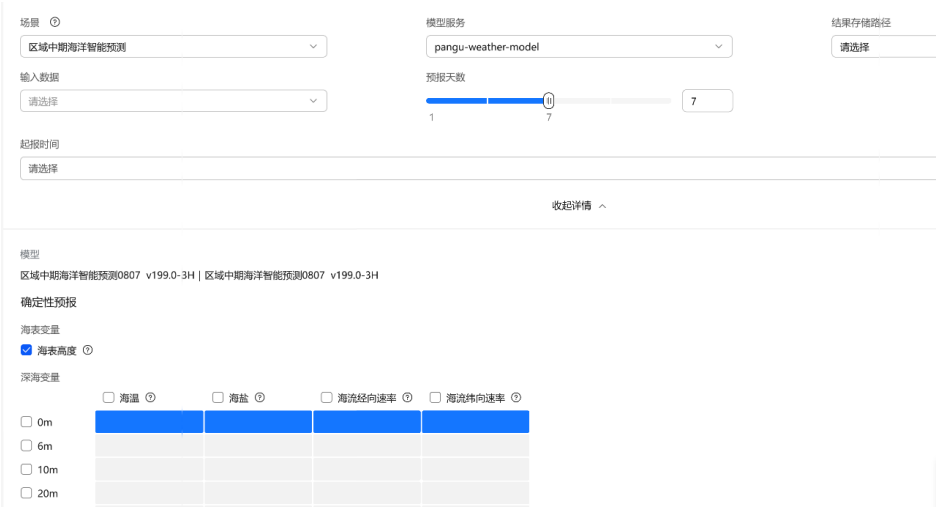
- **海洋类预测场景**的参数配置，请参考**表5-16**。

表 5-16 科学计算大模型能力调测参数说明（海洋类预测）

参数	说明
场景	支持选择全球海洋要素、区域海洋要素、全球海洋生态、全球海浪高度。 ● 全球海洋要素：实现预测全球范围内海面高度， 温度、盐度、海流速度纬向分量和海流速度经向分量变量。 ● 区域海洋要素：实现预测特定区域范围内海面高度， 温度、盐度、海流速度纬向分量和海流速度经向分量变量。 ● 全球海洋生态：实现预测全球范围内的叶绿素浓度、硅藻浓度等8种生态变量。 ● 全球海浪高度：实现预测有效波高的变量。
模型服务	支持选择用于启动推理作业的模型。
结果存储路径	用于存放模型推理结果的OBS路径。
输入数据	支持选择用于存放作为初始场数据的文件路径。
预报天数	支持选择以起报时间点为开始，对海洋模型预测参数进行预报的天数，范围为1~14天。
起报时间	支持选择多个起报时间作为推理作业的开始时间，每个起报时间需为输入数据中存在的时间点。
海表变量	用于描述海洋表面及其生态系统状态的具体指标，尤其是在海洋模型中用于模拟海洋生态和物理过程的输入变量。包括海平面气压、海表高度、总叶绿素浓度、叶绿素浓度、硅藻浓度、颗石藻浓度、蓝藻浓度、铁浓度、硝酸盐浓度、混合层深度、海表高度、有效波高等指标。不同模型的指标以页面展示为准。
深海变量	用于描述海洋深层的物理和化学特性，这些参数在海洋模型中用于模拟海洋内部的动态和状态。包括海温、海盐、海流径向速率、海流纬向速率等。
输出设置	用于选择是否输出图片结果。

海洋类预测场景的参数配置示例如下：

图 5-11 调测科学计算大模型示例（海洋类预测）



5.4.2 使用 API 调用科学计算大模型

预置模型或训练后的模型部署成功后，可以使用API调用科学计算大模型。

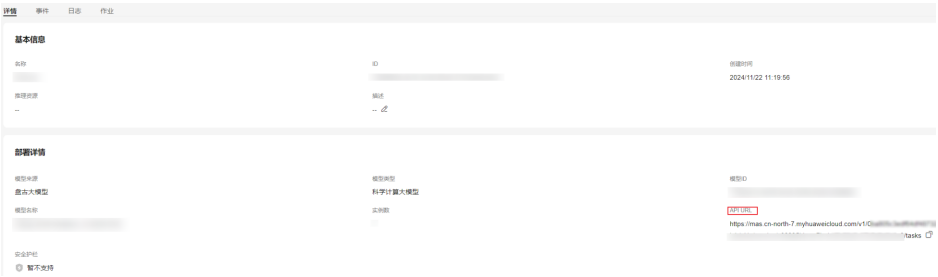
获取调用路径

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

   在左侧导航栏中选择“模型开发 > 模型部署”。

   - 获取已部署模型的调用路径。在“我的服务”页签，单击状态为“运行中”的模型名称，在“详情”页签，可获取模型调用路径，如图5-12。

     图 5-12 获取已部署模型的调用路径

   - 获取预置服务的调用路径。在“预置服务”页签中，选择所需调用的科学计算大模型，单击“调用路径”，在“调用路径”弹窗可获取模型调用路径，如图5-13。

图 5-13 获取预置服务的调用路径

使用 Postman 调用 API

1. 在Postman中新建POST请求，并填入模型调用路径，详见**获取调用路径**。
2. 调用API有两种认证方式，包括Token认证和AppCode认证。其中，AppCode认证的使用场景为当用户部署的API服务期望开放给其他用户调用时，原有Token认证无法支持，可通过AppCode认证调用请求。

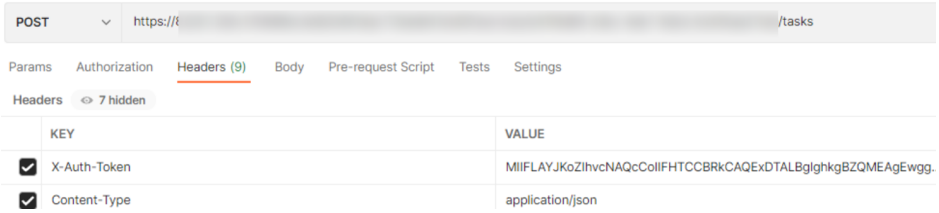
   参考**表5-17**填写请求Header参数。

表 5-17 请求 Header 参数填写说明

认证方式	参数名	参数值
Token认证	Content-Type	application/json
	X-Auth-Token	Token值，参考《API参考》文档“如何调用REST API > 认证鉴权 > Token认证”章节获取Token。
AppCode认证	Content-Type	application/json
	X-Apig-AppCode	AppCode值，获取AppCode步骤如下： 1. 登录ModelArts Studio平台，进入所需空间。 2. 在左侧导航栏中选择“模型开发 > 应用接入”，单击界面右上角“创建应用接入”。 3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。 4. 在“应用接入”列表的“APP Code”操作列中可获取APPCode值。

如**图5-14**，为Token认证方式的请求Header参数填写示例。

图 5-14 配置请求参数



3. 在Postman中选择“Body > raw”选项，参考以下代码填写请求Body。API参数说明详见《API参考》文档。

```
{
   "name": "test-task624",
   "input": {
      "type": "obs",
      "data": [
         {
            "bucket": "pangu-weather-data",
            "path": "test/"
         }
      ]
   },
   "output": {
      "obs": {
         "bucket": "pangu-weather-test",
         "path": "output/"
      }
   },
      "config": {
      "start_time_begin": "2022010100",
      "start_time_end": "2022010106",
      "start_time_interval_hours": 6,
      "forecast_lead_hours": 168
   }
}
```

4. 单击Postman界面“Send”，发送请求。科学计算大模型API调用成功后，会返回任务id参数**task_id**，可获取任务ID参数值。
5. 在Postman中新建一个GET请求，填入域名（将**获取调用路径**中获取的URL去除末尾的“/tasks”即为该域名），设置请求Header参数和任务ID参数。单击Postman界面的“Send”发送请求，以获取科学计算大模型的调用结果。

```
查询科学计算大模型调用详情API
GET /tasks/{task_id}
```

6 开发盘古专业大模型

6.1 部署专业大模型

6.1.1 创建专业大模型部署任务

平台支持部署预置的专业大模型，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，参考表6-1完成部署参数设置。

表 6-1 专业大模型部署参数说明

参数分类	部署参数	参数说明
部署配置	模型来源	选择“盘古大模型”。
	模型类型	选择“专业大模型 > BI专业大模型”或“专业大模型 > 搜索专业大模型”。
	部署模型	在“从资产选模型”选择所需模型。
	部署方式	云上部署：算法部署至平台提供的资源池中。
安全护栏	选择模式	安全护栏保障模型调用安全。
	选择类型	当前支持安全护栏基础版，内置了默认的内容审核规则。
资源配置	计费模式	包年包月计费模式。

参数分类	部署参数	参数说明
	实例数	设置部署模型时所需的实例数。
订阅提醒	订阅提醒	该功能开启后，系统将在任务状态更新时，通过短信或邮件将提醒发送给用户。
基本信息	服务名称	设置部署任务的名称。
	描述（选填）	设置部署任务的描述。

4. 参数填写完成后，单击“立即部署”。

6.1.2 查看专业大模型部署任务详情

部署任务创建成功后，可以查看大模型部署的任务详情，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，完成**创建专业大模型部署任务**后，可以查看模型的部署状态。

   当状态显示为“运行中”时，表示模型已成功部署。此过程可能需要较长时间，请耐心等待。
3. 可单击模型名称可进入详情页，查看模型的部署详情、部署事件、部署日志等信息。

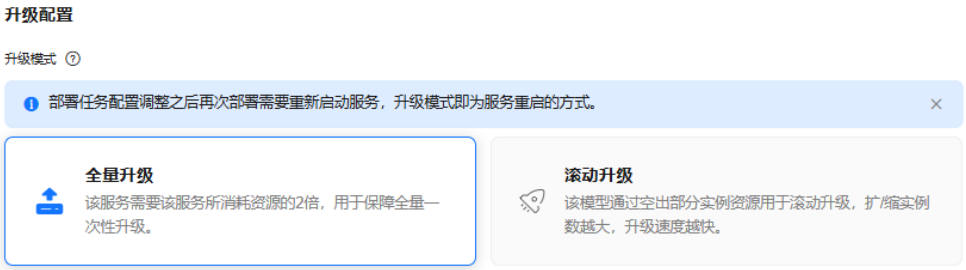
6.1.3 管理专业大模型部署任务

模型更新

完成**创建专业大模型部署任务**后，可以替换已部署的模型并升级配置，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“模型更新”，进入“模型更新”页面。
4. 在“可修改配置 > 部署模型”中，可选择模型以替换当前已部署的模型。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 6-1 升级模式



说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

修改部署配置

完成**创建专业大模型部署任务**后，可以修改已部署模型的描述信息并升级配置，但不可替换模型。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“修改部署”，进入“修改部署”页面。
4. 在“可修改配置”中，可修改已部署模型的描述信息。
5. 在“升级配置”中，选择以下两种升级模式：
   - **全量升级**：新旧版本服务同时运行，直至新版本完全替代旧版本。在新版本部署完成前，旧版本仍可使用。需要该服务所消耗资源的2倍，用于保障全量一次性升级。
   - **滚动升级**：部分实例资源空出用于滚动升级，逐个或逐批停止旧版本并启动新版本。滚动升级时可修改实例数。选择缩实例升级时，系统会先删除旧版本，再进行升级，期间旧版本不可使用。

图 6-2 升级模式

说明

升级配置后，需重新启动该部署任务，升级模式即为重启的方式。

模型部署实例扩缩容

完成**创建专业大模型部署任务**后，可以对已部署模型的实例进行扩缩容，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击模型名称，进入模型详情页面。
3. 单击右上角“扩缩容”，进入“扩缩容”页面，修改实例数，单击“确定”。

7 开发盘古大模型提示词工程

7.1 什么是提示词工程

提示词工程简介

提示词工程（Prompt Engineering）是一个较新的学科，应用于开发和优化提示词（Prompt），帮助用户有效地将大语言模型用于各种应用场景和研究领域。掌握提示词工程相关技能将有助于用户更好地了解大语言模型的能力和局限性。

提示词工程不仅是关于设计和研发提示词，它包含了与大语言模型交互和研发的各种技能和技术。提示工程在实现和大语言模型交互、对接，以及理解大语言模型能力方面都起着重要作用。用户可以通过提示词工程来提高大语言模型的安全性，还可以赋能大语言模型，如借助专业领域知识和外部工具来增强大语言模型的能力。

提示词基本要素

您可以通过简单的提示词（Prompt）获得大量结果，但结果的质量与您提供的信息数量和完善度有关。一个提示词可以包含您传递到模型的指令或问题等信息，也可以包含其他种类的信息，如上下文、输入或示例等。您可以通过这些元素来更好地指导模型，并因此获得更好的结果。提示词主要包含以下要素：

- **指令**：希望模型执行的特定任务或指令，如总结、提取、生成等。
- **上下文**：包含外部信息或额外的上下文信息，引导语言模型更好地响应。
- **输入数据**：用户输入的内容或问题。
- **输出指示**：指定输出的类型或格式。

提示词所需的格式取决于您希望语言模型完成的任务类型，并非所有以上要素都是必须的。

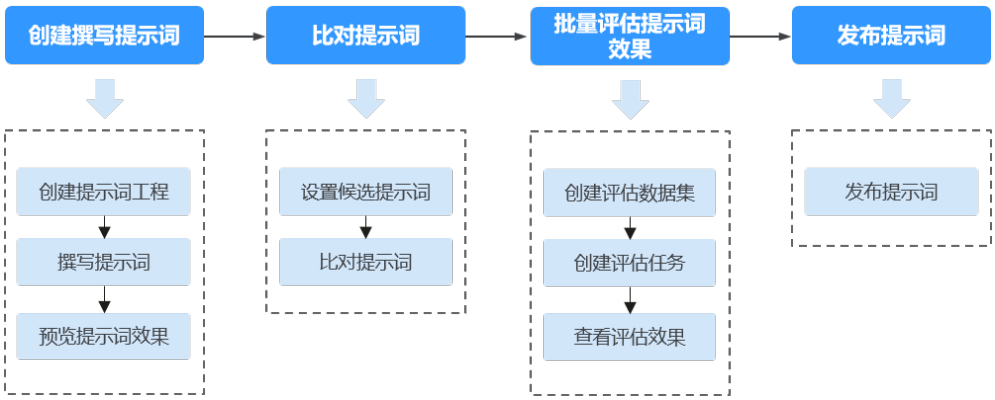
提示词工程使用流程

ModelArts Studio大模型开发平台可以辅助用户进行提示词撰写、比较和评估等操作，并对提示词进行保存和管理。

表 7-1 功能说明

功能	说明
提示词工程任务管理	提示词工程平台以提示词工程任务为管理维度，一个任务代表一个场景或一个调优需求，在提示词工程任务下可以进行提示词的调优、比较和评估。 提示词工程任务管理支持工程任务的创建、查询、修改、删除。
提示词撰写	提示词调优支持对提示词文本的编辑、提示词变量设置、提示词结果生成和调优历史记录管理。
提示词候选	提示词候选支持用户对调优后初步筛选的提示词进行候选管理，每个工程任务下可以保存上限9个候选提示词，进一步基于候选提示词进行比较和评估。
提示词比较	提示词比较支持选择两个候选提示词对其文本和参数进行比较，支持对选择的候选提示词设置相同变量值查看效果。
提示词评估	提示词评估以任务维度管理，支持评估任务的创建、查询、修改、删除。支持创建评估任务，选择候选提示词和需要使用的变量数据集，设置评估算法，执行任务自动化对候选提示词生成结果和结果评估。
提示词管理	提示词管理支持用户对满意的候选提示词进行保存管理，同时支持提示词的查询、删除。

图 7-1 提示词工程使用流程



7.2 获取提示词模板

平台提供了多种任务场景的提示词模板，可以帮助用户更好地利用大模型的能力，引导模型生成更准确、更有针对性的输出，从而提高模型在特定任务上的性能。

在创建提示词工程前，可以先使用预置的提示词模板，或基于提示词模板进行改造 。如果提示词模板满足不了使用需求可再单独创建。

提示词模板可在平台“Agent 开发 > 提示词工程 > 提示词模板”中获取。

7.3 撰写提示词

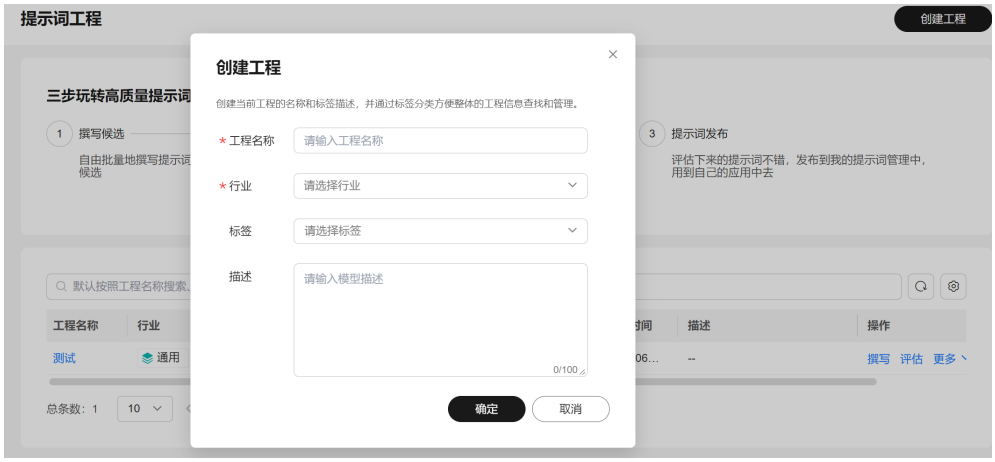
7.3.1 创建提示词工程

通过精心设计和优化提示词，可以引导大模型生成用户期望的输出。提示词工程任务的目标是通过设计和实施一系列的实验，来探索如何利用提示词来提高大模型在各种任务上的表现。

撰写提示词前需要先创建提示词工程，用于对提示词进行统一管理。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”，单击界面右上角“创建工程”。
3. 输入工程名称、描述，选择行业、标签后。单击“确定”完成工程创建。

图 7-2 创建提示词工程



7.3.2 撰写提示词

提示词是用来引导模型生成的一段文本。撰写的提示词应该包含任务或领域的关键信息，如主题、风格、格式等。

撰写提示词时，可以设置提示词变量。即在提示词中通过添加占位符{{ }}标识表示一些动态的信息，让模型根据不同的情况生成不同的文本，增加模型的灵活性和适应性。例如，将提示词设置为“你是一个旅游助手，需要给用户介绍旅行地的风土人情。请介绍下{{location}}的风土人情。”在评估提示词效果时，可以通过批量替换{{location}}的值，来获得模型回答，提升评测效率。

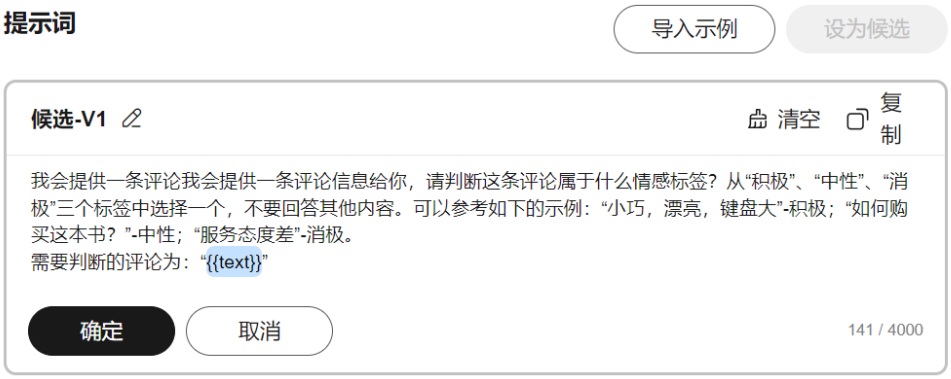
同时，撰写提示词过程中，可以通过设置模型参数来控制模型的生成行为，如调整温度、核采样、最大Token限制等参数。模型参数的设置会影响模型的生成质量和多样性，因此需要根据不同的场景进行选择。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 7-3 提示词工程

4. 在提示词撰写区域输入提示词文本，可以插入若干个变量，变量需要使用占位符{{ }}标识。

图 7-4 撰写提示词



5. 撰写完成后，单击“确定”，平台会自动识别插入的变量。提示词中识别的变量将展示在变量定义区域。

   变量名称可以进行修改，如添加备注信息以便更好理解变量的作用。

图 7-5 变量定义

说明

变量定义区域展示的是整个工程任务下定义的变量信息，候选提示词中关联的变量也会进行展示，候选提示词相关操作请参见**设置候选提示词**。

同一个提示词工程中，定义的变量不能超过20个。

6. 在“模型”区域，单击“设置”，设置提示词输入的模型和模型参数。

图 7-6 模型设置

7.3.3 预览提示词效果

提示词撰写完成后，可以通过输入具体的变量值，组成完整的提示词，查看不同提示词在模型中的使用效果。

1. 在撰写提示词页面，找到页面右侧变量输入区域，在输入框中输入具体的变量值信息。

   输入变量值后预览区域会自动组装展示提示词。也可以直接选择已创建的变量集填入变量值信息，变量集是一个excel文件，每行数据是需要输入的变量值信息，可以通过“导入”功能进行上传。

图 7-7 效果预览

2. 单击“查看效果”，输出模型回复结果，用户可以基于预览的效果调整提示词文本和变量。

7.4 横向比较提示词效果

7.4.1 设置候选提示词

用户可以将效果较好的提示词设为候选提示词，并对提示词进行比对，以查看其效果。

说明

每个工程任务下候选提示词上限9个，达到上限9个时需要删除其他候选提示词才能继续添加。

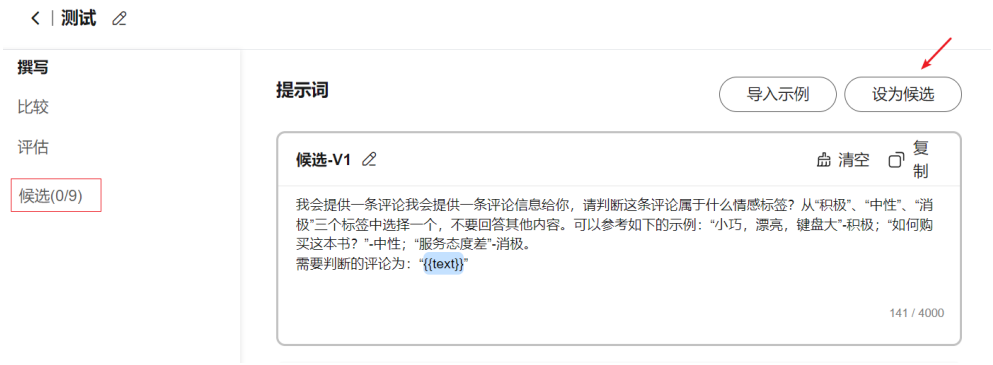
1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 7-8 提示词工程

4. 在提示词撰写区域，单击“设为候选”，将当前撰写的提示词设置为候选提示词。

   候选状态的提示词将保存至左侧导航栏的“候选”中。

图 7-9 设为候选



7.4.2 横向比较提示词效果

将设置为候选的提示词横向比对，获取提示词的差异性和效果。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示词开发”。

3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

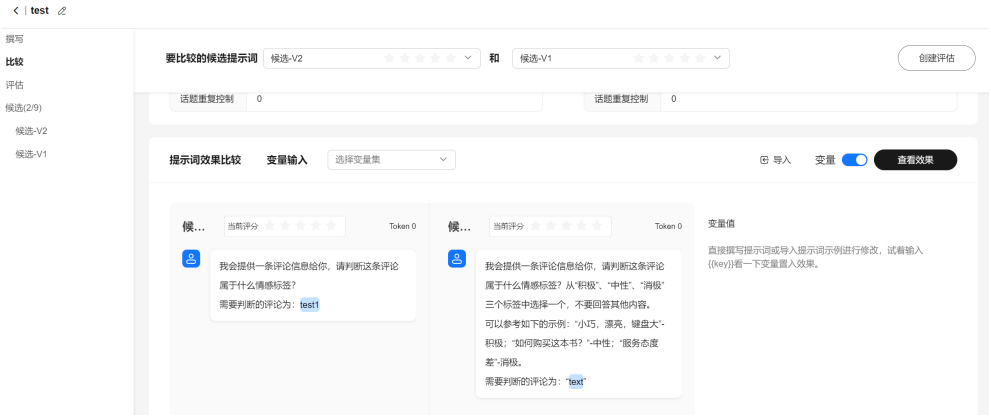
图 7-10 提示词工程

4. 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，勾选需要进行横向比对的提示词，并单击“横向比较”。

图 7-11 横向比较

5. 进入到横向比较页面，下拉页面至“提示词效果比较”模块，比较提示词的效果，输入相同的变量值，查看两个提示词生成的结果。

图 7-12 横向比对提示词效果



7.5 批量评估提示词效果

7.5.1 创建提示词评估数据集

批量评估提示词效果前，需要先上传提示词变量数据文件用于创建对应的评估数据集。

提示词变量是一种可以在文本生成中动态替换的占位符，用于根据不同的场景或用户输入生成不同的内容。其中，变量名称可以是任意的文字，用于描述变量的含义或作用。

提示词评估数据集约束限制

- 上传文件限xlsx格式。
- 数据行数不小于10行，不大于50行。
- 数据不允许相同表头，表头数量小于20个。
- 数据单条文本长度不超过1000。

说明

创建数据集时会对相关限制条件进行校验。

数据参考格式如下：

图 7-13 数据参考格式

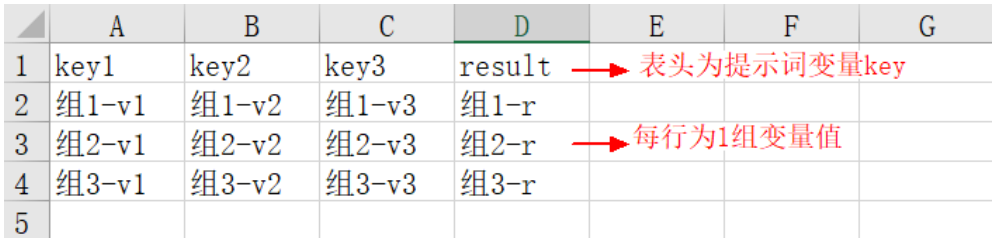
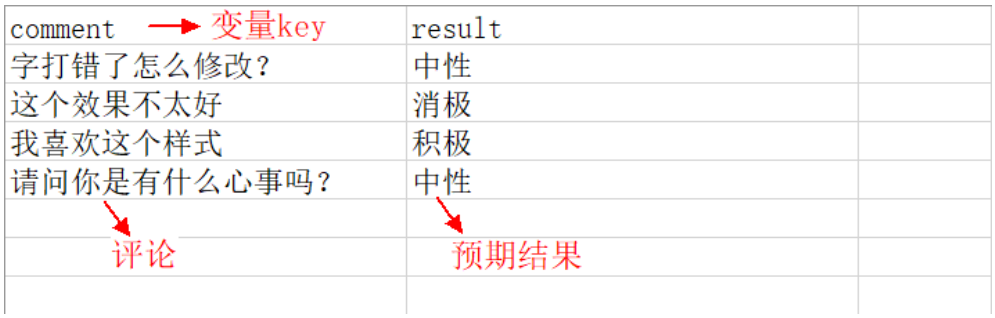


图 7-14 数据示例



创建提示词评估数据集

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示用例管理”，单击页面右上角“创建提示用例集”。

图 7-15 提示用例管理

3. 在“创建数据集”页面完成数据集的上传。

图 7-16 创建提示词评估数据集

7.5.2 创建提示词评估任务

选择候选提示词进行批量自动化评估，步骤如下：。

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。
4. 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，勾选需要进行横向比对的提示词，并单击“创建评估”。

图 7-17 创建评估

5. 选择评估使用的变量数据集和评估方法。
   - 评估用例集：根据选择的数据集，将待评估的提示词和数据集中的变量自动组装成完整的提示词，输入模型生成结果。
   - 评估方法：根据选择的评估方法，对模型生成结果和预期结果进行比较，并根据算法给出相应的得分。

图 7-18 创建提示词评估任务

6. 单击“确定”，评估任务自动进入执行状态。

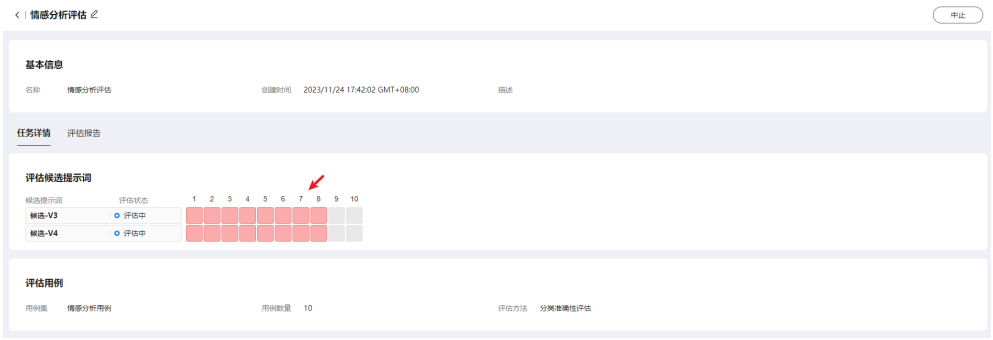
7.5.3 查看提示词评估结果

1. 评估任务创建完成后，会跳转至“评估”页面，在该页面可以查看评估状态。

图 7-19 查看提示词评任务状态

2. 单击“评估名称”，进入评估任务详情页，可以查看详细的评估进度，例如在图7-20中有10条评估用例，当前已评估8条，剩余2条待评估。

图 7-20 查看评估进展



3. 评估完成后，可以查看每条数据的评估结果。

   在评估结果中，“预期结果”表示变量值（问题）所预设的期望回答，“生成结果”表示模型回复的结果。通过比对“预期结果”、“生成结果”的差异可以判断提示词效果。

7.6 发布提示词

通过**横向比较提示词效果**和**批量评估提示词效果**，如果找到高质量的提示词，可以将这些提示词发布至“提示词模板”中。

1. 在提示词“候选”页面，选择质量好的提示词，并单击“保存到模板库”。

图 7-21 保存提示词至模板库

2. 进入“Agent 开发 > 提示词工程 > 提示词模板”页面，查看发布的提示词。

8 开发盘古大模型 Agent 应用

8.1 Agent 开发平台介绍

Agent 开发平台简介

Agent开发平台是基于NLP大模型，致力打造智能时代集开发、调测和运行为一体的AI应用平台。无论开发者是否拥有大模型应用的编程经验，都可以通过Agent平台快速创建各种类型的智能体。Agent开发平台旨在帮助开发者高效低成本的构建AI应用，加速领域和行业AI应用的落地。

- 针对“零码”开发者（无代码开发经验），平台提供了Prompt智能生成、插件自定义等能力，方便用户快速构建、调优、运行属于自己的大模型应用，仅需几步简单的配置即可创建属于自己的Agent应用。
- 对于“低码”开发者（有一定代码开发经验），可以通过工作流方式，适当编写一定代码，来构建逻辑复杂、且有较高稳定性要求的Agent应用，开发者也可以灵活组合各个节点，包含大模型节点、意图识别节点、提问器节点、插件节点等，通过“拖拉拽”的方式快速搭建一个工作流。

Agent 开发平台功能及优势

Agent开发平台具有能力扩展、自定义知识库、灵活的工作流设计和全链路信息调测评估等特点。

- 能力扩展：平台可以集成多种插件，插件能够有效扩展Agent的能力边界。
  - 预置插件：平台当前为用户提供了“Python解释器”插件，支持开发者直接将插件添加到Agent中，丰富Agent的能力。
  - 自定义插件：平台支持开发者创建自定义插件。支持开发者将工具、Function或者API通过配置方式快速创建为一个插件，并供Agent调用。
- 自定义知识库：平台提供了知识库功能来管理和存储数据，支持为AI应用提供自定义数据，并与之进行互动。多种格式的本地文档（支持docx、pptx、pdf等）都可以导入至知识库。
- 灵活的工作流设计：平台提供灵活的工作流设计，用于开发者处理逻辑复杂、且有较高稳定性要求的任务流。 支持“零码”和“低码”开发者通过“拖拉拽”的方式快速搭建一个工作流，创建一个应用。

Agent 开发平台应用场景

当前，基于Agent开发平台可以构建两种类型的应用，一种是针对文本生成、文本检索的知识型Agent，如搜索问答助手、代码生成助手等，执行主体在大模型；另一种是针对复杂工作流场景的流程型Agent，如金融分析助手、网络检测助手等。

- 知识型Agent：以大模型为任务执行核心，用户通过配置Prompt、知识库等信息，实现工具自主规划与调用，优点是可零码开发，对话过程更为智能，缺点是当大模型受到输入限制，难以执行链路较长且复杂的流程。
- 流程型Agent：以工作流为任务执行核心，用户通过在画布上对节点进行“拖拉拽”即可搭建出任务流程，场景的节点包括大模型节点、意图识别节点、提问器节点、插件节点、判断节点、代码节点、消息节点，优点是可扩展能力强，用户适当使用低码开发，缺点是对话交互智能度不高，复杂场景下分支多，难以维护。

8.2 编排与调用应用

8.2.1 应用介绍

在Agent开发平台上，用户可以构建两种类型的应用：

- **知识型Agent**：以大模型为任务执行核心，适用于文本生成和文本检索任务，如搜索问答助手、代码生成助手等。用户通过配置Prompt、知识库等信息，使得大模型能够自主规划和调用工具。
  - **优点**：零代码开发，对话过程智能化。
  - **缺点**：大模型在面对复杂的、长链条的流程时可能会受到输入长度限制，难以有效处理较为复杂的工作流。
- **流程型Agent**：以工作流为任务执行核心，用户可以通过在画布上“拖拽”节点来搭建任务流程。支持编排的节点类型包括：大模型节点、知识检索节点、意图识别节点、插件节点、判断节点、代码节点、消息节点、提问器节点。
  - **优点**：高度可扩展，支持低代码开发。
  - **缺点**：对话交互的智能度较低，复杂场景下流程分支较多，维护难度较大。

8.2.2 手动编排应用

Agent平台支持为应用配置插件、工作流技能，支持接入知识库，还可增加应用的对话体验，详见**创建与管理插件**、**编排工作流**、**创建与管理知识库**。

应用编排流程见**表8-1**。

表 8-1 应用编排流程

操作步骤	说明
步骤1：创建应用	创建一个新应用。
步骤2：配置提示词	在应用中配置大模型所需的Prompt。
步骤3：添加插件	为应用添加插件技能。
步骤4：添加工作流	为应用添加工作流技能。

操作步骤	说明
步骤5：添加知识库	为应用添加知识库。
步骤6：配置对话体验	为应用配置优化体验，提升用户体验。
步骤7：调试应用	调试应用的各个模块，确保其功能和表现符合预期。

说明

Agent应用支持的模型类型为**NLP大模型**。

步骤 1：创建应用

创建应用步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击“Agent开发”，进入Agent开发平台。
2. 单击左侧导航栏“工作台”，在“应用”页签，单击右上角“创建应用”。
3. 在“创建应用”页面，填写应用名称与应用描述，单击页面左下角的图片可修改应用图标，单击“确定”，进入应用编排页面。

步骤 2：配置提示词

创建应用后，需要撰写提示词（Prompt），为Agent设定人设、目标、核心技能、执行步骤。

应用会根据盘古NLP大模型对提示词的理解，来选择使用插件、工作流或知识库，响应用户问题。因此，一个好的提示词可以让模型更好地理解并执行任务，应用效果与提示词息息相关。

配置提示词步骤如下：

1. 在“提示词”模块，需要在输入框中填写Prompt提示词。
2. 可依据模板填写Prompt，单击“ ”，输入框中将自动填入角色指令模板。单击“ ”，可使用**获取提示词模板**中的提示词模板。

图 8-1 提示词

3. 提示词填写完成后可通过大模型进行优化，单击“ ”，可在 “Prompt优化”窗口中复制优化后的提示词，单击“确定”。

   注意，使用智能优化提示词功能前，请先在页面右上角选择需要使用的模型。

图 8-2 配置大模型

步骤 3：添加插件

应用支持添加插件技能，可添加“预置插件”和“个人插件”。添加插件可以为应用配备更多技能，建议插件数量不超过5个。

如果需要添加“个人插件”，请确保已完成**创建插件**操作。

添加插件的步骤如下：

1. 在“技能 > 插件”模块，单击“ ”。
2. 在“添加插件”窗口，选择预置插件或个人插件，单击“ ”进行添加，再单击“确定”。
3. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

图 8-3 配置应用插件

步骤 4：添加工作流

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

如果需要添加工作流，请确保已完成**编排工作流**操作。

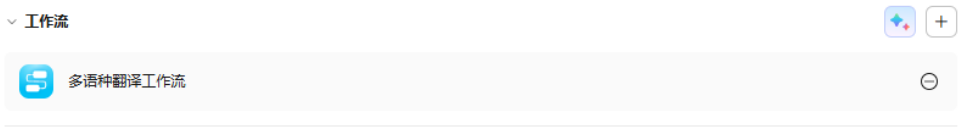
添加工作流的步骤如下：

1. 在“技能 > 工作流”模块，单击“ ”。
2. 在“添加工作流”窗口，单击“ ”进行添加，再单击“确定”。

图 8-4 添加工作流

3. 添加插件后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 8-5 已添加工作流



步骤 5：添加知识库

应用支持添加知识库。发送消息时，应用能够引用知识库中的内容回答用户问题，当前仅支持关联1个知识库。

如果需要添加知识库，请确保已完成**创建知识库**操作。

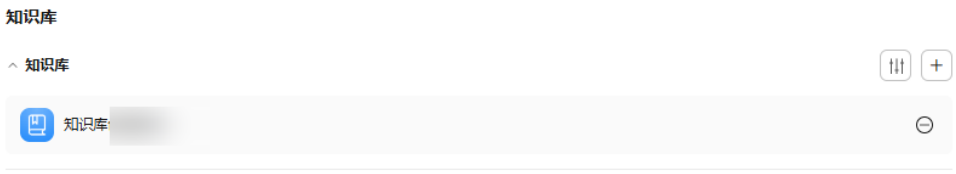
添加知识库的步骤如下：

1. 在“知识库”模块，单击“ ”。
2. 在“添加知识库”窗口，单击“ ”进行添加，再单击“确定”。

图 8-6 添加知识库

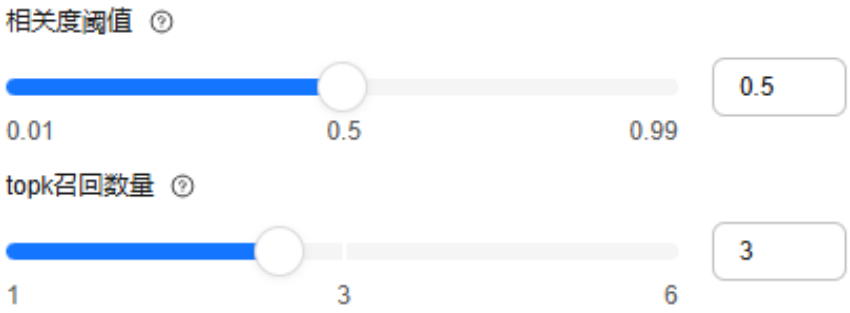
3. 添加知识库后，可在“知识库”中查看当前已添加的知识库。

图 8-7 已添加知识库



4. 可单击“ ”对知识库进行高级配置，包括相关度阈值与topk召回数量。
   - 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。

图 8-8 知识库高级配置



步骤 6：配置对话体验

应用支持配置对话体验功能，该功能可以提升用户与应用之间的互动质量和个性化体验，包括开场白、推荐问题与追问。

- 开场白：开场白是用户与应用进行首次交互时，应用主动向用户展示的一段内容。
- 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。
- 追问：在每轮回复后，默认根据对话内容提供提问建议。

配置对话体验的步骤如下：

1. 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”智能添加开场白。

   例如，“您好！我是您的智能助手，很高兴为您提供帮助。请告诉我，今天有什么我可以为您做的吗？”
2. 在“对话体验 > 推荐问题”中，可填写推荐问题，也可单击“智能添加 > 确定”智能添加推荐问题。推荐问题至多配置3条。

   例如，“请告诉我您需要什么帮助？如：帮我预定会议室、帮我查询天气预报。
3. 在“对话体验 > 追问”中，可选择是否开启“追问”功能，若开启，模型在每轮回复后，默认根据对话内容提供提问建议。
4. “对话体验”配置完成后，可在右侧“预览调试”中查看当前配置的开场白与推荐问题。

步骤 7：调试应用

创建应用后，平台支持对应用执行过程的进行预览与调试。

调试应用的步骤如下：

1. 在页面右上角单击“ ”，参考**表8-2**配置大模型参数。

表 8-2 大模型参数配置

参数	说明
模型选择	选择要使用的大模型，不同的模型效果存在差异。 该模型需提前部署，步骤请参见**创建NLP大模型部署任务**。
模式选择	用于配置大模型的输出多样性。 包含取值： ● 精确的：模型的输出内容严格遵循指令要求，可能会反复讨论某个主题，或频繁出现相同词汇。 ● 平衡的：平衡模型输出的随机性和准确性。 ● 创意性的：模型输出内容更具多样性和创新性，某些场景下可能会偏离主旨。 ● 自定义：自定义大模型输出的温度和核采样值，生成符合预期的输出。
温度	调高温度会使得模型的输出更多样性和创新性，反之，降低温度会使输出内容更加遵循指令要求但减少多样性，取值范围为0到1之间。 ● 调高温度，会使得模型的输出更多样性和创新性。 ● 降低温度，会使输出内容更加遵循指令要求但减少多样性。 在基于事实的问答场景，可以使用较低的回复随机性数值，以获得更真实和简洁的答案；在创造性的任务例如小说创作，可以适当调高回复随机性数值。建议不要与核采样同时调整。
核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值。核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性，取值范围为0.1到1之间。
携带上下文轮数	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。
输出模式	当前应用支持输出文本、Markdown两种模式的回答。

2. 在右侧“预览调试”的文本框中输入对话，应用将根据对话生成相应的回答。
3. 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。

图 8-9 查看调试结果

8.2.3 调用应用

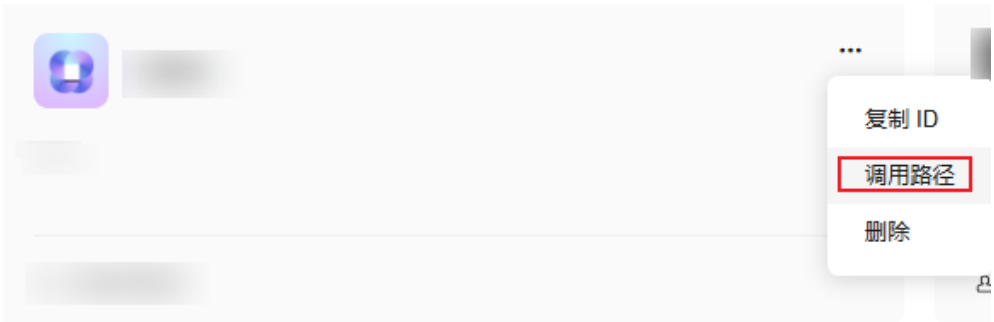
Agent开发平台支持将创建好的应用进行API调用。在调用应用前，请先参考**手动编排应用**，完成创建应用操作。

获取调用路径

应用的调用路径获取步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击“Agent开发”，进入Agent开发平台。
2. 在“工作台 > 应用”页面，单击所需应用的“ ··· > 调用路径”。

   图 8-10 获取应用调用路径-1

3. 在“调用路径”页面，单击“复制路径”即可获取调用路径。

   其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 8-11 获取应用调用路径-2

使用 Postman 调用 API

1. 获取Token。参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。
2. 在Postman中新建POST请求，并填入应用的调用路径，详见**获取调用路径**。
3. 填写请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**步骤1**中获取的Token值。
   - 参数名为stream，参数值为true。当前应用仅支持流式调用。
4. 在Postman中选择“Body > raw”选项，请求Body填写示例如下。其中，query参数为用户提出的问题，作为应用的输入。

```
{
  "query": "预定15:00到16:00的A12会议室"
}
```

5. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示应用API调用成功，响应示例如下：

```
data:{"event":"start","data":{},"createdTime":1733821291867,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"好的
"},"createdTime":1733821304670,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":",
"},"createdTime":1733821304671,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"我
"},"createdTime":1733821304671,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"需要先
"},"createdTime":1733821304671,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"查询
"},"createdTime":1733821304672,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"一下
"},"createdTime":1733821304672,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"A"},"createdTime":1733821304672,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"12"},"createdTime":1733821304673,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}
```


data:{"event":"message","data":{"answer":"会议室
"},"createdTime":1733821304673,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"在
"},"createdTime":1733821304673,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"15"},"createdTime":1733821304673,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821304674,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":{"answer":"到
"},"createdTime":1733821304674,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"16"},"createdTime":1733821304674,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821304675,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":{"answer":"的状态
"},"createdTime":1733821304675,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"。
"},"createdTime":1733821304675,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"请
"},"createdTime":1733821304675,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"稍
"},"createdTime":1733821304676,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"等
"},"createdTime":1733821304676,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"。
"},"createdTime":1733821304676,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
"},"createdTime":1733821304676,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
meeting"},"createdTime":1733821304676,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"_"},"createdTime":1733821304677,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"room"},"createdTime":1733821304677,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"_status"},"createdTime":1733821304677,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"_query"},"createdTime":1733821304677,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"|"},"createdTime":1733821304678,"conversationId":"7795ee1b-b145-4e21-8e02-
d14f973b6410"}

data:{"event":"message","data":
{"answer":"{\""},"createdTime":1733821304678,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"meeting"},"createdTime":1733821304678,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"Room"},"createdTime":1733821304678,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\":\""},"createdTime":1733821304678,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"A"},"createdTime":1733821304679,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"12"},"createdTime":1733821304679,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\",\""},"createdTime":1733821304679,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"start"},"createdTime":1733821304679,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\":\""},"createdTime":1733821304679,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"15"},"createdTime":1733821304680,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821304680,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\",\""},"createdTime":1733821304680,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"end"},"createdTime":1733821304680,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\":\""},"createdTime":1733821304681,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"16"},"createdTime":1733821304681,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821304681,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\"}"},"createdTime":1733821304681,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
"},"createdTime":1733821304682,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"function_call_end","data":{},"createdTime":1733821304684,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"function_call","conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410","createdTime":1733821304685,"data":{"answer":{"function_call":{"name":"meeting_room_status_query","arguments":"{\"meetingRoom\": \"A12\", \"start\": \"15:00\", \"end\": \"16:00\"}"},"is_workflow":false,"time_consumption":{"total_latency":12.81,"overall_latency":12.82},"memory_variables":{}}}}

data:{"event":"api_exec_data","data":{"answer":{"role":"function","name":"meeting_room_status_query","content":{"result":"空闲
"},"time_consumption":{"plugin_latency":0.0,"overall_latency":13.05}}},"createdTime":1733821304918,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"start","data":{},"createdTime":1733821304918,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
"},"createdTime":1733821307034,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
reserve"},"createdTime":1733821307035,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"_"},"createdTime":1733821307035,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"meeting"},"createdTime":1733821307035,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"_"},"createdTime":1733821307036,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"room"},"createdTime":1733821307036,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"|"},"createdTime":1733821307036,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"{\""},"createdTime":1733821307037,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"meeting"},"createdTime":1733821307037,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"Room"},"createdTime":1733821307037,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"\":"},"createdTime":1733821307037,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
\""},"createdTime":1733821307037,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"A"},"createdTime":1733821307038,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"12"},"createdTime":1733821307038,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\","},"createdTime":1733821307038,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
\""},"createdTime":1733821307038,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"start"},"createdTime":1733821307039,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\":"},"createdTime":1733821307039,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
\""},"createdTime":1733821307039,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"15"},"createdTime":1733821307039,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821307039,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\","},"createdTime":1733821307040,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
\""},"createdTime":1733821307040,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"end"},"createdTime":1733821307040,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\":"},"createdTime":1733821307040,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
\""},"createdTime":1733821307040,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"16"},"createdTime":1733821307041,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":":00"},"createdTime":1733821307041,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":
{"answer":"\"}"},"createdTime":1733821307041,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"
"},"createdTime":1733821307041,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"function_call_end","data":{},"createdTime":1733821307044,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"function_call","conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410","createdTime":1733821307044,"data":{"answer":{"function_call":

{"name":"reserve_meeting_room","arguments":"{\"meetingRoom\": \"A12\", \"start\": \"15:00\", \"end\": \"16:00\"}"},"is_workflow":false,"time_consumption":{"total_latency":2.12,"overall_latency":15.18},"memory_variables":{}}}}

data:{"event":"api_exec_data","data":{"answer":{"role":"function","name":"reserve_meeting_room","content":{"result":"会议室预定成功"},"time_consumption":{"plugin_latency":0.0,"overall_latency":15.25}}},"createdTime":1733821307119,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"start","data":{},"createdTime":1733821307120,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"已"},"createdTime":1733821308735,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"成功"},"createdTime":1733821308735,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"为您"},"createdTime":1733821308736,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"预定"},"createdTime":1733821308736,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"15"},"createdTime":1733821308736,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":":00"},"createdTime":1733821308736,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"到"},"createdTime":1733821308737,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"16"},"createdTime":1733821308737,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":":00"},"createdTime":1733821308737,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"的"},"createdTime":1733821308737,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"A"},"createdTime":1733821308737,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"12"},"createdTime":1733821308738,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"会议室"},"createdTime":1733821308738,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"message","data":{"answer":"。"},"createdTime":1733821308738,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"done","data":{},"createdTime":1733821308741,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"statistic_data","data":{"answer":{"total_latency":1.62,"model_latency":null,"wait_latency":null,"overall_latency":16.87}},"createdTime":1733821308741,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

data:{"event":"summary_response","data":{"answer":{"role":"assistant","content":"已成功为您预定15:00

到16:00的A12会议室。"}},"createdTime":1733821308741,"conversationId":"7795ee1b-b145-4e21-8e02-d14f973b6410"}

8.2.4 管理应用

Agent开发平台支持对应用执行获取应用ID、删除、导入、导出操作。

获取应用 ID、删除应用

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 应用”页面。
4. 单击“ ··· > 复制ID”，可获取当前应用ID。
5. 单击“ ··· > 删除”，可删除当前应用。

   说明

   删除应用属于高危操作，删除前，请确保该应用不再使用。

导出、导入应用

平台支持导出和导入应用。导出应用时，将同步导出应用关联的插件和工作流等配置。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 应用”页面。
4. 导出应用。
   a. 单击页面右上角“导出”。
   b. 在“导出应用”页面选择应用，单击“导出”。应用将以一个jsonl格式的文件下载至本地。
5. 导入应用。
   a. 单击页面右上角“导入”。
   b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。
   c. 选择导入文件后，选择解析内容。

      平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。
   d. 单击“导入”，导入成功的应用将在“工作台 > 应用”中展示。

8.3 编排与调用工作流

8.3.1 工作流介绍

Agent开发平台的工作流由多个节点构成，节点是组成工作流的基本单元。平台支持多种节点，包括开始、结束、大模型、意图识别、提问器、插件、判断、代码、知识检索和消息节点。

创建工作流时，每个节点需要配置不同的参数，如输入和输出参数等，开发者可通过拖、拉、拽可视化编排更多的节点，实现复杂业务流程的编排，从而快速构建应用。

工作流方式主要面向目标任务包含多个复杂步骤、对输出结果成功率和准确率有严格要求的复杂业务场景。

在编排工作流时，可以使用以下节点进行功能设计：

- 开始节点：开始节点是工作流的起始节点，用户输入的信息由开始节点传入。
- 结束节点：结束节点是工作流的最终节点，用于定义整个工作流的输出信息。
- 大模型节点：用于在工作流中引入大模型能力。
- 意图识别节点：用于根据用户的输入进行意图分类并导向后续不同的处理流程。
- 提问器节点：提供了在对话过程中向用户收集更多信息的能力。
- 插件节点：用于引入API插件，根据节点的输入，执行用户定义的插件，将插件执行结果作为节点的输出。
- 判断节点：编排应用时作为分支切换节点，可以根据输入满足的判断条件，指定执行对应的工作流分支。
- 代码节点：用于引入代码执行器，根据节点的输入，执行指定Python代码，节点的输出是代码执行的结果信息。
- 知识检索节点：可以根据输入参数从指定知识库内召回匹配的信息。
- 消息节点：定义一段文本内容，在工作流的执行过程中向用户发送该内容的消息。

8.3.2 编排工作流

Agent平台支持对工作流编排多个节点，以实现复杂业务流程的编排。

工作流包含两种类型：

- 对话型工作流。面向多轮交互的开放式问答场景，基于用户对话内容提取关键信息，输出最终结果。适用于客服助手、工单助手、娱乐互动等场景。
- 任务型工作流。面向自动化处理场景，基于输入内容直接输出结果，无中间的对话交互过程。适用于内容生成、批量翻译、数据分析等场景。

说明

任务型工作流不支持配置**消息节点**和**提问器节点**。

工作流编排流程见**表8-3**。

表 8-3 工作流编排流程

操作步骤	说明
创建工作流（必选）	创建一个新的工作流。

操作步骤	说明
开始节点配置说明（必选）	设定工作流的起始点。
大模型节点配置说明	将大模型节点加入工作流，用于处理复杂的自然语言理解或生成任务。
意图识别节点配置说明	配置该节点来分析用户输入，识别其意图，以便后续处理。
提问器节点配置说明	配置一个提问器节点，用于向用户或系统提出问题，获取所需信息。
插件节点配置说明	将外部API等集成到工作流中，以扩展功能或调用外部接口。
判断节点配置说明	设置条件判断逻辑，根据不同情况分支到不同的流程路径。
代码节点配置说明	配置自定义代码逻辑，用于处理特定的业务需求或复杂运算。
消息节点配置说明	向用户展示中间过程的消息输出能力。
知识检索节点配置说明	配置知识检索节点。
结束节点配置说明（必选）	设定工作流的结束点，标志流程的完成或终止。
试运行工作流（必选）	进行工作流的调试，确保各节点正常运行。

创建工作流（必选）

大模型工作流应用可以将NLP大模型编排至工作流中，编排完成后可以使用大模型回答用户问题。

创建工作流的步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击“Agent开发”，进入Agent开发平台。
2. 单击左侧导航栏“工作台”，在“工作流”页签，单击右上角“创建工作流”。
3. 在“创建工作流”页面，选择工作流类型。填写工作流名称、英文名称与工作流描述。
4. 单击页面左下角的图片可修改工作流图标，单击“确定”，进入工作流编排页面。

开始节点配置说明（必选）

开始节点是工作流的起始节点，包含用户输入信息，用于触发一个工作流，是每个工作流的入口节点。开始节点不支持新增或者删除。

开始节点为**必选**节点，需要配置于所有场景。

开始节点配置步骤如下：

1. 拖动左侧任意节点至画布中，以显示开始节点。
2. 单击画布中的开始节点以打开节点配置页面。
3. 开始节点的参数默认已配置，不支持修改开始节点的参数。

图 8-12 开始节点配置图

大模型节点配置说明

大模型节点提供了使用大模型的能力，可在节点中配置已部署的模型，用户可以通过编写Prompt、设置参数让模型处理相应任务。

大模型节点为**可选**节点，若无需配置，可跳过该步骤。

大模型节点配置步骤如下：

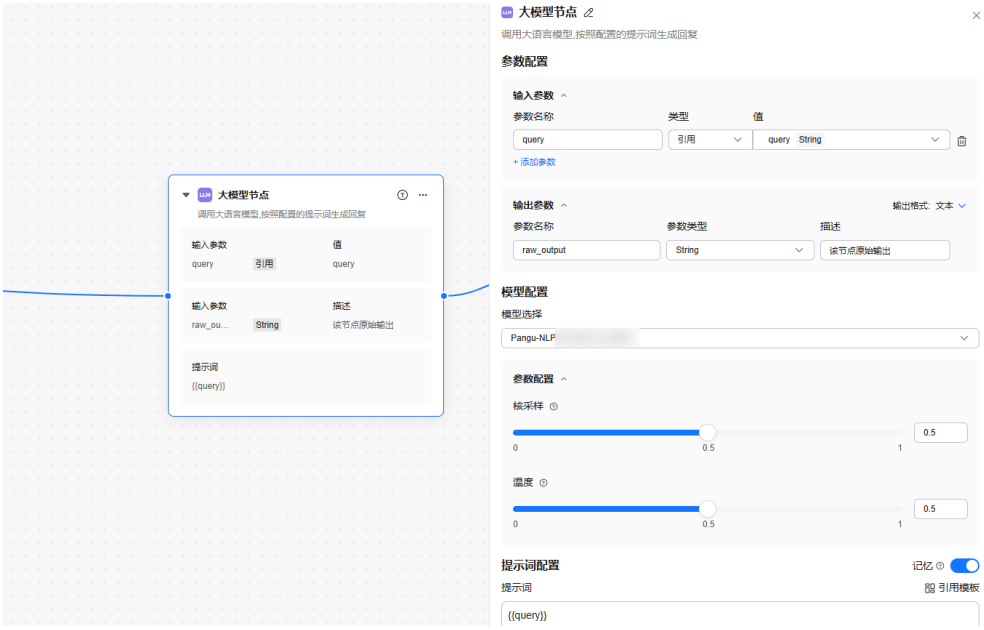
1. 拖动左侧“大模型”节点至画布中，单击该节点以打开节点配置页面。
2. 参照表8-4，完成大模型节点的配置。

表 8-4 大模型节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。

配置类型	参数名称	参数说明
	输出参数	该参数用于解析大模型节点的输出，并提供给后序节点的输出参数引用。 ● 参数名称：参数的名称长度必须大于等于1个字符，并且字符只允许为下面三种类型： – 字母（A-Z或a-z） – 数字（0-9） – 特殊字符：_ **说明** 用户自定义输出参数名称不允许与内置输出参数rawOutput同名。大模型节点有一个内置输出参数rawOutput，代表该节点未经解析的原始输出，与大模型节点相连的后序节点可以直接引用该输出。 ● 参数类型：输出参数的类型，可选String、Integer、Number、Boolean。 ● 描述：对于该输出参数的描述。 ● 输出格式：支持输出的格式包括文本、Markdown、JSON。
模型配置	模型选择	选择已部署的模型。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
提示词配置	提示词	配置提示词，并选择是否打开记忆功能。 配置提示词时，支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。 ● 提示词：大模型的系统提示词，用于指导模型更好的完成任务。 ● 记忆：聊天记忆，打开后可记录多轮对话的内容。默认关闭。

图 8-13 大模型节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接大模型节点和其他节点。

意图识别节点配置说明

意图识别节点通过大模型推理分析用户输入，匹配预定义的意图关键字类别，并根据识别结果引导至相应的处理流程，通常位于工作流的前置位置。

意图识别节点为**可选**节点，若无需配置，可跳过该步骤。

意图识别节点配置步骤如下：

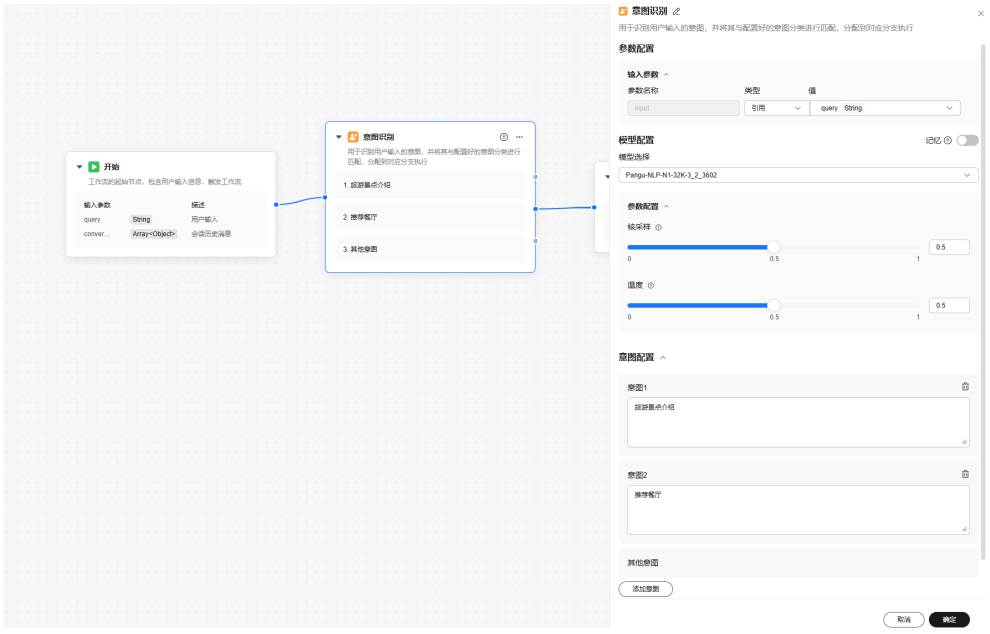
1. 拖动左侧“意图识别”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表8-5**，完成意图识别节点的配置。

表 8-5 意图识别节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：默认名称input，为固定值，不可编辑。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
模型配置	模型选择	选择已部署的模型。

配置类型	参数名称	参数说明
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
意图配置	意图1	配置相关意图关键字信息。 在意图输入框中输入意图描述信息，描述信息为针对该类别的描述语句或者关键词，也将作为大模型进行推理和分类的依据。意图数量最多为21个（包含默认的“其他”意图）。
高级配置	提示词	高级配置项供进阶开发者修改提示词，如果不配置将会使用系统默认值。提示词的撰写可能影响到意图识别节点的准确性。

图 8-14 意图识别节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接意图识别节点和其他节点。

提问器节点配置说明

提问器节点为开发者提供了收集用户问题所需信息的功能。该节点会循环执行，直到收集到所有必需的信息为止。

提问器节点为**可选**节点，若无需配置，可跳过该步骤。

提问器节点配置步骤如下：

1. 拖动左侧“提问器”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表8-6**，完成提问器节点的配置。

表 8-6 提问器节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 示例：输入参数为“pre_assigned_meeting_rooms”，希望用户在指定的多个选项中选出一个，后续问题配置为“有以下几个会议室供您选择：{{pre_assigned_meeting_rooms}}，请选择您想预订的会议室”。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	输出参数	该参数用于解析大模型节点的输出，并提供给后序节点的输出参数引用。 • 参数名称：只允许输入字母、数字、下划线、短横线。 • 中文名称：不允许为空。 • 类型：输出参数的类型，可选String、Integer、Number、Boolean。 • 默认值：输出参数的默认值，若开启“参数提取 > 是否提取”功能，则默认值不允许为空。 • 描述：对于该输出参数的描述。 • 参数提取：开启后，可增加输出参数的配置，并对参数中文名进行额外配置。关闭参数提取，输出为用户最近一轮（即回答当前提问器）的对话输入。 – 中文名称：若开启“参数提取 > 是否提取”功能，可额外配置中文名称。 – 参数校验：可自定义参数校验规则对输出参数规范性进行校验。规则包括参数名称、校验类型及校验规则。 – 是否提取：开启后该参数必须提取到或使用默认值，关闭则该参数允许为空或者使用默认值。 – 反思：在参数提取之后，会根据参数描述与用户指令，对打开反思开关的参数，独立调用大模型进行反思并修正当前提取的结果。 • 引用插件：支持导入已有插件的参数信息。

配置类型	参数名称	参数说明
模型配置	模型选择	选择已部署的模型。
	核采样	模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	温度	用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
问题配置	问题	该参数将在对话框中原样呈现给用户。如未配置此处，将由大模型根据输出参数描述，自动生成包含所有问题关键词的一个问题。
	最大回复轮数	该参数指在与用户交互过程中，模型能够持续进行对话而不丧失上下文或性能的最大回合数。
输入配置	对话历史	开启对话历史后，当前工作流的上下文会带入提问器。
高级配置	指令	提供大模型额外的提示信息，用于更准确的提取参数，例如指定被提取参数的格式要求。
深度定制	示例配置	用户配置的示例内容，配置后会在大模型的请求中添加“#示例 {{用户配置的内容}}”。
输入改写	日期时间改写	开启后，用户问题涉及日历、日期及时间相关内容时，系统将进行运算，补充具体时间点，以便大模型更准确的理解。

3. 节点配置完成后，单击“确定”。
4. 连接大模型节点和其他节点。

插件节点配置说明

插件节点使开发者可以在工作流中实现与外部环境的交互，以拥有更强大的能力，完成更复杂的任务。

插件类型包括预置插件和个人插件。

- 预置插件：平台预置了代码解释器插件，支持开发者直接将插件添加到工作流或应用中，丰富其能力。
- 自定义插件：平台允许开发者创建自定义插件，支持将API通过配置方式快速创建为插件，并供Agent调用。

插件节点为**可选**节点，若无需配置，可跳过该步骤。

插件节点配置步骤如下：

1. 拖动左侧“插件”节点至画布中，在“个人插件”或“预置插件”页签单击“  ”，将插件添加至画布中。

   预置插件为平台内置的插件

   个人插件为用户自定义的插件，创建插件步骤详见**创建插件**。
2. 单击画布中已添加的“插件”节点，参照**表8-7**，完成插件节点的配置。

表 8-7 插件节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：从插件元信息中导入，用户无需手动添加。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	输出参数	输出参数所有信息从插件元信息中导入，用户无需手动添加。

图 8-15 插件节点配置示例

3. 节点配置完成后，单击“确定”。
4. 连接插件节点和其他节点。

判断节点配置说明

判断节点是一个IF-ELSE节点，提供了多分支条件判断的能力，用于设计分支流程。

当向该节点输入参数时，节点会判断输入是否符合“参数配置”中预设的条件，符合则执行“IF”对应的工作流分支，否则执行“ELSE”对应的工作流分支。

每个分支条件支持添加多个判断条件（且、或），同时支持添加多个条件分支。

判断节点为**可选**节点，若无需配置，可跳过该步骤。

判断节点配置步骤如下：

1. 拖动左侧“判断”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表8-8**，完成判断节点的配置。

表 8-8 判断节点配置说明

配置类型	参数名称	参数说明
参数配置	IF	IF分支由**[参数名称 比较条件 比较对象 值]**组成一条件表达式。 ● 参数名称：条件表达式左边部分，需要选择来自前序节点的输出参数。 ● 比较条件：条件表达式中间部分，当前支持的比较条件有：长度大于、长度大于等于、长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。其中，针对不同的参数名称，将展示不同的比较条件，具体可以前端页面为准。 ● 比较对象、值：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值。 – 输入：支持用户自定义取值。 ● 添加条件：单击“添加条件”，在当前分支添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 – 单击“且”或“或”，可以切换该分支表达式的运算逻辑。
	ELSE	该参数将不满足其他条件分支的内容输出，并提供给后序节点的输出参数引用。
	添加分支	可以添加新的分支ELSE IF，新分支的配置方式与IF分支相同。

图 8-16 判断节点配置示例

3. 节点配置完成后，单击“确定”。
4. 连接判断节点和其他节点。

代码节点配置说明

代码节点支持编写Python代码来处理文本、复杂逻辑判断等。

代码节点可以增强开发人员的灵活性，使其能够在工作流程中以嵌入自定义Python脚本的方式操作变量。通过配置选项，开发者可以指定所需的输入和输出变量，并编写相应的执行代码。

代码节点为**可选**节点，若无需配置，可跳过该步骤。

代码节点配置步骤如下：

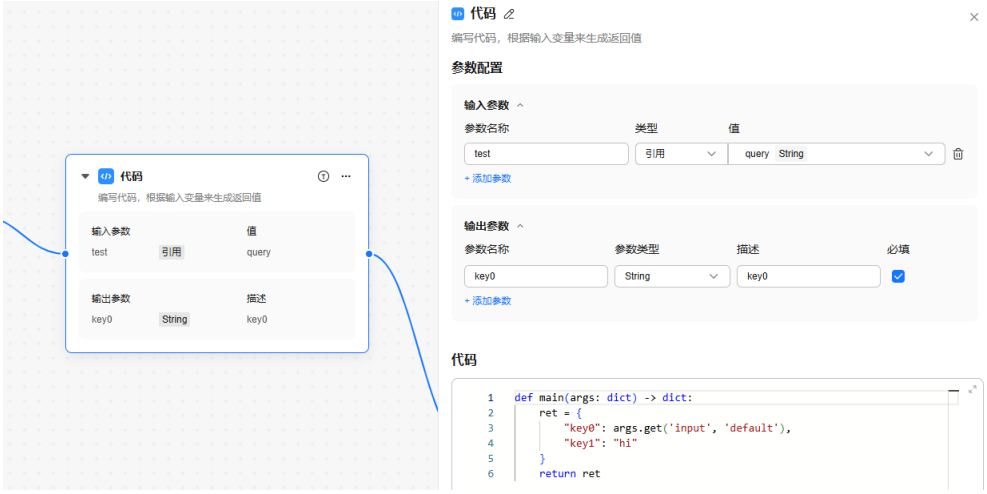
1. 拖动左侧“代码”节点至画布中，单击该节点以打开节点配置页面。
2. 参照表8-9，完成代码节点的配置。

表 8-9 代码节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值。 – 输入：支持用户自定义取值。
	输出参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 参数类型：输出参数的类型，可选String、Integer、Number、Boolean。 • 描述：对于该输出参数的描述。 • 必填：选择当前输出参数是否必填。

<table>
<tr><th>配置类型</th><th>参数名称</th><th>参数说明</th></tr>
<tr><td>代码</td><td>-</td><td>编写Python代码，代码配置示例如下：
文本拼接示例代码。<pre>def main(args: dict) -> dict:
# 注意在输入参数中定义名为input1的变量
input1 = args.get('input1')
# 注意在输入参数中定义名为input2的变量
input2 = args.get('input2')
res = {
# 注意在输出参数中定义名为res的变量
"res": input1 + input2,
}
return res</pre>
复杂逻辑判断示例代码。<pre>def main(args: dict) -> dict:
import re
# 注意在输入参数中定义input1参数
input1 = args.get('input1')
# 判断是否满足要求：非空、以字母开头、只包含数字字母下划线
if input1 and bool(re.match(r'^[A-Za-z][A-Za-z0-9_]*$', input1)):
return {
# 注意在输出参数中定义res
'res': "输入字符串满足要求"
}
else:
return {
# 注意在输出参数中定义res
'res': "输入字符串不满足要求"
}</pre>
数学计算示例代码。<pre>def main(args: dict) -> dict:
# 注意在输入参数中定义名为input1的变量
input1 = args.get('input1')
try:
input1 = int(input1)
return {
# 注意输出参数中定义res变量
'res': input1 * input1
}
except Exception as e:
return {
# 注意输出参数中定义res变量
'res': "输入类型错误或者数字大小超出限制"
}</pre>
说明
编写代码时不要更改第一行函数定义以及输入输出定义。</td></tr>
</table>

图 8-17 代码节点配置示例



3. 节点配置完成后，单击“确定”。
4. 连接代码节点和其他节点。

消息节点配置说明

消息节点可提供中间过程的消息输出能力，通过定义一段文本内容，在工作流的执行过程中向用户发送该内容的消息。

消息节点配置步骤如下：

1. 拖动左侧“消息”节点至画布中，单击该节点以打开节点配置页面。
2. 参照表8-10，完成大模型节点的配置。

表 8-10 消息节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
指定回复	-	可撰写指定的回复信息，并以{{参数名称}}的形式插入变量。

3. 节点配置完成后，单击“确定”。
4. 连接消息节点和其他节点。

知识检索节点配置说明

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，节点支持选择用户创建的知识库，创建步骤请详见**创建知识库**。

知识检索节点配置步骤如下：

1. 拖动左侧“知识检索”节点至画布中，单击该节点以打开节点配置页面。
2. 参照**表8-11**，完成大模型节点的配置。

表 8-11 知识检索节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：输入参数固定只有1个，参数名称为query且不可修改。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。
	知识库	支持上传用户所建立的知识库。
	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。 取值范围为0.01~0.99。
	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。 取值范围为1~6。
输出参数	-	知识检索节点输出的参数output_list为一个数组，包含文档名称（document_name）、检索到的内容（content）以及得分（score）。

图 8-18 知识检索节点配置示例

3. 节点配置完成后，单击“确定”。
4. 连接消息节点和其他节点。

结束节点配置说明（必选）

结束节点是工作流的最终节点。当工作流执行完成后，需要结束节点用于输出工作流的执行结果。结束节点不支持新增或者删除，该节点后不支持添加其他节点。

结束节点可能会有多个输入，但是只能有一个输出值，因此需要开发者在“指定回复”中合并多个输入值为一个输出值。

结束节点为**必选**节点，需要配置于所有场景中。

结束节点配置步骤如下：

1. 单击画布中的结束节点以打开节点配置页面。
2. 参照**表8-12**，完成结束节点的配置。

表 8-12 结束节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值。 – 输入：支持用户自定义取值。

配置类型	参数名称	参数说明
指定回复	-	可撰写指定的回复信息，并以{{参数名称}}的形式插入变量。 • 支持用户将多个输入变量合并成一个字符串输出，使用{{参数名称}}代指上述定义的输入参数。 例如，已定义输入参数**end_input值为hello**，定义“指定回复”内容为**{{end_input}} world**，则最终的输出即为**hello world**。

图 8-19 结束节点配置示例

3. 节点配置完成后，单击“确定”。
4. 连接其他节点和结束节点。

试运行工作流（必选）

Agent开发平台支持对整个工作流进行试运行，也支持对工作流的单个节点进行调试。

- 试运行工作流：
  a. 工作流编排完成后，单击右上角“试运行”，在对话框中输入问题，等待返回试运行结果。
  b. 在试运行过程中，可以单击右上角“ 调试 ”查看调试结果，包括运行结果与调用详情。
  c. 如果试运行失败，常见报错与解决方案请详见**Agent开发常见报错与解决方案**。

图 8-20 调试结果示例

- 单节点调试，以调试“意图识别”节点为例：
  a. 在工作流编排页面，单击意图识别节点的“Ⓣ”，进入单节点调试页面。
  b. 编写输入参数内容，单击“开始运行”。

  图 8-21 编写输入参数内容

  c. 可在“运行结果”页面查看当前节点的运行结果。
  d. 若运行成功，节点处也将显示“运行成功”字样。

图 8-22 单节点调试运行成功示例

8.3.3 调用工作流

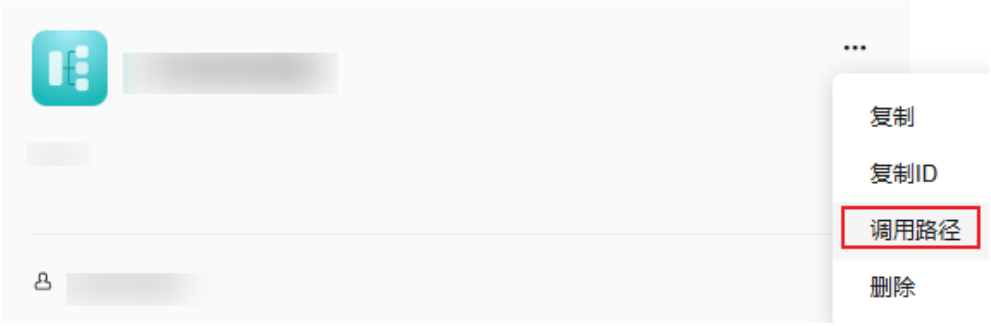
工作流试运行成功后，可以使用API调用该工作流。

获取调用路径

工作流的调用路径获取步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 在“工作台 > 工作流”页面，单击所需工作流的“ ··· > 调用路径”。

   图 8-23 获取工作流调用路径-1

4. 在“调用路径”页面，单击“复制路径”即可获取调用路径。

   其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 8-24 获取工作流调用路径-2

使用 Postman 调用 API

1. 获取Token。参考《API参考》文档“如何调用REST API > 认证鉴权”章节获取Token。
2. 在Postman中新建POST请求，并填入工作流的调用路径，详见**获取调用路径**。
3. 填写请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**步骤1**中获取的Token值。
   - 参数名为stream，参数值为true。当前工作流仅支持流式调用。
4. 在Postman中选择“Body > raw”选项，请求Body填写示例如下。

   其中，inputs参数为用户提出的问题，作为工作流的输入。plugin_id参数为插件ID，获取方式详见**管理插件**。

```
{
  "inputs": {
    "query": "你好"
  },
  "plugin_configs": [
    {
      "plugin_id": "xxxxxxxxx",
      "config": {
        "key": "value"
      }
    }
  ]
}
```

5. 单击Postman界面“Send”，发送请求。当接口返回状态为200时，表示应用API调用成功，响应示例如下：

   提问器节点返回示例：

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "messages": [
        {
            "role": "assistant",
            "content": "请您提供年龄相关的信息"
        }
    ],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734336269313,
    "end_time": 1734336270908
}
```


结束节点返回示例：

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "outputs": {
        "responseContent": "18"
    },
    "messages": [],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734337068533,
    "end_time": 1734337082545
}
```

8.3.4 管理工作流

Agent开发平台支持对工作流执行复制、获取工作流ID、删除、导入、导出操作。

获取工作流 ID、删除工作流

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 工作流”页面。
4. 单击“ ··· > 复制”，可复制当前工作流。
5. 单击“ ··· > 复制ID”，可获取当前工作流ID。
6. 单击“ ··· > 删除”，可删除当前工作流。

说明

删除应用属于高危操作，删除前，请确保该工作流不再使用。

导出、导入工作流

平台支持导出和导入工作流。导出工作流时，将同步导出工作流关联的插件等配置。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 工作流”页面。
4. 导出工作流。
   a. 单击页面右上角“导出”。
   b. 在“导出工作流”页面选择工作流，单击“导出”。工作流将以一个jsonl格式的文件下载至本地。
5. 导入工作流。
   a. 单击页面右上角“导入”。
   b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。
   c. 选择导入文件后，选择解析内容。
      平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。

d. 单击“导入”，导入成功的工作流将在“工作台 > 工作流”页面中展示。

8.4 创建与管理插件

8.4.1 插件介绍

在Agent开发平台中，插件是大模型能力的重要扩展。通过模块化方式，插件能够为大模型提供更多专业技能和复杂任务处理能力，使其在多样化的实际场景中更加高效地满足用户需求。

通过插件接入，用户可以为应用赋予大模型本身不具备的能力。插件提供丰富的外部服务接口，当任务执行时，模型会根据提示词感知适用的插件，并自动调用它们，从外部服务中获取结果并返回。这样的设计使得Agent能够智能处理复杂任务，甚至跨领域解决问题，实现对复杂问题的自动化处理。

Agent开发平台支持两种类型的插件：

- 预置插件：平台为开发者和用户提供了预置插件，直接可用，无需额外开发。例如，平台提供的“Python解释器插件”能够根据用户输入的问题自动生成Python代码，并执行该代码获取结果。此插件为Agent提供了强大的计算、数据处理和分析功能，用户只需将其添加到应用中，即可扩展功能。
- 自定义插件：为了满足更个性化的需求，平台允许开发者创建自定义插件，支持将API通过配置方式快速创建为插件，并供Agent调用。这样，开发者可以根据特定需求为应用增加专属功能。

8.4.2 创建插件

创建插件的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 单击左侧导航栏“工作台”，在“插件”页签，单击右上角“创建插件”。
4. 在“创建插件”页面，填写插件名称与插件描述，单击图片可上传插件图标，单击“下一步”。
5. 在“配置信息”页面，参照**表8-13**完成信息配置。

表 8-13 插件信息配置说明

参数名称	参数说明
插件URL	插件服务的请求URL地址。 ● URL协议只支持HTTP和HTTPS。 ● 系统会校验URL地址是否为标准的URL格式。 ● URL对应的IP默认不应为内网，否则会导致注册失败。仅在非商用环境部署时，才允许支持内网URL，且需要通过相关的服务的启动配置项关闭内网屏蔽。

参数名称	参数说明
请求方法	插件服务的请求方式，POST或GET。
权限校验	选择调用API时是否需要通过鉴权才可以调用。 • 无需鉴权：API可以公开访问，不需要任何形式的身份验证或授权。 • 用户级鉴权：需要用户提供身份验证信息来访问API。 需填写密钥位置，即密钥是从Header中读取还是Query中读取。并设置密钥鉴权参数名、密钥来源参数名，以确保系统能够正确地提取和使用鉴权信息。 • API Key：在调用API时提供一个唯一的API Key进行鉴权。 需填写密钥位置，即密钥是从Header中读取还是Query中读取。并设置API Key的密钥鉴权参数名和密钥值。
请求头	填写API的请求头信息，例如： • Key：Content-Type • Value：application/json

图 8-25 API 请求信息配置示例

说明

自定义插件使用HTTP服务，或不增加鉴权方式可能存在安全风险。

6. 单击“下一步”，在“参数信息”页面，参照表8-14完成参数配置。

表 8-14 插件参数配置说明

参数类型	参数名称	参数说明
请求参数	参数封装	开启后，会将请求参数封装为一个列表（数组）结构，可适配入参为数组格式的插件接口。 示例：原参数列表：{"a":"string", "b":1}，开启封装后的参数列表：[{"a":"string", "b":1}]
	参数名称	参数的名称，参数名称会作为大模型解析参数含义的依据。
	中文名称	该参数的中文名称。
	参数类型	该参数值的数据类型，String、Integer、Number等多种类型支持选择。
	位置	当前参数在请求信息中的位置，可选Body、Headers或Query。
	默认值	参数的默认值。
	描述	参数的描述，尽可能准确的描述参数的含义和要求，可提升Agent提取参数的准确率。
	参数校验	可设置当前参数的校验规则。
	必填	指定该参数是否为必填项。
响应参数	参数封装	开启后，会将请求参数封装为一个列表（数组）结构，可适配入参为数组格式的插件接口。 示例：原参数列表：{"a":"string", "b":1}，开启封装后的参数列表：[{"a":"string", "b":1}]
	参数名称	响应参数的名称，参数名称会作为大模型解析大模型输出结果的依据。
	参数描述	响应参数的名称，参数描述会作为大模型解析大模型输出结果的依据。
	参数类型	该参数值的数据类型，String、Integer、Number等多种类型支持选择。
	是否提取	开启后则该参数必须提取到，关闭则该参数允许为空或者使用默认值。

图 8-26 填写 API 请求、响应参数

8.4.3 管理插件

Agent开发平台支持对插件执行获取插件ID、删除、导入、导出操作。

获取插件 ID、删除插件

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 插件”页面。
4. 单击“ *** > 复制ID”，可获取当前插件ID。
5. 单击“ *** > 删除”，可删除当前插件。

说明

删除应用属于高危操作，删除前，请确保该插件不再使用。

导出、导入插件

平台支持导出和导入插件。

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 插件”页面。
4. 导出插件。

a. 单击页面右上角“导出”。
b. 在“导出插件”页面选择工作流，单击“导出”。插件将以一个jsonl格式的文件下载至本地。

5. 导入插件。
a. 单击页面右上角“导入”。
b. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。
c. 选择导入文件后，选择解析内容。
平台将自动解析jsonl文件。如果解析的文件在平台中已存在，勾选该文件将自动覆盖平台现有文件。
d. 单击“导入”，导入成功的插件将在“工作台 > 插件”页面中展示。

8.5 创建与管理知识库

8.5.1 知识库介绍

平台提供了知识库功能来管理和存储数据，支持为应用提供自定义数据，并与之进行互动。

知识库支持导入以下格式的本地文档：

- 文本文档数据。支持上传常见文本格式，包括：txt、doc、docx、pdf、ppt、pptx格式。
- 表格数据。支持上传常见的表格文件格式，便于管理和分析结构化数据，包括：xlsx、xls、csv格式。

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。

8.5.2 创建知识库

创建知识库的步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 单击左侧导航栏“工作台”，在“知识库”页签，单击右上角“创建知识库”。
4. 在“创建知识库”页面，填写知识库名称与描述，单击图片可上传知识库图标，单击“下一步”。
5. 在“文件类型”页面，选择文件类型。
   - 导入文本文档数据。支持上传txt、doc、docx、pdf、ppt、pptx格式的文本文档，要求单个文件不超过10M。
   - 导入表格数据。支持上传xlsx、xls、csv格式的表格数据，要求单个文件不超过10M。
6. 单击“点此上传”上传本地文件至知识库。支持单次上传文件个数不超过300个。
7. 上传完成后，单击“确定”，完成知识库的创建。

说明

知识库创建完成后，如果想在当前知识库中继续上传文件，可单击该知识库进入详情页面，再单击右上角“继续上传”，上传本地文件。

知识库命中测试

平台支持对创建的知识库进行命中测试，以评估知识库的效果和准确性。

命中测试通过将用户的查询与知识库中的内容进行匹配，最终输出与查询相关的信息，并根据匹配的程度进行排序。

知识库命中测试步骤如下：

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 知识库”页面，单击所需知识库，进入知识库基本信息页面，单击右上角“命中测试”。
4. 在文本框中输入问题，单击“命中测试”，页面下方将展示多条匹配的内容，并按照匹配分值降序排列。

   用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。
5. 单击“查看历史”，可以查看用户输入的历史问题。

8.5.3 管理知识库

Agent开发平台支持对知识库执行获取知识库ID、删除、命中测试操作。

新增、删除知识库中知识文档

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 知识库”页面。
4. 单击所需知识库，进入详情页面。
   - 新增知识库中知识文档。单击右上角“继续上传”，可上传本地文档至当前知识库。
   - 删除知识库中知识文档。在“知识文档”中单击操作列“删除”可删除当前知识文档。

获取知识库 ID、删除知识库

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，跳转至Agent开发平台。
3. 进入“工作台 > 知识库”页面。

- 单击“ ··· > 复制ID”，可获取当前知识库ID。
- 单击“ ··· > 删除”，可删除当前知识库。

说明

删除应用属于高危操作，删除前，请确保该知识库不再使用。

8.6 Agent 开发常见报错与解决方案

工作流常见错误码与解决方案

工作流常见报错及解决方案请详见表8-15。

表 8-15 工作流节点常见报错与解决方案

模块名称	错误码	错误描述	解决方案
开始节点	101501	开始节点全局配置未传入值。	开始节点错误，请联系客服解决。
结束节点	101531	结束节点初始化失败。	检查结束节点配置，可能为校验报错。
	101532	结束节点模板拼接失败。	先检查模板占位符与输入是否匹配，请联系客服解决。
	101533	结束节点流式处理失败。	请联系客服解决。
大模型节点	101561	大模型节点初始化失败。	检查大模型节点配置，可能为校验报错。
代码节点	101591	代码组件初始化失败。	检查代码节点配置，可能为校验报错。
	101592	代码节点安全沙箱请求失败。	请联系客服解决。
	101593	代码节点安全沙箱执行失败。	检查代码的语法是否有误，检查是否用到了未引用的变量。
	101594	代码组件安全沙箱其他报错。	请联系客服解决。
	101595	代码节点执行失败未知错误。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
消息节点	101651	消息组件初始化失败。	检查消息节点配置，可能为校验报错。
	101652	消息节点缺少模板信息。	配置消息节点的提示词模板。
	101653	消息节点模板拼接错误。	先检查模板占位符与输入是否匹配，若仍无法解决，请联系客服解决。
	101654	消息组件执行失败。	请联系客服解决。
	101655	消息组件异步执行失败。	请联系客服解决。
意图识别节点	101098	意图识别prompt模板请求失败。	检查模板占位符与输入是否匹配。
	101097	意图识别调用大模型的prompt不符合模型输入的规范。	检查输入的prompt格式，消息的角色和内容。
	101096	意图识别调用大模型失败。	检查消息的格式，内容以及大模型服务是否正常。
	101095	意图识别用户query输入/引用解析失败。	检查用户query格式和内容。
	101094	意图识别prompt模板构建失败。	检查内置模板以及输入的system prompt格式与内容。
提问器节点	101043	当单个提问器内的对话轮数超过预设轮数上限时触发该错误码，对话状态回到开始节点状态。	可通过调大对话轮数上限解决。
	101047	初始化深度定制前后处理模块失败时触发该错误码。	可检查护栏配置是否符合要求。
	101048	执行深度定制用户回复改写（前处理）失败时触发该错误码。	可检查前处理护栏代码。
	101049	执行深度定制大模型生成的参数取值改写（后处理）失败时触发该错误码。	可检查后处理护栏代码。

模块名称	错误码	错误描述	解决方案
	101050	执行默认护栏（时间参数解析）失败时触发该错误码。	可检查支持处理的时间类型是否超出支持范围。
	102053	提示词模板有误时触发该错误码。	检查提示词模板是否格式有误。
	103004	大模型推理失败时触发该错误码。	请检查模型服务是否可以正常运行。
插件节点	101741	插件组件初始化失败。	检查插件组件配置，可能为校验报错。
	101742	工作流插件节点参数类型转换时出错。	根据error message确定具体转换出错的参数名称，并确认类型是否正确。
	101743	工作流插件节点的input在插件定义中不存在。	检查插件定义和对应的组件定义是否匹配。
	101744	插件定义了response，但实际插件执行结果与定义不一致。	检查插件response定义和实际插件执行结果是否匹配。
	101745	工作流插件节点执行出错。	插件执行出错，可以根据具体的error message信息定位。如果message无有效信息，说明该错误属于未捕获到的异常。
	105001	插件执行时发生了无法捕获的异常。	检查插件本身是否可用。
	105004	插件定义时check param error。	根据对应error message信息确定具体出错的参数定义。
	105005	插件定义不合法。	插件定义时的数据不合法，例如字段定义超出最长长度，具体根据error message判断。
	105008	插件内部错误。	请联系客服解决。
	105010	插件运行时鉴权出错。	可根据error message信息确定具体出错的鉴权问题，并检查鉴权信息的传递和插件鉴权定义是否正确。
	105011	插件运行返回的响应代码非200。	可根据报信息查看实际的http返回码。
	105012	插件request请求超时。	插件请求超时，检查插件服务。

模块名称	错误码	错误描述	解决方案
	105013	插件返回结果过大。	当前支持10M大小的返回，超过此大小会报错。
	105014	插件request proxy error。	请检查插件服务是否有问题导致无法连接。
认证鉴权	110000	认证失败。	查看认证配置。
	110001	用户信息获取失败。	查看用户信息是否正确配置。
工作流	112501	工作流认证失败。	查看认证配置。
	112502	缺少必要参数。	从打印日志可以看出当前缺失何种参数。
	112503	工作流连接数据库失败。	请联系客服解决。
	112504	缺少必要权限。	查看当前用户权限。
	112513	工作流流程中存在死循环。	检查工作流画布。
	112514	工作流被引用，无法删除。	查看知识型应用中是否引用了该工作流。
	112600	workflow ir转化失败	需要查看工作流配置是否正确。
	112941	获取workflow对话历史失败	请联系客服解决。

9 管理盘古大模型空间资产

9.1 盘古大模型空间资产介绍

在ModelArts Studio大模型开发平台的空间资产中，包括数据和模型两类资产。这些资产为用户提供了集中管理和高效操作的基础，便于用户实现统一查看和操作管理。

- **数据资产**：用户已发布的数据集将作为数据资产存放在空间资产中。用户可以查看数据集的详细信息，包括数据格式、大小、配比比例等。同时，平台支持数据集的删除等管理操作，使用户能够统一管理数据集资源，以便在模型训练和分析时灵活调用，确保数据资产的规范性与安全性。
- **模型资产**：平台提供的模型资产涵盖了预置或训练后发布的模型，所有这些模型将存放于空间资产中进行统一管理。用户可查看预置模型的历史版本和操作记录，还可以执行模型的进一步操作，包括训练、压缩、部署等。此外，平台支持导出和导入盘古大模型的功能，使用户能够将其他局点的盘古大模型迁移到本局点，便于模型资源共享。

9.2 管理盘古数据资产

数据资产介绍

数据资产是指在平台中被纳入管理、存储并可供使用的数据集。

数据资产包含以下两种形式：

- **用户自行发布的数据集**。

  用户可以通过“数据工程 > 数据发布 > 数据流通”功能将数据集发布为数据资产。发布的数据集支持查看详细信息、编辑、删除以及发布至AI Gallery等操作。
- **从AI Gallery订阅的数据资产**。

  除了用户自行发布的数据集，平台还提供了从AI Gallery中订阅数据资产的功能。AI Gallery提供了模型、数据集、AI应用等AI数字资产的共享，为企业级或个人开发者等群体，提供安全、开放的共享及交易环节。

发布数据资产至 AI Gallery

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击数据资产（状态为“未发布到Gallery”）操作列的“发布到Gallery”，对数据资产进行发布。
4. 在“发布到AI Gallery”页面填写AI Gallery资产名称与描述，选择可订阅区域约束与可看范围，单击“确定”，发布数据资产至AI Gallery。
5. 数据资产列表页将显示发布数据资产的状态：
   - 如果状态为“发布中”，表示该资产正在同步至AI Gallery，请耐心等待。
   - 如果状态为“发布成功”，表示该资产已同步至AI Gallery，可单击操作列“查看发布信息”以查看该资产的发布信息。
   - 如果状态为“发布失败”，表示该资产未成功同步至AI Gallery，可单击“发布到Gallery”重新对数据资产进行发布。

从 AI Gallery 订阅数据资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击右上角“订阅数据”，在“从AI Gallery订阅”页面选择需订阅的数据资产，单击“下一步”。
4. 填写资产名称与资产描述后，单击“确定”实现数据资产的订阅。
5. 数据资产列表页将显示订阅数据资产的状态：
   - 如果状态为“订阅中”，表示该资产正从AI Gallery同步中，请耐心等待。
   - 如果状态为“订阅成功”，表示该资产已从AI Gallery订阅成功，可单击操作列“查看订阅信息”以查看该资产的订阅信息。
   - 如果状态为“订阅失败”，表示该资产未成功从AI Gallery订阅，可单击“重新订阅”重新从AI Gallery订阅数据资产。
6. 订阅成功后的数据资产，将在“数据工程 > 数据获取 > 原始数据集”中显示，可执行后续的数据加工及发布操作。

管理数据资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“空间资产 > 数据”。
3. 单击“任务发布”页签，可对用户自行发布的数据集执行以下操作：
   - 查看基本信息。单击具体数据资产，可查看资产的配比详情、数据详情等基本信息。
   - 发布至Gallery。单击操作列的“发布至Gallery”，可发布数据资产至AI Gallery。
   - 查看发布信息。单击“查看发布信息”，查看该资产的发布信息（该操作需提前发布该数据资产至AI Gallery）。
   - 编辑属性。单击操作列的“更多 > 编辑属性”，可编辑数据资产的名称、描述以及资产可见性。

- 删除。单击操作列的“更多 > 删除”，可删除当前数据资产。
- 取消发布至Gallery。单击操作列的“更多 > 取消发布至Gallery操作”，可将已发布至AI Gallery的数据资产取消发布（该操作需提前发布该数据资产至AI Gallery）。

4. 单击“AI Gallery”页签，可对从AI Gallery订阅的数据资产执行以下操作：
   - 查看订阅信息。单击具体数据资产或操作列的“查看订阅信息”，查看该资产的名称描述等订阅信息。
   - 编辑属性操作。单击操作列的“更多 > 编辑属性”，可编辑数据资产的名称、描述以及资产可见性。
   - 删除操作。单击操作列的“更多 > 删除”，可删除当前数据资产。
   - 重新订阅。如果订阅失败，单击操作列的“重新订阅”，可重新从AI Gallery订阅所需数据资产。

9.3 管理盘古模型资产

模型资产介绍

用户在平台中可试用、已订购或训练后发布的模型，将被视为模型资产并存储在空间资产内，方便统一管理与操作。用户可以查看模型的所有历史版本及操作记录，从而追踪模型的演变过程。同时，平台支持一系列便捷操作，包括模型训练、压缩和部署，帮助用户简化模型开发及应用流程。这些功能有助于用户高效管理模型生命周期，提高资产管理效率。

模型资产包含以下两种形式：

- **预置模型。**
  用户在平台中可试用、已订购的预置模型。
- **用户自行发布的模型**。
  用户可以将训练完成的模型发布为模型资产。发布的模型支持查看详细信息、编辑属性、删除、导出、导入等操作。

管理模型资产

1. 登录ModelArts Studio大模型开发平台，在“我的空间”模块，单击进入所需空间。
2. 在左侧导航栏中选择“空间资产 > 模型”。
3. 单击“预置”页签，在预置模型列表，单击模型，可对预置的模型资产执行以下操作：
   - 查看模型历史版本。在“版本列表”页面，可查看模型的各个版本。
   - 训练、压缩、部署操作。在“版本列表”页面，可对不同版本模型执行训练、压缩或部署操作。单击相应按钮，将跳转至相关操作页面。
   - 查看操作记录。在“操作记录”页面，可查看当前模型的操作记录。

- 单击“本空间”页签，可对用户在当前空间发布的模型执行以下操作：
  - 查看模型信息。单击模型名称，进入模型信息页面，可产模型的基本信息与操作记录。
  - 编辑属性。单击操作列的“编辑属性”，可修改模型资产名称、描述以及资产可见性。

- 训练、压缩、部署。可在模型列表页面，对模型执行训练、压缩或部署操作。单击相应按钮，将跳转至相关操作页面。

导出盘古大模型至其他局点

导出盘古大模型至其他局点前，请确保当前空间为该用户所创建的空间。

模型训练发布完成后，可以通过导出模型功能将本局点训练的模型导出，导出后的模型可以通过**导入其他局点盘古大模型**，导入至其他局点进行使用。

以从环境A迁移模型到环境B为例：

1. 登录环境B的ModelArts Studio大模型开发平台，在“空间资产 > 模型”页面，单击右上角的“导入模型”。
2. 在“导入模型”页面，下载用户证书。

图 9-1 下载用户证书

3. 登录环境A的ModelArts Studio大模型开发平台，在“空间资产 > 模型 > 本空间”页面，单击支持导出的模型名称，右上角的“导出模型”。
4. 在“导出模型”页面，选择需要导出的模型，应设置导出模型时对应的导出位置（OBS桶地址），添加从环境B中下载的用户证书。设置完成后单击“确定”导出模型。

图 9-2 导出模型

导入其他局点盘古大模型

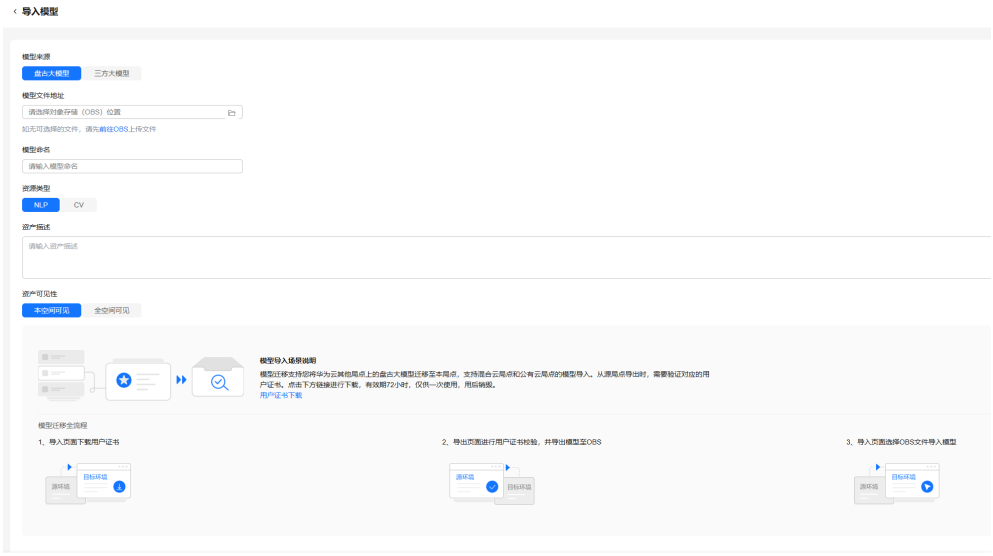
导入盘古大模型前，请确保当前空间为该用户所创建的空间。

导入模型功能可以将其他局点训练的模型导入本局点进行使用。

导入模型前，请参考**导出盘古大模型至其他局点**完成模型导出操作。

1. 登录ModelArts Studio大模型开发平台，在“空间资产 > 模型”页面，单击右上角的“导入模型”。
2. 在“导入模型”页面，模型来源选择“盘古大模型”。输入模型对应的obs地址和模型命名、选择资源类型、输入资产描述并设置资产可见性后，单击“确定”，启动导入模型任务。

图 9-3 导入模型



10 管理盘古大模型资源池

10.1 创建边缘资源池

边缘部署是指将模型部署到用户的边缘设备上，这些设备通常是用户自行采购的服务器，通过ModelArts服务纳管为边缘资源池，然后利用盘古大模型服务将模型部署到这些边缘资源池中。

ModelArts边缘节点是ModelArts平台提供的用于部署边缘服务的终端设备。创建边缘资源池之前需先创建ModelArts边缘节点。节点创建完成后，同步下载证书和边缘Agent固件，及时将固件复制到节点上，并执行注册命令完成设备的注册。

创建边缘资源池的流程见**表10-1**。

表 10-1 创建边缘资源池

操作步骤	说明
准备工作	说明创建边缘资源池的前期准备。
步骤1：注册边缘资源池节点	说明注册边缘资源池节点步骤。
步骤2：搭建边缘服务器集群	说明搭建边缘服务器集群的步骤。
步骤3：安装Ascend插件	说明安装Ascend插件指导。
步骤4：创建证书	说明创建负载均衡所需证书步骤。
步骤5：创建负载均衡	说明创建负载均衡步骤。

说明

- ModelArts Studio大模型开发平台当前仅部分模型支持边缘部署，详见《产品介绍》“模型能力与规格”章节。
- 使用边缘部署功能需要在ModelArts服务中开通“边缘资源池”功能，该功能为**白名单特性**，需要联系ModelArts客服进行开通。
- **创建边缘资源池操作较为复杂，建议联系盘古客服进行协助。**

准备工作

本章节的边缘部署操作以largemodel集群为例，示例集群信息如下表。

表 10-2 示例集群信息

集群名	节点类型	节点名	规格	备注
largemodel	controller	ecs-edge-XXXX	鲲鹏通用计算型|8vCPUs|29GiB|rc3.2xlarge.4镜像 EulerOS 2.9 64bit with ARM for Tenant 20230728 base 2.9.15	公网IP：100.85.220.207 root密码：/ CPU架构：aarch64（登录设备，执行arch命令查看）
	worker	bms-panguXXXX	CPU:Kunpeng 内存：24*64GB DDR4 RAM(GB) 本地磁盘：3*7.68TB NVMe SSD 扩展配置：2*100GE +8*200GE 类型：physical.kat2e.48xlarge.8.313t.ei.pod101 euler2.10_arm_sdi3_1980b_hc_sdi5_b080_20230831v2	公网IP：100.85.216.151 root密码：/ CPU架构：aarch64（登录设备，执行arch命令查看）

1. 依赖包下载。
   - docker下载：**https://download.docker.com/linux/static/stable**

   选择对应cpu架构下载，docker版本选在19.0.3+。
   - K3S下载：**https://github.com/k3s-io/k3s/releases/tag/v1.21.12%2Bk3s1**

   按照对应cpu架构下载二进制文件以及air-gap镜像。
2. npu驱动和固件安装。

   执行命令**npu-smi info**查看驱动是否已安装。如果有回显npu卡信息，说明驱动已安装。

   详情请参见**昇腾官方文档**。
3. hccn too网卡配置。
   a. 执行如下命令，查看是否有回显网卡信息。如果有，则说明网卡已经配置，否则继续操作下面步骤。
   ```
   cat /etc/hccn.conf
   ```
   b. 执行如下命令，查看npu卡数。
   ```
   npu-smi info
   ```

c. 执行如下命令（地址自行配置）：

```
hccn_tool -i 0 -ip -s address 192.168.0.230 netmask 255.255.255.0
hccn_tool -i 1 -ip -s address 192.168.0.231 netmask 255.255.255.0
hccn_tool -i 2 -ip -s address 192.168.0.232 netmask 255.255.255.0
hccn_tool -i 3 -ip -s address 192.168.0.233 netmask 255.255.255.0
hccn_tool -i 4 -ip -s address 192.168.0.234 netmask 255.255.255.0
hccn_tool -i 5 -ip -s address 192.168.0.235 netmask 255.255.255.0
hccn_tool -i 6 -ip -s address 192.168.0.236 netmask 255.255.255.0
hccn_tool -i 7 -ip -s address 192.168.0.237 netmask 255.255.255.0
```

d. 执行命令**cat /etc/hccn.conf**，确保有如下回显网卡信息，则配置完成。

```
[root@bms-pang[redacted] k3s]# hccn_tool -i 0 -ip -s address 192.168.0.230 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 1 -ip -s address 192.168.0.231 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 2 -ip -s address 192.168.0.232 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 3 -ip -s address 192.168.0.233 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 4 -ip -s address 192.168.0.234 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 5 -ip -s address 192.168.0.235 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 6 -ip -s address 192.168.0.236 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]# hccn_tool -i 7 -ip -s address 192.168.0.237 netmask 255.255.255.0
[root@bms-pang[redacted] k3s]#
[root@bms-pang[redacted] k3s]# cat /etc/hccn.conf
address_0=192.168.0.230
netmask_0=255.255.255.0
address_1=192.168.0.231
netmask_1=255.255.255.0
address_2=192.168.0.232
netmask_2=255.255.255.0
address_3=192.168.0.233
netmask_3=255.255.255.0
address_4=192.168.0.234
netmask_4=255.255.255.0
address_5=192.168.0.235
netmask_5=255.255.255.0
address_6=192.168.0.236
netmask_6=255.255.255.0
address_7=192.168.0.237
netmask_7=255.255.255.0
[root@bms-pangu30037210 k3s]#
```

4. 配置NFS网盘服务。

   大模型采用镜像+模型分开的方式部署时，需要有一个节点来提供nfs网盘服务，创建部署时通过nfs挂载的方式访问模型。

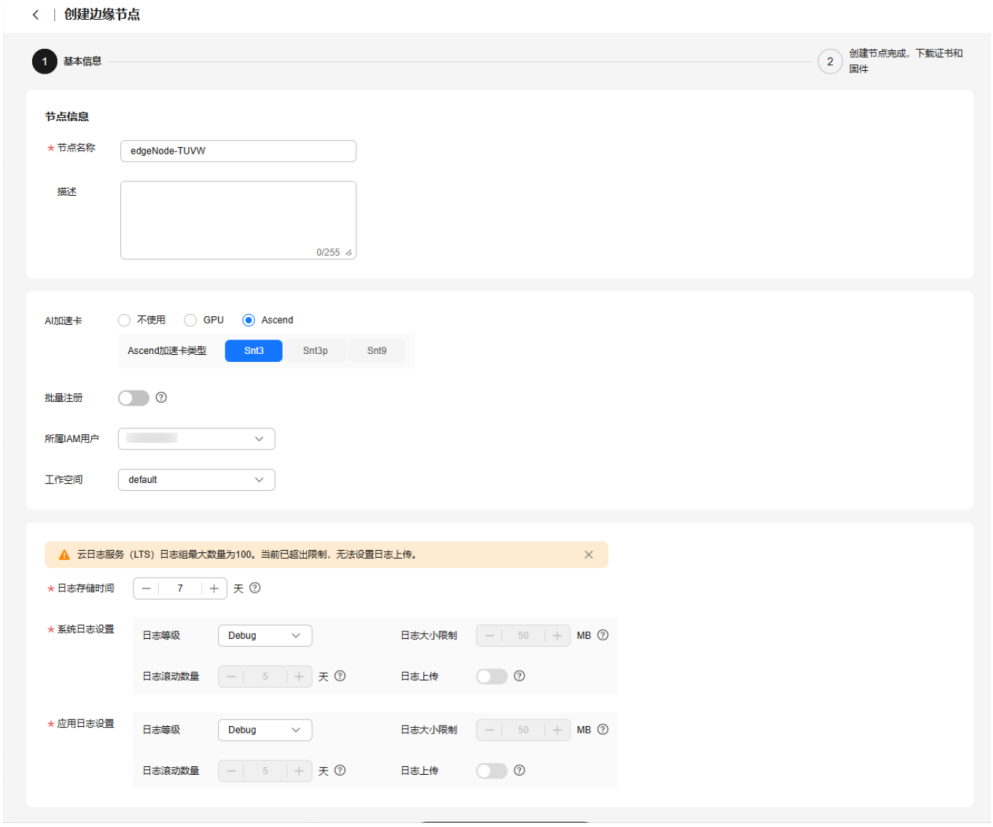
步骤 1：注册边缘资源池节点

1. 进入ModelArts服务，选择所需空间。
2. 在左侧列表中选择“资源管理 > 边缘资源池 Edge”，在“节点”页签中，单击“创建”。

图 10-1 创建边缘资源池节点

3. 在“创建边缘节点”页面中，填写节点名称，配置AI加速卡与日志信息。
   - 如果节点有npu设备需选择“AI加速卡 > Ascend”，并选择加速卡类型。
   - 如果节点没有加速卡，则选择“AI加速卡 > 不使用”。

图 10-2 配置边缘节点参数



4. 基本信息配置完成后，单击“确定”，单击“立即下载”，下载设备证书和Agent固件，并将设备证书与Agent固件分别重命名为**license.tgz**、**hilens-agent.tgz**。

图 10-3 下载所需证书

步骤 2：搭建边缘服务器集群

1. 执行如下命令，生成docker证书。

```
bash cluster_install-ascend.sh generate_docker_cert --pkg-path=/home/hilens/pkgs
```

说明

请注意，该命令只需执行一次，如果已有相关证书，请跳过该步骤。

2. 基于**准备工作**与**步骤1：注册边缘资源池节点**，按照以下目录结构存放下载文件，注意修改下载文件的命名。其中，docker下的certs证书会自动生成，一般无需修改。

```
pkgs // 包目录，用户自行命名
    docker
        docker.tgz    // docker 二进制文件，要求版本>19.0.3
        certs         // 使用generate命令生成的证书，指定--pkg-path后会自动创建到certs目录
            ca.crt
            server.crt
            server.key
    k3s
        k3s // k3s可执行文件
        agent
            images
                k3s-airgap-images-[arm64|amd64].tar.gz //k3s离线镜像
    hilens-agent
        hilens-agent.tgz // hilens agent固件包
        license.tgz // hilens 设备license
```

3. 工作节点执行命令如下：

```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=worker --host-ip=192.168.0.209
```

主控节点执行命令如下：

```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=controller --host-ip=192.168.0.150
```

- **cluster_install-ascend.sh**脚本主要用于安装docker、hdad和k3s，请联系盘古客服获取。
- **pkg-path**是步骤2中整合的安装包文件目录。
- **host-ip**是设备在集群中的ip，一般为内网ip。
- **node-type**是集群节点类型。其中，worker表示工作节点，controller表示主控节点。

4. 在服务器执行如下命令，判断docker是否安装成功。

```
systemctl status docker
```

```
[root@bms-pangu        hilens-agent]# systemctl status docker
● docker.service - Docker Application Container Engine
   Loaded: loaded (/usr/lib/systemd/system/docker.service; enabled; vendor preset: disabled)
   Active: active (running) since Wed 2023-11-01 14:23:17 CST; 2h 32min ago
     Docs: https://docs.docker.com
 Main PID: 1629388 (dockerd)
    Tasks: 65
   Memory: 31.0M
   CGroup: /system.slice/docker.service
           ├─1629388 /usr/bin/dockerd -H fd://
           └─1629412 containerd --config /var/run/docker/containerd/containerd.toml --log-level info
```

5. 在服务器执行如下命令，判断edge agent是否安装成功。

```
hdactl info
```

```
[root@bms-pangu        hilens-agent]# hdactl info
{
 "device_info": {
  "device_id": "hilens-f1bf45d27c464413b6b53fa40d9a1c0d",
  "device_ip": "192.168.0.209",
  "device_name": "edgeNode-910b",
  "device_type": "General ARM Device",
  "domain": "",
  "is_active": true,
  "obs": "",
  "projectId": "51fae1a6ef804316abeaadb80e302339"
 },
 "dm": {
  "state": "success",
  "connect_err": "",
  "last_message": {
   "topic": "pong",
   "timestamp": 1698829033
  },
  "channel_available": true,
  "address": "wss://modelarts-dev.cn-north-7.myhuaweicloud.com/v3/login"
 }
}
```

6. 配置NFS网盘服务。

   a. **安装NFS服务**

   该步骤需要设备联网下载软件依赖包。

   - **Ubuntu系统**

     在线安装：

     ```
     sudo apt install nfs-kernel-server
     ```

   - **Euler OS系统**

     在线安装：

     ```
     sudo yum install nfs-utils
     ```

     说明

     若需离线安装，请联系盘古客服。

   - 防火墙需要打开rpc-bind和nfs的服务（可选）：

     ```
     sudo firewall-cmd --zone=public --permanent --add-service={rpc-bind,mountd,nfs}
     sudo firewall-cmd --reload
     ```

   b. **创建网盘共享目录**

   该路径的存储空间能够存储大模型文件。此处设置/var/docker/hilens作为网盘根目录，将会在容器里访问该路径。执行：

   ```
   sudo mkdir -p /var/docker/hilens
   ```

   该路径的访问权限需设置为：1000:100（与/etc/exports配置保持一致），执行：

   ```
   sudo chmod 750 /var/docker/hilenschown -R 1000:100 /var/docker/hilens
   ```


查看权限，执行：

```
ls -l /var/docker | grep hilens
```

c. **添加网盘访问权限**

配置nfs-server访问白名单和网盘共享文件的路径。执行：

```
sudo vim /etc/exports
```

添加如下配置：

```
/var/docker/hilens 172.xxx.0.0/24(rw,no_all_squash,anonuid=1000,anongid=100,fsid=0)
```

172.xxx.0.0/24为集群内网IP网段（登录主控节点，使用hdactl info命令查看IP地址。比如查得IP地址为172.16.0.22，可配置为172.16.0.0/24网段）。

其中，

- /var/docker/hilens：网盘根目录路径。
- 192.168.0.0/24：客户端IP范围，表示IP在192.168.0.0/24范围的所有节点，都可以访问 /var/docker/hilens 。* 代表所有，即没有限制。也可以填写具体某个节点的IP。
- rw：权限设置，可读可写。
- anonuid：为映射的匿名用户id，anongid为映射的匿名用户组，也就是挂载进容器后，在容器中看到的文件属主。
- no_all_squash：可以使用普通用户授权。

执行**:wq**命令保存设置后，刷新nfs配置。执行：

```
exportfs -rv
```

d. **启动NFS和rpcbind**

设置服务开机启动、启动服务：

```
systemctl enable nfs-server && systemctl enable rpcbind && systemctl start rpcbind nfs-server
```

执行如下命令，验证以上配置内容是否正确。如下图，表示配置正确，即NFS服务安装成功。

```
showmount -e localhost
```

e. **验证NFS配置（可选）**

在非NFS服务节点创建目录：

```
sudo mkdir ~/data
```

执行挂载：

```
sudo mount -t nfs 192.168.xx.xxx:/var/docker/hilens ~/data
```

挂载后，可以使用以下命令查看：

```
mount
```

回显如下，则成功：

```
...
...
192.168.0.150:/var/docker/hilens on ~/data type nfs4
```

f. **测试NFS功能**

在客户端向共享目录创建一个文件：

```
cd ~/data
sudo touch a
```

在NFS服务端192.168.0.150查看所创建的文件：

```
cd /var/docker/hilens
ls -l
```


g. **yum源配置**

如果yum install使用正常，请忽略该章节。

i. 配置yum内源地址

- 备份yum配置

```
mkdir -p /etc/yum.repos.d/bak/ mv -f /etc/yum.repos.d/*.repo /etc/yum.repos.d/bak/
```

- x86 eulerOS配置

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base]
name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/
x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1
[EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-
mirrors.huawei.com/install/euleros/2.10/devel_tools/x86_64/ gpgkey=file:///etc/pki/rpm-
gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1  [EulerOS_2.10_updates]
name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/
updates/x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0
enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

- arm eulerOS系统配置

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base]
name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/
aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1
[EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-
mirrors.huawei.com/install/euleros/2.10/devel_tools/aarch64/ gpgkey=file:///etc/pki/rpm-
gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1  [EulerOS_2.10_updates]
name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/
updates/aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1
enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

ii. 配置DNS

```
cat >> /etc/resolv.conf<<"EOF" nameserver 10.189.32.59 nameserver 10.72.55.103
nameserver 10.98.48.39 EOF
```

证书一般自动下载，但是有遇到有的服务器没有自动安装或者已经安装有问题的情况下，直接复制执行即可。

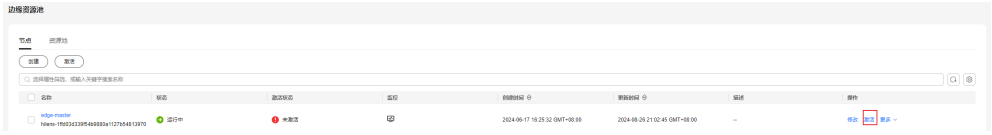
```
 cat > /etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS <<"EOF" -----BEGIN PGP PUBLIC KEY
BLOCK-----  mQENBFx4vK4BCACrq4PA8EZEr5XH08bfh9rlms
+QDZsJYhqqIXWx4qsZ8dAqdWLH
O2Dm0HYk17xVwTZjXUYH9rz1gn2bGa5At4xTpH7FHMDpNG8DfwC6UpMKEmGGvy/S
OfL4fl6Yq2tCHAx3LrHXO9PGigafz5XMDtBySI4ixOR4M/w8IaIPEyN0BfGC6DQ5
PZlXMPDOc5VW1NhxwH0u0yHH1lLITKBAEifTQa8+3YZY54PWvBbYxOcCCS63FOTn
pk0wWuwm2JqozDBxV/w8Ty0c25+y4FTiUGzOj2e/3K1ls+Zs/tTxf8asKpFH/dYN
ffXmdkPLDCHRGkTyrLPdghoFmVg4XhrOhsf/ABEBAAG0LkV1bGVyT1MgKEV1bGVy
T1MgMi4wIFNQOCkgPGV1bGVyb3NAaHVhd2VpLmNvbT6JAVQEEwEIAD4WIQS4wSv5
6CWqOEUVIHBxp+kP4ux1vAUCXHi8rgIbAwUJA8JnAAULCQgHAgYVCgkICwIEFgID
AQIeAQIXgAAKCRBxp+kP4ux1vN/UB/4jy2kRiFznSXfWz5SX0szLOf1FgDssdRZG
xASHonAJqrV19mkG2pNTkgip9LZQsCqLbxj5FV+TMm1o+6jubd9qRPePlV2Tpc0T
m3cDmpcZbW/XrFh4dLdN644TwKDAPcK8TK/wOepFVtjhx1Qc8o+8nuoFhmsMoKkf
AA1DYDDCZbpbgMQwMV3yKh002CFRlCnVMylqOi8U+FrYfphnsYfujXpKu9g+FmO6
ju0xVhxyFOVCEicamKiel3ZS9z06+PIFL3KP/nKC7pu0tfaxogJjCh8y+ZIF/FJU
ygHpZYKQJyJNO8Do7AucudbruOxqGqhD2BIBtX1JNP/hkKj4w+OJuQENBFx4vK4B
CAD1SYnEWn0mf3umbucovVYHaywJqhErB3ia+Ykq3InKvIORf1reiCRvVuse2wZr
cnPWEKeRE7tTEIZ5hCuLYQzaqngqVkwqLbRR6vtxiDhTWNgJH9+GuokgjotQ3/7T
AXH7AwG57OPp6vvaiazCDjhy5t3Vr1snWkiwWkJR2GFRkuwKu7FDLjc1n2dx4zLF
zRZJa5TTAR1zrHWgVwkLxgq0+eJHWq7eHFw1SBjmc4Vs4z/QI+Q+3rkVBiqCmr11 /
XQz9ZePdOl/8fCuNh4l480c7AFlFt8IvKBP6Kh+jShxbED2NJpRV04MhV6SyA69 ixd/
VKtPJMRcRg3phyuuCedrABEBAAGJATwEGAEIACYWIQS4wSv56CWqOEUVIHBx p
+kP4ux1vAUCXHi8rgIbDAUJA8JnAAAKCRBxp+kP4ux1vBGeB/4ubYvxZ6/apb+i
MCtRIuA15PWEwVFTVfKirvEliY4fAjw5HSlfrnN4FV/OCTIRHecuNBNBfLt78DoK
08x7fYtEBqIN6pDanjsSwbPhuzhz6m4C/GWLqqDi8SCaVTQsJqKc2QHjr7CaBIuo GRlB84/pOq
+kGAnMZPhCjy52K9x5zRpq6zTUpV5XPeLCBn6Kc8GW1Lk6K1eXsn09
KdIhb9JqTDtQx0eOS1p0fJlTb68Pj406IYJ16FaXmMTcYvep6HhVATQGbPDuIepd
BK12nQEDrezkGR5vH9nMraQuZTvADuRFFgQvZQ5QMYAzMa0RhQTHKMOwCMq8mBq7
3Csgrxwa =kXkt -----END PGP PUBLIC KEY BLOCK----- EOF
```

7. 配置hda.conf配置文件信息

a. 登录nfs服务节点，执行如下命令：
```
vi /etc/hilens/hda.conf
```
b. 增加如下配置：
```
hilens.nfs.server.ip=192.168.0.150
hilens.nfs.mount.dir=/home/mind/model
hilens.nfs.source.dir=/var/docker/hilens
```
其中，**server.ip**是nfs存储节点内网ip，**mount.dir**是大模型默认挂载路径，**source.dir**是大模型下载路径。
c. 配置完成后，执行如下命令重启固件：
```
systemctl restart hdad
```

8. 进入ModelArts服务，选择所需空间。进入“边缘资源池 Edge > 节点”，在当前设备节点操作列单击“激活”，节点状态将从“未激活”转为“已激活”。

图 10-4 激活边缘节点



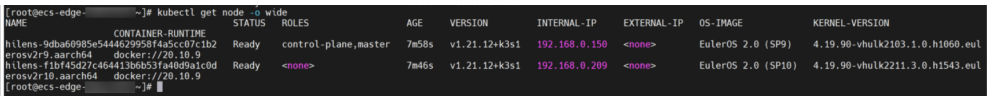
9. 进入“边缘资源池 > 资源池”，单击“创建”。填写资源池名称，选择“ModelArts边缘节点”，在“主控节点”处单击“添加”，选择要添加的主控节点，单击“确定”。

图 10-5 添加主控节点

10. 在“工作节点”处单击“添加”，选择要添加的工作节点，单击“确定”。

图 10-6 添加工作节点

11. 单击“立即创建”，可在资源池列表中查看节点的状态。如果状态为“运行中”，则创建成功。
12. 在主控节点执行如下k8s命令，验证边缘池创建结果：
    a. 执行如下命令建立软连接。
    ```
    ln -s /home/k3s/k3s /usr/bin/kubectl
    ```
    b. 执行如下命令查看节点状态。
    ```
    kubectl get node -o wide
    ```
    c. 如果所有节点状态STATUS为“Ready”，则说明集群创建成功。



步骤 3：安装 Ascend 插件

详情请参考官方文档：**https://www.hiascend.com/document/detail/zh/mindx-dl/60rc2/clusterscheduling/clusterschedulingig/clusterschedulingig/dlug_installation_001.html**

步骤 4：创建证书

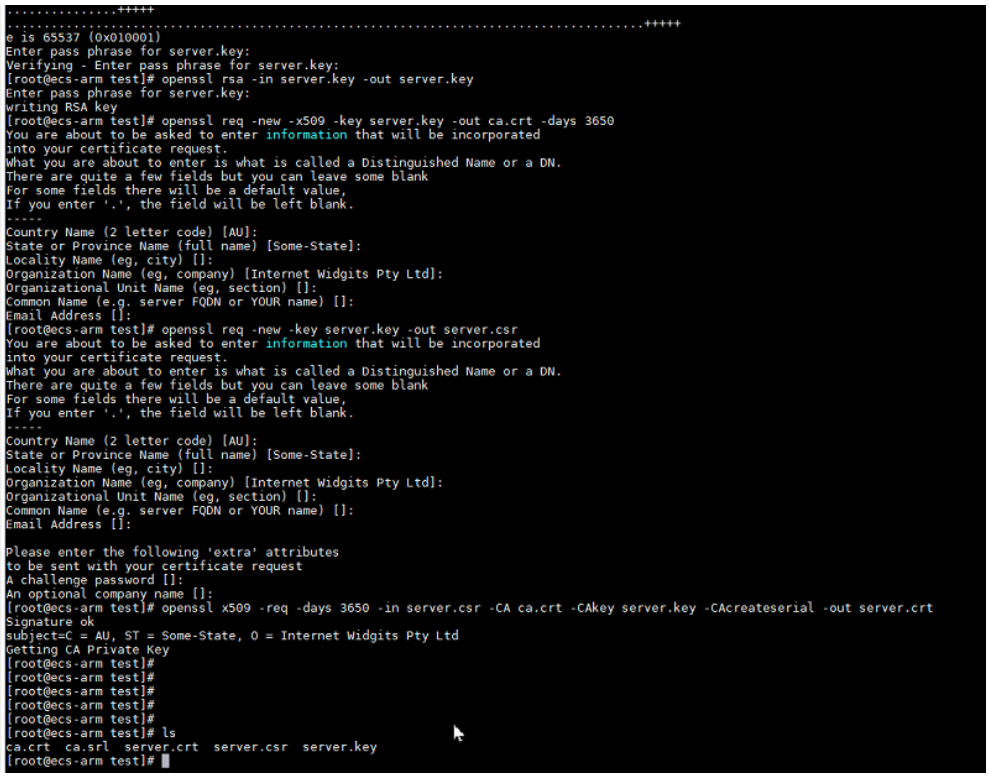
如**图10-7**，如果在“边缘资源池”页签提示无可用的证书，可以参考以下方法创建证书。

图 10-7 无可用证书

1. 准备一台Linux系统的服务器（已安装OpenSSL），依次执行以下命令制作证书。
   执行命令时会提示输入至少四位的密码，例如：123456，需记住密码后续步骤会使用。

```
1. 生成server.key命令：
openssl genrsa -des3 -out server.key 2048
2. 生成不加密的server.key:
openssl rsa -in server.key -out server.key
3. 生成ca.crt命令：
openssl req -new -x509 -key server.key -out ca.crt -days 3650
4. 生成server.csr命令：
openssl req -new -key server.key -out server.csr
5. 生成server.crt命令：
openssl x509 -req -days 3650 -in server.csr -CA ca.crt -CAkey server.key -CAcreateserial -out server.crt
```


图 10-8 命令执行示例

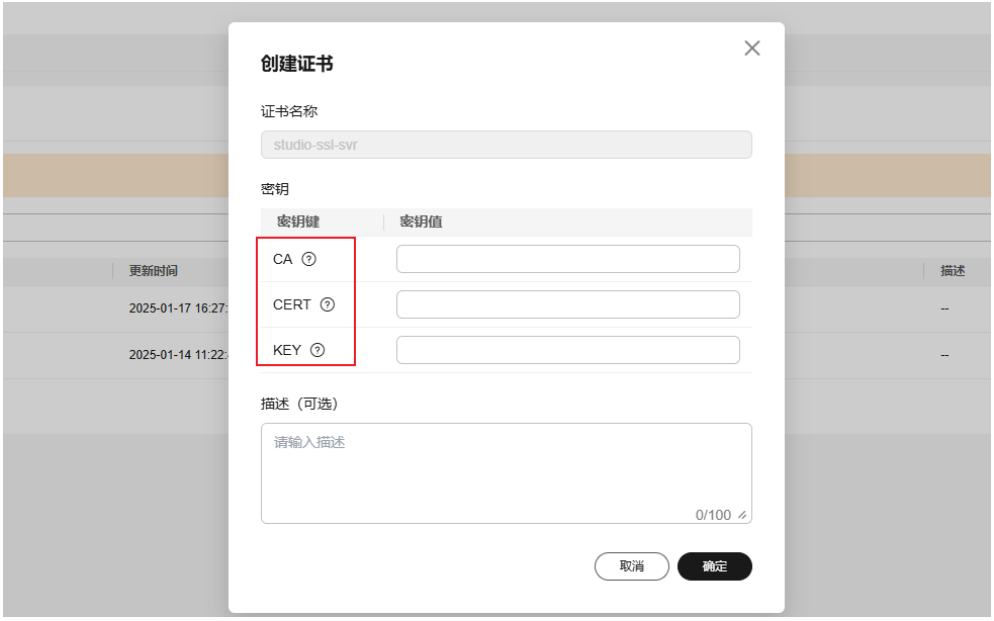


2. 证书制作完成后，执行ls命令可查看生成的证书文件。证书文件与ModelArts Studio平台中证书填写项对应关系如下。

   ca.crt -- CA
   server.crt -- CERT
   server.key -- KEY

3. 在ModelArts Studio平台首页，单击右上角“设置”，在“资源池管理 > 边缘资源池”页签单击“创建证书”，填写相应证书参数及描述。

图 10-9 证书密钥



4. 可通过view命令查看证书密钥值，例如view server.crt。输入ModelArts Studio平台的密钥值需要经过base64加密（echo -n "复制的密钥值内容" | base64）。

步骤 5：创建负载均衡

在边缘资源池创建完成后，需要返回ModelArts Studio平台“首页 > 设置”中为边缘池创建负载均衡、创建监控插件。

创建负载均衡步骤如下：

1. 登录ModelArts Studio平台，单击右上角“设置”。
2. 在“资源池管理 > 边缘资源池”页签查看已经创建的边缘资源池，单击操作列“创建负载均衡”。

图 10-10 边缘资源池

3. 填写负载均衡名称（按mas-xxx命名填写），设置侦听端口（取值在30000到40000之间）和协议类型（调用推理模型时使用http还是https请求）。设置完成后单击“确定”。

图 10-11 创建负载均衡