盘古大模型
3.0.0
用户指南

文档版本 01
发布日期 2024-12-02

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、科学计算大模型等服务，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、清洗、配比和管理等功能，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、部署、推理等功能。通过高效推理性能和跨平台迁移工具，保障模型在不同环境中的稳定、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应用开发与丰富的开发SDK，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

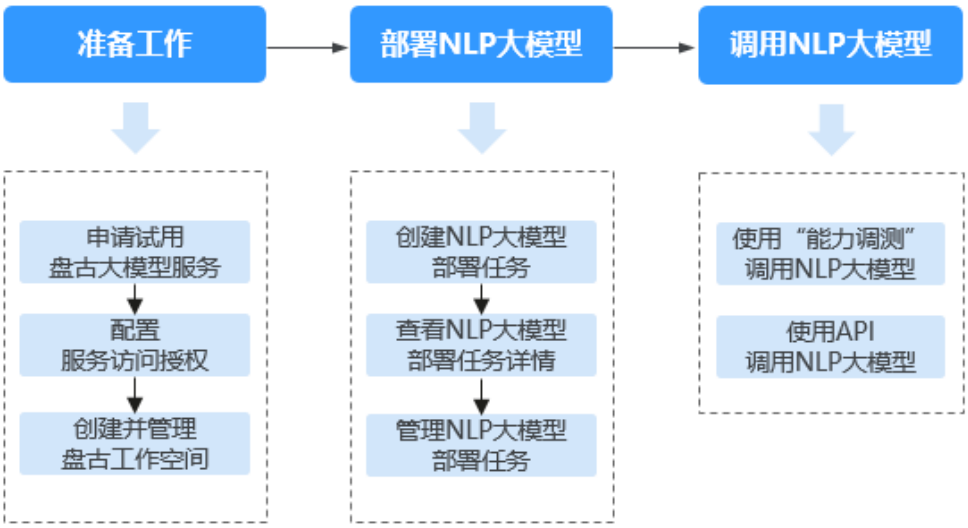


表 1-1 NLP 预置模型使用流程表

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**
	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**、《快速入门》“使用盘古预置NLP大模型进行文本对话”

流程	子流程	说明	操作指导
	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、《快速入门》“调用盘古NLP大模型API实现文本对话”

图 1-2 科学计算预置模型使用流程表

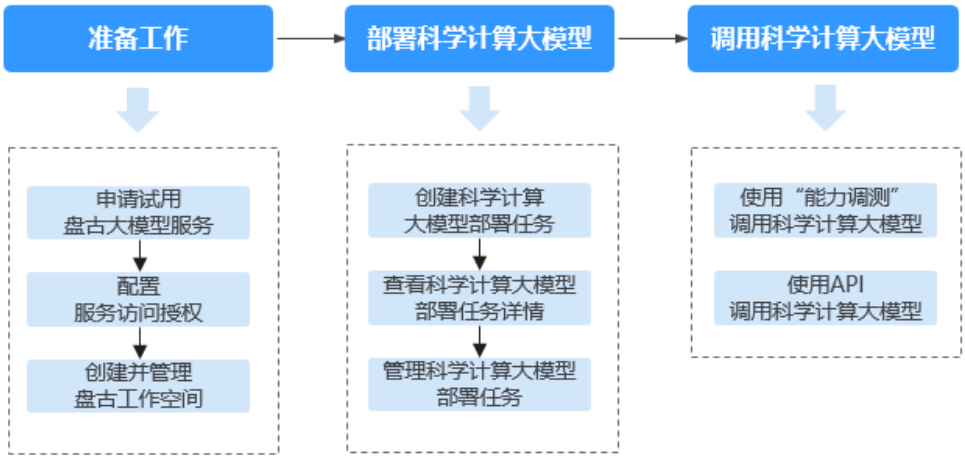


表 1-2 科学计算预置模型使用流程表

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
	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

流程	子流程	说明	操作指导
	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

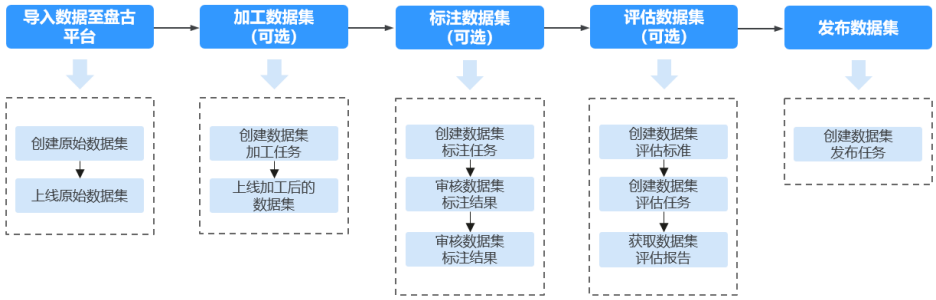


表 1-3 数据工程使用流程表

流程	子流程	说明
导入数据至盘古平台	创建原始数据集	数据集是指用于模型训练或评测的一组相关数据样本，上传至平台的数据将被创建为原始数据集进行统一管理。
	上线原始数据集	在正式发布数据集前，需要执行上线操作。
加工数据集（可选）	创建数据集加工任务	数据集中若存在异常数据，可通过数据集加工功能去除异常字符、表情符号、个人敏感内容等。
	上线加工后的数据集	对加工后的数据集执行上线操作。

流程	子流程	说明
标注数据集（可选）	创建数据集标注任务	创建数据集标注任务，并对数据集执行标注操作，标注后的数据可以用于模型训练。
	审核数据集标注结果	对数据集的标注结果进行审核。
	上线标注后的数据集	对标注后的数据集执行上线操作。
评估数据集（可选）	创建数据集评估标准	创建数据集评估标准。评估文本通顺性、信息充分性、内容有效性等。
	创建数据集评估任务	创建数据集质量评估任务，并基于评估标注对数据逐一评估其质量，评估后的数据可以用于模型训练。
	获取数据集评估报告	查看数据集评估任务的进展和数据集质量。
发布数据集	创建数据集发布任务	创建数据集发布任务，并进行正式的数据集发布操作，可用于后续的训练任务。 平台支持发布的数据集格式为**默认格式、盘古格式**，可按需进行数据集格式转换。 • **默认格式**：平台默认的格式。 • **盘古格式**：训练盘古大模型时，需要进行数据集格式转换。当前仅**文本类、图片类数据集支持转换为盘古格式。**

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、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

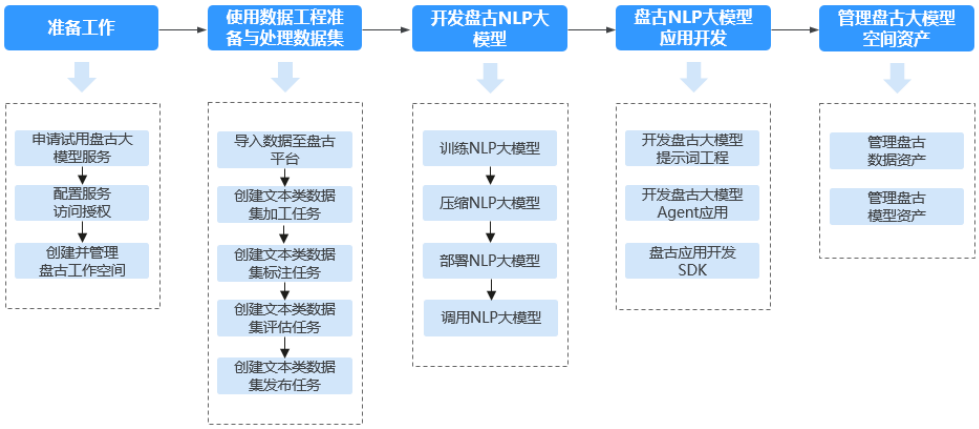


表 1-4 NLP 大模型开发流程表

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
	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
使用数据工程准备与处理数据集	导入数据至盘古平台	将用户数据导入至盘古平台的过程。	导入数据至盘古平台
	创建文本类数据集加工任务	数据集中若存在异常数据，可通过数据集加工功能去除异常字符、表情符号、个人敏感内容等。	创建文本类数据集加工任务
	创建文本类数据集标注任务	创建数据集标注任务，并对数据集执行标注操作，标注后的数据可以用于模型训练。	创建文本类数据集标注任务
	创建文本类数据集评估任务	评估文本通顺性、信息充分性、内容有效性等。	创建文本类数据集评估任务
	创建文本类数据集发布任务	创建数据集发布任务，并进行正式的数据集发布操作，可用于后续的训练任务。 平台支持发布的数据集格式为**默认格式、盘古格式**，可按需进行数据集格式转换。 • **默认格式：**平台默认的格式。 • **盘古格式：**训练盘古大模型时，需要进行数据集格式转换。当前仅**文本类、图片类数据集支持转换为盘古格式。**	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

流程	子流程	说明	操作指导
	部署NLP大模型	部署后的模型可进行调用操作。	部署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
管理盘古大模型空间资产	管理盘古数据资产	管理已发布的数据集。	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

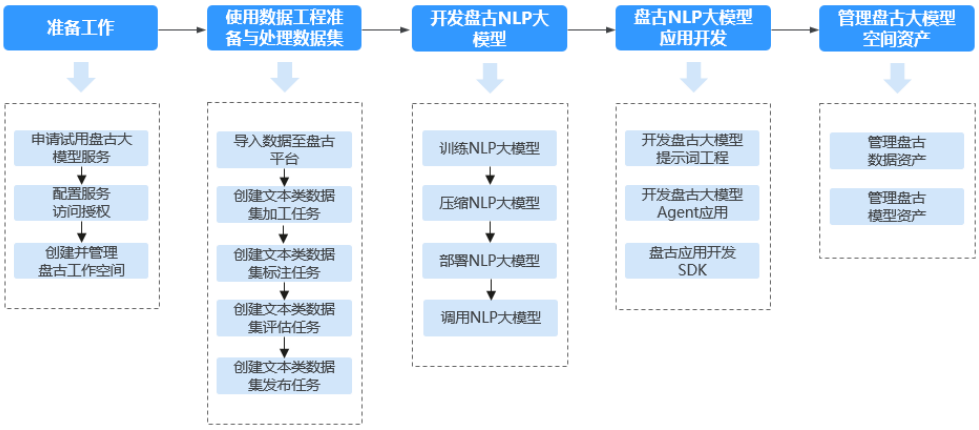


表 1-5 科学计算大模型开发流程表

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**
	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**
使用数据工程准备与处理数据集	导入数据至盘古平台	将用户数据导入至盘古平台的过程。	**导入数据至盘古平台**
	创建气象类数据集加工任务	数据集中若存在异常数据，可通过数据集加工功能去除异常字符、表情符号、个人敏感内容等。	**创建气象类数据集加工任务**
	创建气象类数据集发布任务	创建数据集发布任务，并进行正式的数据集发布操作，可用于后续的训练任务。	**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	部署后的模型可进行调用操作。	**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**

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，申请通过后试用盘古大模型服务。

1. 登录ModelArts Studio大模型开发平台首页。
2. 在首页单击“试用咨询”，申请试用盘古大模型服务。

图 2-1 申请试用

2.2 配置服务访问授权

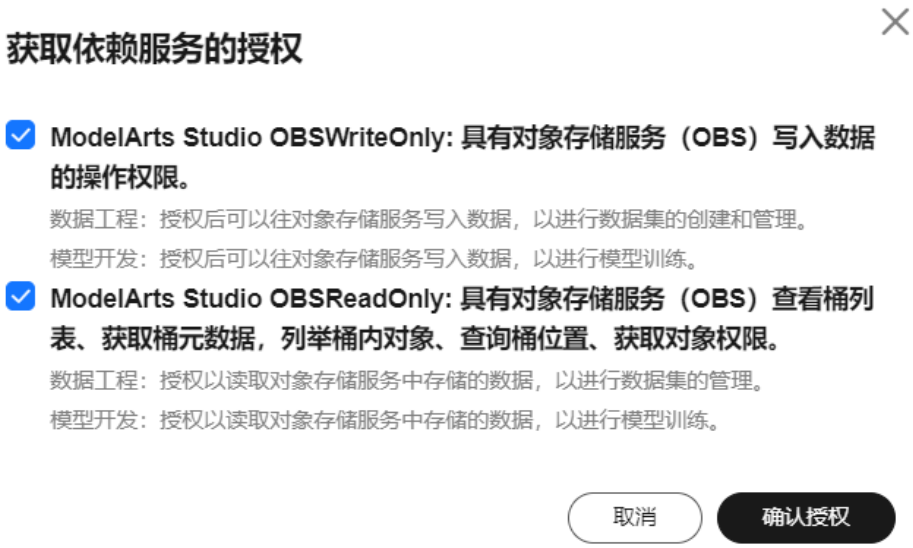
配置 OBS 访问授权

盘古大模型服务使用对象存储服务（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。

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“获取依赖服务的授权”弹窗选中授权，并单击“确认授权”。

图 2-2 配置 OBS 访问授权提示

图 2-3 配置 OBS 访问授权方式 1



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

图 2-4 配置 OBS 访问授权方式 2

2.3 创建并管理盘古工作空间

2.3.1 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。

2.3.2 创建并管理盘古工作空间

创建盘古工作空间

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

图 2-5 创建空间

4. 单击创建好的空间，进入ModelArts Studio大模型开发平台，平台支持数据工程、模型开发、Agent开发等功能。

   如果用户具备多个空间的访问权限，可在页面左上角单击 切换空间。

图 2-6 切换空间

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，进入所需操作空间。

图 2-7 进入操作空间

2. 单击左侧导航栏的“空间管理”，在“空间设置”页签中可执行如下操作：
   - 可修改当前空间的名称与描述。
   - 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。

2.3.3 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。

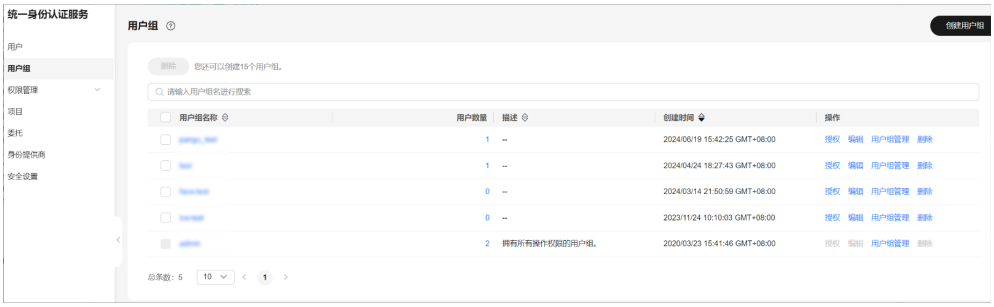
如果华为云账号已经能满足您的要求，不需要创建独立的IAM用户（子用户）进行权限管理，您可以**跳过本章节**，不影响您使用盘古的其他功能。

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

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

图 2-8 创建用户组



3. 在“创建用户组”界面，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 2-9 用户组授权

5. 参考表2-1，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 2-1 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 2-10 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 2-11 设置最小授权范围

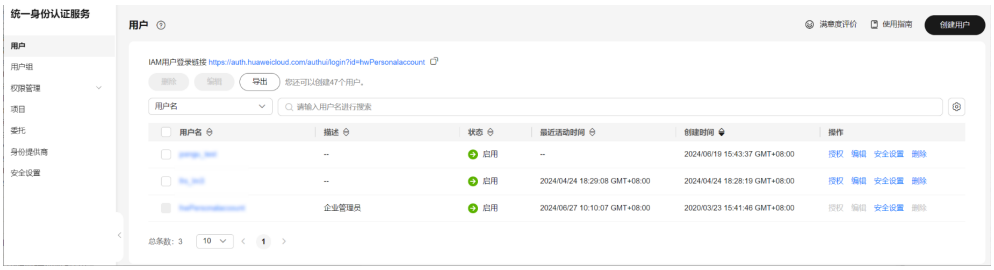
7. 单击“完成”，完成用户组授权。

图 2-12 完成授权

创建盘古子用户

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”。

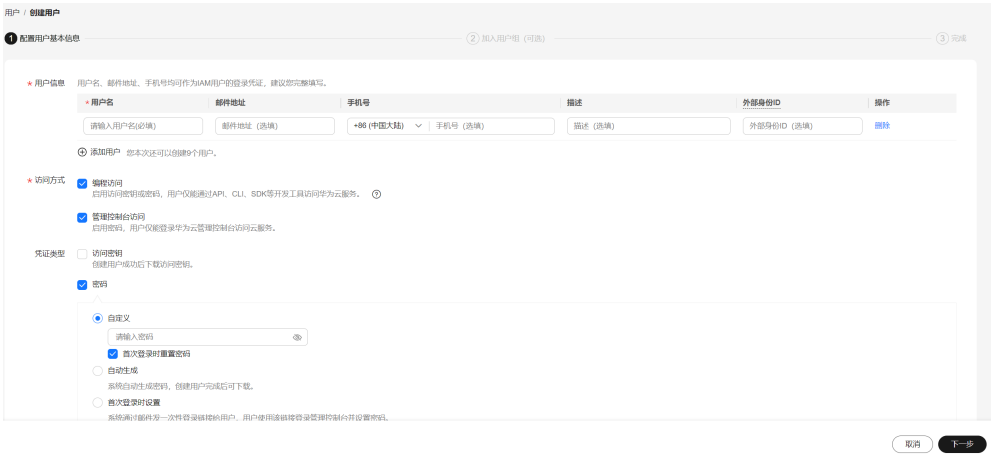
图 2-13 创建用户



3. 配置用户基本信息，单击“下一步”。

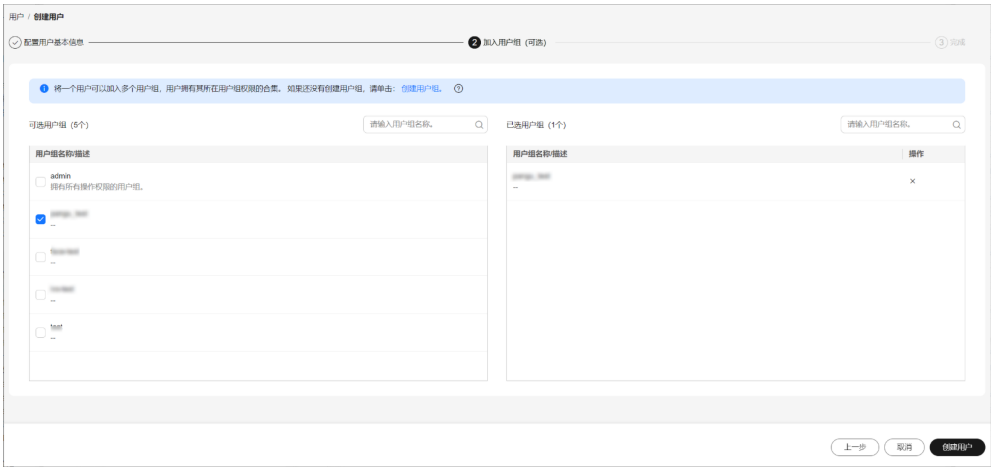
   配置用户信息时，需要勾选“编程访问”，如果未勾选此项，会导致IAM用户无法使用盘古服务API、SDK。

图 2-14 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 2-15 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. 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。

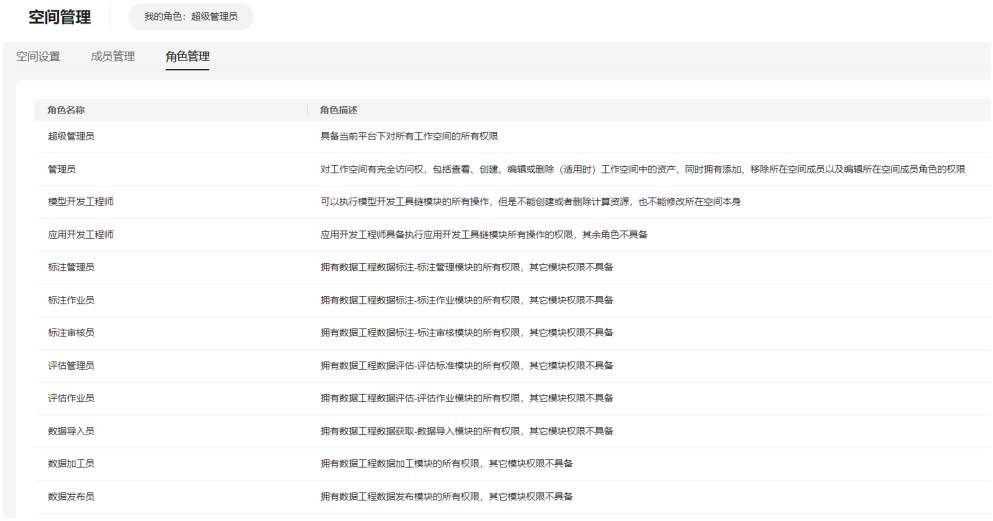
图 2-16 添加成员

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 2-17 角色管理



3. 单击进入“成员管理”页签。
4. 单击用户列表操作栏的“编辑”。
5. 勾选需要赋予用户的角色，单击“确认”。

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 2-18 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 2-19 删除操作二次确认

3 使用数据工程准备与处理数据集

3.1 数据工程介绍

数据工程简介

数据工程是ModelArts Studio大模型开发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、标注、评估和发布等过程，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。

数据工程所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，涉及从不同来源和格式的数据导入到平台。ModelArts Studio大模型开发平台提供多种高效灵活的数据接入方式，支持本地上传、通过OBS服务将数据导入平台。平台支持的多种数据类型包括文本、图片、视频等，能够满足不同行业和业务需求的多样化数据接入方式。用户还可以根据业务需求上传自定义格式的数据，极大地提升了数据获取的灵活性和可扩展性。通过这一功能，用户能够方便快捷地将大量数据导入平台，为后续的数据处理和模型训练打下良好的基础。
- **数据加工**：数据加工是确保数据质量的关键步骤。平台提供一系列数据清洗、过滤、转换等加工操作，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。针对不同类型的数据集，平台设计了专用的加工算子（即为特定数据处理任务预定义的操作模块，如文本去重、格式转换、异常处理等），通过这些算子能够高效地处理各类数据。对于文本类数据集，平台还支持用户自定义加工算子，以进一步满足特定场景下的需求。目前这一自定义算子功能仅适用于文本类数据集。通过加工操作，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。
- **数据标注**：在大模型的训练中，数据标注至关重要。平台不仅支持对无标签数据进行手动标注或重新标注，还支持对图片、视频类数据集通过AI预标注技术提升标注效率。AI预标注功能通过自动化的方式为数据集生成初步的标签，用户可以在此基础上进行人工审核和修正，从而大幅度减少人工标注的工作量和时间成本。此外，AI预标注不仅提高了标注效率，还能减少人为错误，提高标注的一致性和准确性。标注质量的提高直接增强了训练数据的有效性，确保训练模型时能获得更高质量的学习数据，从而推动模型性能的提升。
- **数据评估**：数据的质量直接决定了大模型的表现，因此，数据质量评估在整个数据工程中占有重要地位。ModelArts Studio大模型开发平台提供了强大的数据质

量评估工具，能够对处理后的数据集进行深入分析，评估其准确性、完整性和一致性。平台生成详细的数据质量评估报告，帮助用户全面了解数据的健康状况。数据评估结果能够为后续的数据优化提供明确指导，帮助用户在数据发布前进行最后的质量把关，确保数据集的可靠性，为大模型的训练提供高质量的基础数据。

- **数据发布**：数据发布是数据工程流程的最后一步。平台支持将经过加工、标注和评估的数据集以多种格式进行发布，包括默认格式、盘古格式（适用于训练盘古大模型时）。这些格式支持用户在不同的AI平台和业务场景中使用，确保数据在不同模型训练系统中的兼容性与流畅使用。目前，发布多种数据集格式的功能仅支持文本类和图片类数据集。

数据工程架构图如下：

图 3-1 数据工程架构图

通过集成数据获取、加工、标注、评估和发布的完整流程，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**。

表 3-1 平台支持的数据类型

数据类型	数据内容	数据文件格式要求
文本类	文档	支持txt、mobi、epub、docx、pdf，详见**文本类数据集格式要求**。
	网页	支持html，详见**文本类数据集格式要求**。
	预训练文本	支持jsonl，详见**文本类数据集格式要求**。
	单轮问答	支持jsonl、csv，详见**文本类数据集格式要求**。
	单轮问答（人设）	支持jsonl、csv，详见**文本类数据集格式要求**。

数据类型	数据内容	数据文件格式要求
	多轮问答	支持jsonl，详见**文本类数据集格式要求**。
	多轮问答（人设）	支持jsonl，详见**文本类数据集格式要求**。
	问答排序	支持jsonl、csv，详见**文本类数据集格式要求**。
图片类	图片	支持图片、tar，详见**图片类数据集格式要求**。
	图片+Caption	图片支持tar，Caption支持jsonl，详见**图片类数据集格式要求**。
	图片+QA对	图片支持tar，QA对支持jsonl，详见**图片类数据集格式要求**。
视频类	视频	支持mp4、avi，详见**视频类数据集格式要求**。
气象类	海洋气象	支持nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2，详见**气象类数据集格式要求**。
预测类	时序	支持csv，详见**预测类数据集格式要求**。
	回归分类	支持csv，详见**预测类数据集格式要求**。
其他类	用户自定义	支持构建CV场景中包含图片和标注文件的图像分类数据集，如图片+CV标注、视频+CV标注等类型，详见**其他类数据集格式要求**。

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	数据标注	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，数据的采集、清洗、标注、评估、发布等环节，成为数据开发中不可或缺的重要步骤。

在ModelArts Studio开发平台中，数据工程功能提供了完整的解决方案，用于高效构建和管理数据集，其操作流程见图3-2、表3-3。这种全面的数据准备机制，确保了数据质量的可靠性，为各类模型开发奠定了坚实的基础。

图 3-2 数据集准备与处理流程图

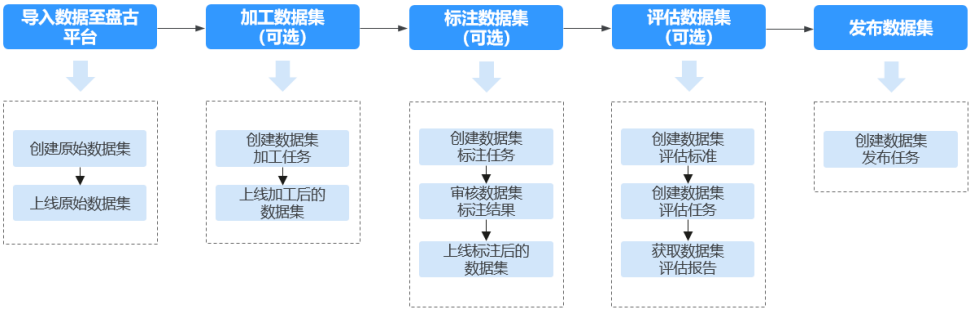


表 3-3 数据集准备与处理流程表

流程	子流程	说明
导入数据至盘古平台	创建原始数据集	数据集是指用于模型训练或评测的一组相关数据样本，上传至平台的数据将被创建为原始数据集进行统一管理。
	上线原始数据集	在正式发布数据集前，需要执行上线操作。
加工数据集（可选）	创建数据集加工任务	当数据集中存在异常数据、噪声数据、或不符合分析需求的数据时，可以通过加工数据集进行处理，包括但不限于数据提取、过滤、转换、打标签等操作。
	上线加工后的数据集	对加工后的数据集执行上线操作。
标注数据集（可选）	创建数据集标注任务	创建数据集标注任务，并对数据集执行标注操作，标注后的数据可以用于模型训练。
	审核数据集标注结果	对数据集的标注结果进行审核。
	上线标注后的数据集	对标注后的数据集执行上线操作。
评估数据集（可选）	创建数据集评估标准	创建数据集评估标准。可以评估文本通顺性、图文内容一致性、视频清晰度等。

流程	子流程	说明
	创建数据集评估任务	创建数据集质量评估任务，并基于评估标注对数据逐一评估其质量，评估后的数据可以用于模型训练。
	获取数据集评估报告	查看数据集评估任务的进展和数据集质量。
发布数据集	创建数据集发布任务	创建数据集发布任务，并进行正式的数据集发布操作，可用于后续的训练任务。 平台支持发布的数据集格式为**默认格式、盘古格式、自定义格式**，可按需进行数据集格式转换。 ● 默认格式：平台默认的格式。 ● 盘古格式：训练盘古大模型时，需要进行数据集格式转换。当前仅文本类、图片类数据集支持转换为盘古格式。 ● 自定义格式：文本类数据集可以使用自定义脚本进行数据格式转换。

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	数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
网页	html	数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
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。"} ● 数据集最大100万个文件，单文件最大2GB，整个数据集最大1.5TB。

文件内容	文件格式	文件要求
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" • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
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，具体格式示例如下： [{"context":"你好","target":"你好，请问有什么可以帮助你"},{"context":"请介绍一下盘古大模型","target":"盘古大模型，是华为推出盘古系列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。"}] • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
问答排序	jsonl、csv	• jsonl格式：context表示问题，targets的回答1、回答2、回答3表示答案的优劣顺序，最好的答案排在最前面。targets内容的数量至少为2个，且最多为6个，具体格式示例如下： { "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，具体格式示例如下： "问题","回答1","回答2","回答3" • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
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，具体格式示例如下： {"system":"机智幽默","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，具体格式示例如下： {"机智幽默","你好，请介绍自己","哈哈，你好呀，我是你的聪明助手。"} • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。

文件内容	文件格式	文件要求
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，具体格式示例如下： [{"system":"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》。"}] • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。

3.3.2 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，创建时支持导入**mp4或avi**格式文件，同一文件夹下mp4或avi格式的所有视频文件会被同时上传导入，具体格式要求详见**表3-5**。

表 3-5 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。 • 数据集最大1000万个文件，单文件最大100GB，整个数据集最大100TB。

3.3.3 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**图片**、**图片+Caption**、**图片+QA对**三种类型的数据，具体格式要求详见**表3-6**。

表 3-6 图片类数据集格式要求

文件内容	文件格式	文件要求
图片	tar、图片目录	• 图片：支持jpg、jpeg、png、bmp类型，单张图片大小不能超过5M，图片总大小不能超过500MB。 • tar：tar包内图片支持jpg、jpeg、png、bmp图片类型，每个tar包不超过500MB。 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。

文件内容	文件格式	文件要求
图片+Caption	图片支持tar，Caption支持jsonl	• 图片+Caption指的是一张图片和与之相关的文字描述，Caption是对图片内容的简短说明或解释，帮助人们理解图片所表达的信息。 • 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）。 • Caption：jsonl格式，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，具体格式示例如下： {"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"} • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB，具体格式示例如下： 

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+QA对是指将一张图片和与之相关的问题及答案配对在一起，用于训练模型让其能够理解图片内容并回答与图片相关的问题。 • 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）。 • QA对：jsonl格式，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，具体格式示例如下： {"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"}]} • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB，具体格式示例如下： 

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3-7**。

表 3-7 气象类数据集格式要求

文件内容	文件格式	文件要求
海洋气象	nc、cdf、netcdf、gr、gr1、grb、grib、grb1、grib1、gr2、grb2、grib2	• 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。 • 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）。

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，每个数据点都有一个时间戳，表示数据在时间上的位置。它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。数据集中的多个特征变量帮助预测目标变量，而目标变量为连续数值，非离散类别。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3-8**。

表 3-8 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字眼，如train01.csv；验证数据名称需包含eval字眼；测试数据名称需包含test字眼。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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
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字眼，如train01.csv；验证数据名称需包含eval字眼；测试数据名称需包含test字眼。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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。

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，用户训练模型时如果使用较特殊的数据集，ModelArts Studio大模型开发平台支持导入用户自定义的数据集。

例如，在训练CV类算法（如图片分类、图片分割、图片检测等任务）时，用户需使用“其他”类型的数据集。

其他类数据集可直接执行发布操作，但暂不支持数据加工、标注、评估等操作。

具体格式要求详见**表3-9**。

表 3-9 其他类数据集格式要求

文件内容	文件格式	文件要求
图片+CV标注	图片+分割标注（图片+xml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_2.jpg”，那么标注文件的文件名应为“IMG_2.xml”。具体示例如下： ├─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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。xml标注文件的详细说明请参见**图像分割数据集标注文件说明**。
	图片+分类标注（图片+txt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，标注文件txt中可以放单标签，也可以放多标签。具体示例如下： 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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。详细标注说明请参见**图像分类数据集标注文件说明**。

文件内容	文件格式	文件要求
	图片+二分类标注（图片+txt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，标注文件txt中可以放单标签，也可以放多标签。具体示例如下： 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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。详细标注说明请参见**图像分类数据集标注文件说明**。
	图片+检测标注（图片+xml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_2.jpg”，那么标注文件的文件名应为“IMG_2.xml”。具体示例如下： ├─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 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。xml标注文件的详细说明请参见**物体检测数据集标注文件说明**。
	图片+语义分割标注	• 训练数据为纯图片，要求为png格式。 • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。

文件内容	文件格式	文件要求
	图片+骨骼关键点坐标标注（图片+json）	• 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。具体示例如下： ├─annotations │ train.json │ val.json ├─train │ IMG_20180919_114745.jpg ├─val │ IMG_20180919_114945.jpg • 数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。json标注文件的详细说明请参见**骨骼关键点坐标标注json文件说明**。
视频+CV标注	视频+分类标注	• 数据源样本格式为.mp4格式，标注格式为.txt。每种类别的视频数需要大于50个，类别数量需要大于2，才能进行模型训练。数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。 • 用文本标签对视频文件进行标识，文本和视频放在同一目录下且同名。具体示例如下： dataset-import-example ├─import-dir-1 │ 10.mp4 │ 10.txt │ 11.mp4 │ 11.txt │ 12.mp4 │ 12.txt └─import-dir-2 1.mp4 1.txt 2.mp4 2.txt • 标签文件示例，如1.txt文件内容如下所示： Running

文件内容	文件格式	文件要求

	视频+事件起止时间与类别标注	● 数据源样本为.avi或.mp4格式，标注格式为.json。必须包含两个及以上后缀名字为avi或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。数据集最大100万个文件，单文件最大10GB，整个数据集最大10TB。
		● 支持视频的格式包括常见的mp4和或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。具体示例如下：
		├─file
		│ 1.mp4
		│ 2.avi
		│ annotation.json
		● 具体的json标注文件参考如下：
		{
		'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},
		...
		]
		},
		...
		}
		}

图像分割数据集标注文件说明

图像分割的数据支持格式为ModelArts image segmentation 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_114746.jpg”，那么标注文件的文件名应为“IMG_20180919_114746.xml”。

图像分割的标注文件基于PASCAL VOC格式增加了字段mask_source和mask_color，格式详细说明请参见**表3-10**。

具体示例如下：

```
├─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
```

标注文件的内容示例如下：

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
   <folder>NA</folder>
   <filename>image_0006.jpg</filename>
   <source>
      <database>Unknown</database>
   </source>
   <size>
      <width>230</width>
      <height>300</height>
      <depth>3</depth>
   </size>
   <segmented>1</segmented>
   <mask_source>obs://xianao/out/dataset-8153-Jmf5ylLjRmSacj9KevS/annotation/V001/
segmentationClassRaw/image_0006.png</mask_source>
   <object>
      <name>bike</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <mask_color>193,243,53</mask_color>
      <occluded>0</occluded>
      <polygon>
         <x1>71</x1>
         <y1>48</y1>
         <x2>75</x2>
         <y2>73</y2>
         <x3>49</x3>
         <y3>69</y3>
         <x4>68</x4>
         <y4>92</y4>
         <x5>90</x5>
         <y5>101</y5>
         <x6>45</x6>
         <y6>110</y6>
         <x7>71</x7>
         <y7>48</y7>
      </polygon>
   </object>
</annotation>
```


表 3-10 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示数据源所在目录。
filename	是	被标注文件的文件名。
size	是	表示图像的像素信息。 • width：必选字段，图像的宽度。 • height：必选字段，图像的高度。 • depth：必选字段，图像的通道数。 图像的通道数是指图像中每个像素的颜色信息的维度。常用的RGB图像默认有3个通道。3通道表示彩色图像，每个像素有三个值表示红、绿、蓝三个色彩通道的亮度。常用的还有1通道，表示灰度图像，每个像素只有一个值表示亮度或灰度级别。4通道表示带有透明度的彩色图像，每个像素有四个值表示红、绿、蓝三个色彩通道的亮度以及透明度。
segmented	是	表示是否用于分割。“是”取值为1，“否”取值为0。
mask_source	否	表示图像分割保存的mask路径。
object	是	表示物体检测信息，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，标注内容是否被截断（0表示完整）。 • occluded：必选字段，标注内容是否被遮挡（0表示未遮挡）。 • difficult：必选字段，标注目标是否难以识别（0表示容易识别）。 • confidence：可选字段，标注目标的置信度，取值范围0-1之间。 • bndbox：必选字段，标注框的类型，可选值请参见**表3-11**。 • mask_color：必选字段，标签的颜色，以RGB值表示。

表 3-11 标注框类型描述

type	形状	标注信息
polygon	多边形	各点坐标。 <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6> <x7>100<x7> <y7>100<y7>

图像分类数据集标注文件说明

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：

  ```
  Cat
  ```

  多标签的标签文件示例，如2.txt文件内容如下所示：

  ```
  Cat
  Dog
  ```


物体检测数据集标注文件说明

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3-12**。

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 3-12 PASCAL VOC 格式说明

字段	是否必选	说明
folder	是	表示图像所在的目录名称。
filename	是	被标注文件的文件名。
size	是	表示图像的像素信息。 • width：必选字段，图像的宽度。 • height：必选字段，图像的高度。 • depth：必选字段，图像的通道数。
segmented	是	表示是否用于分割，取值为0或1。0表示没有分割标注，1表示有分割标注。
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3-13。

表 3-13 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>

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

骨骼关键点坐标标注 json 文件说明

骨骼关键点坐标标注基于开源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 3-14 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。

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

字段	是否必选	说明
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。

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。存储在OBS服务中的数据或本地数据导入ModelArts Studio大模型开发平台后，将以数据集的形式进行统一管理。

用户将数据导入至平台后，这些数据会生成一个“原始数据集”，用于对导入的数据进行集中管理和进一步操作。

创建原始数据集

创建原始数据集前，请先按照**数据集格式要求**提前准备数据。如果需要使用OBS服务导入数据，请详见**通过控制台快速使用OBS**。

说明

在使用OBS服务上传数据时，如果遇到网络报错“NET::ERR_CERT_AUTHORITY_INVALID”，是由于域名未绑定有效的SSL证书，导致HTTPS请求被浏览器拦截。可以通过以下方法进行规避：

通过浏览器访问报错的URL链接，根据页面告警提示对链接进行安全认证。认证完成后，只要不清理浏览器缓存，对相同桶域名的访问都不会被拦截。

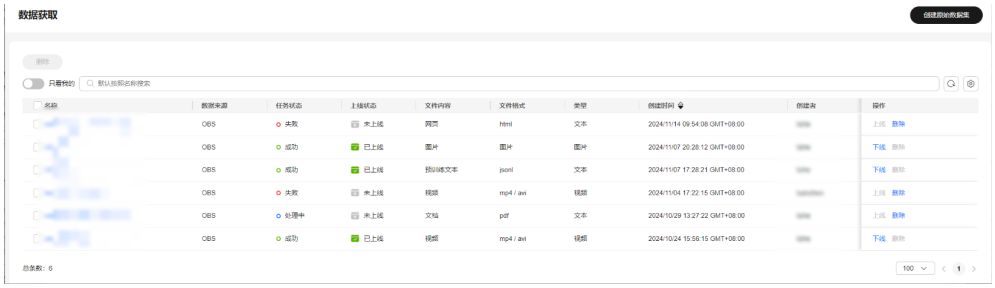
创建原始数据集步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入操作空间。

图 3-3 进入操作空间

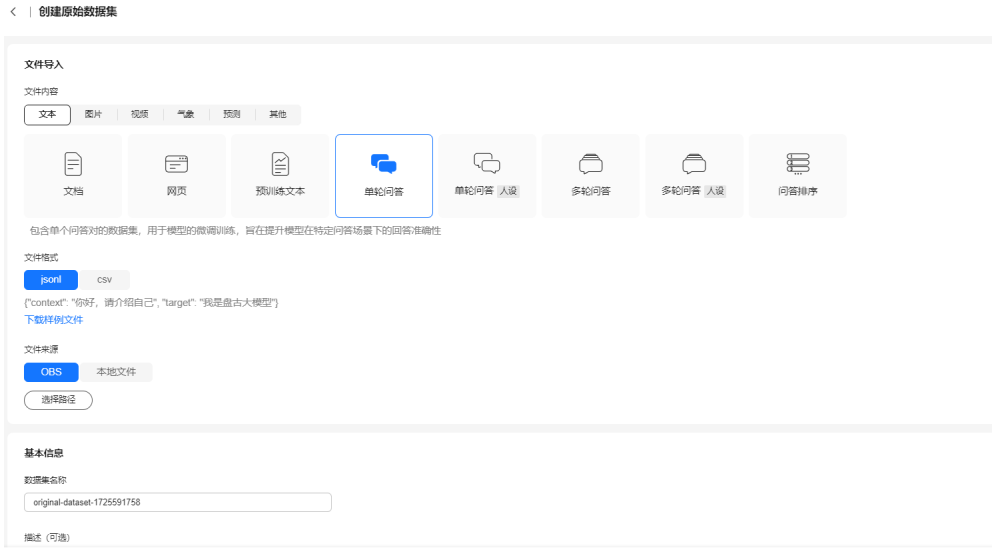
2. 在左侧导航栏中选择“数据工程 > 数据获取”，单击界面右上角“创建原始数据集”。

图 3-4 数据获取



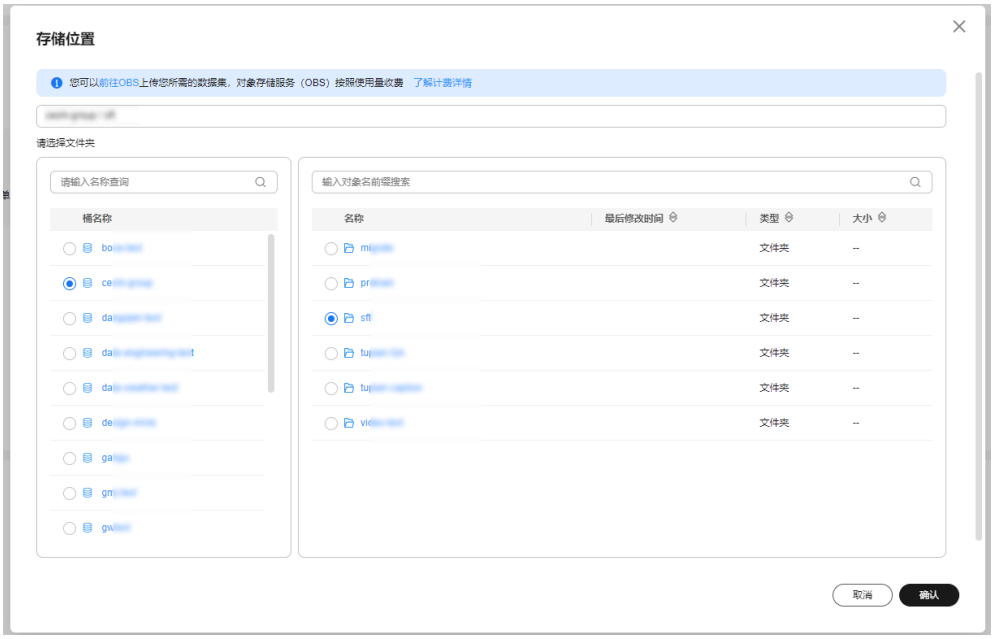
3. 在“创建原始数据集”页面，选择所需“文件内容”、“文件格式”和“文件来源”。

图 3-5 创建原始数据集



4. 单击“选择路径”，在“存储位置”弹窗中选择需导入的数据，单击“确认”。

图 3-6 选择导入的数据



5. 数据集信息设置完成后，填写“数据集名称”和“描述”，并设置“拓展信息”。

   拓展信息包括“标签设置”与“数据版权”：

   - 标签设置。通过标签设置，可以给数据集添加行业、语言、标签信息。
   - 数据版权设置。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。

   图 3-7 设置数据版权

6. 单击页面右下角“立即创建”完成原始数据集的创建操作。创建完成后，自动返回至“数据获取”页面，在该页面可以查看数据集的任务状态，还可执行上线、删除操作。

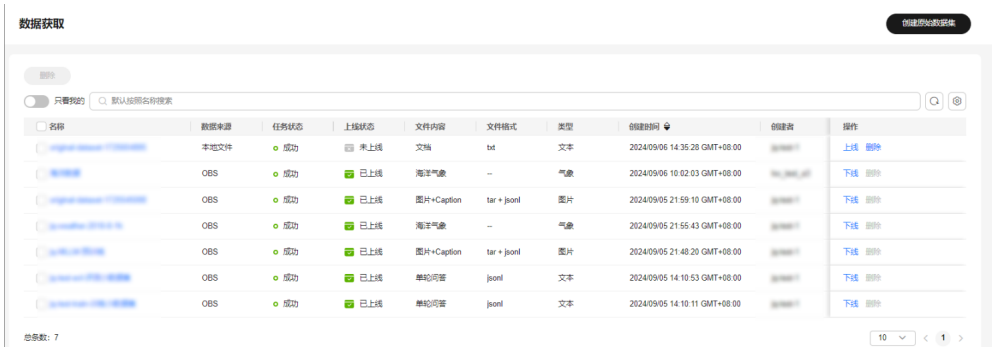
   如果任务状态为“失败”，可能由以下原因导致：

   - 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“创建原始数据集”页面下载数据样例进行比对。

上线原始数据集

原始数据集创建成功后，在“数据获取”页面的操作列单击“上线”，完成原始数据集上线。

图 3-8 上线数据集



说明

只有上线后的数据集才可用于后续的数据加工、标注、评估、发布操作。

管理原始数据集

原始数据集上线成功后，支持查看数据集详情、下载数据集、查看数据血缘、以及对数据集进行删除等操作。

- 支持查看数据集详情。在“数据获取”页面，单击数据集名称，在“基本信息”页签可查看当前数据集的创建人、创建时间等详细信息、行业标签等扩展信息以及该数据集的创建、导入、上线等操作记录。
- 下载数据文件。在“数据获取”页面，单击数据集名称，在“数据文件”页签，单击文件操作列的“下载”，可实现下载数据文件操作。
- 查看数据血缘。在“数据获取”页面，单击数据集名称，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。
- 删除原始数据集。已上线的数据集需先执行下线操作后才可以删除。在“数据获取”页面，单击数据集操作列的“下线”，单击“删除”并进行二次删除确认。

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

数据加工概念

数据加工是数据工程中的核心环节，旨在通过使用数据集加工算子对原始数据进行清洗、转换、提取和过滤等操作，以确保数据符合模型训练的标准和业务需求。

通过这一过程，用户能够优化数据质量，去除噪声和冗余信息，提升数据的准确性和一致性，为后续的模型训练提供更高质量、更有效的输入。数据加工不仅仅是对数据的简单处理，它还针对不同数据类型和业务场景进行有针对性的优化。

ModelArts Studio大模型开发平台提供了强大的数据加工功能，根据不同类型的数据集预置了多种加工算子，如数据提取、转换和过滤等。

数据加工意义

数据加工直接影响到模型训练的质量和效率。通过数据加工，可以确保训练数据具有较高的质量，减少由于数据问题导致的训练误差，从而提高模型的性能。

- **提升数据质量**：数据加工能够去除噪声、修复缺失值和异常值，保证数据的准确性、完整性和一致性，为模型训练提供高质量的输入数据。
- **提高处理效率**：平台预置的多种数据加工算子，帮助用户快速完成数据清洗、转换和处理，减少手动操作，提高数据处理的效率。
- **满足业务需求**：不同类型的数据需要不同的处理方式，平台根据文本、图片、视频、气象等数据类型提供专门的加工工具，满足各种复杂的业务需求。
- **增强模型性能**：通过合适的数据加工，可以提高数据的可用性，进而提升模型的训练效果，使其具备更高的精度和鲁棒性。

总体而言，数据加工不仅帮助用户提升数据处理效率，还通过优化数据质量，支持高效的模型训练，帮助用户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

当前支持加工操作的数据集类型如下：

- 文本类数据集，加工算子清单详见**文本类加工算子能力清单**。
- 视频类数据集，加工算子清单详见**视频类加工算子能力清单**。
- 图片类数据集，加工算子清单详见**表3-17**、**表3-18**。
- 气象类数据集，加工算子清单详见**表3-19**。

3.5.2 数据集加工算子介绍

3.5.2.1 文本类加工算子能力清单

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、数据过滤三类，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中提取内容转换为结构化数据。

算子分类	算子名称	算子描述
	JSON内容提取	从JSON文件（键值对类型文件）中提取出内容。
	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。
	智能文档解析	从PDF（支持扫描版）或图片中提取文本，转化为结构化数据，持文本、表格、表单、公式等内容提取。
数据转换	个人数据脱敏	对文本中的电话号码、邮箱、身份证、车牌号、IP地址、URL地址、MAC地址、护照号、IMEI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	将简体文本转换为繁体，或将繁体文本转换为简体。
	符号标准化	查找数据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"}

算子分类	算子名称	算子描述
	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	日期有数字+中文、全数字、全中文等形式，将不同种类的日期格式对齐到同种格式。
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<p>。 ● 特殊符号，比如● ■ ◆。 ● 乱码和无意义的字符□□□□□。
	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、机密和知识产权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，对长度范围内的数据进行保留。
	冗余信息过滤	查找文本中的冗余信息并替换为空值，不改变数据条目。例如目录封面、图注表注、标注说明、首尾部信息、冗余段落和参考文献等非正文内容。
	N-gram特征过滤	根据如下特征过滤： ● N gram重复率：以N个字符为粒度统计频率大于1的N-gram的个数与所有N-gram的个数比值。 ● Top N gram占比：频率最高N gram占比。

算子分类	算子名称	算子描述
	段落特征过滤	根据如下特征过滤： ● 段落重复率。 ● 段落非中文字符占比。 ● 段落完整性。
	句子特征过滤	根据如下特征过滤： ● 过滤平均句长小于阈值的文档。
	词语特征过滤	根据如下特征过滤： ● 词个数。 ● 平均词长度。
	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。
	段落结尾不完整句子过滤	删除文本中不完整段落和句子。
	广告数据过滤	删除文本中包含广告数据的句子。
	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。

3.5.2.2 视频类加工算子能力清单

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、打标签和评分等。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。

算子分类	算子名称	算子描述
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 图片类加工算子能力清单

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3-18**。

图文类加工算子能力清单

表 3-17 图文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。
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，比如U+0000-U+001F ● 表情符□□ ● 网页标签符号<p> ● 特殊符号，比如● ■ ◆ ● 乱码和无意义的字符□□□□□

图片类加工算子能力清单

表 3-18 图片类加工算子功能表

算子分类	算子名称	算子描述
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/图文数据清洗。
	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。评分范围（0，100），默认评分超过50分的视频可视为涉黄视频。

3.5.2.4 气象类加工算子能力清单

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3-19。

表 3-19 气象类加工算子能力清单

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

3.5.3 加工文本类数据集

3.5.3.1 创建文本类数据集加工任务

创建文本类数据集加工任务前，请先完成“原始数据集”的创建与上线，具体步骤请参见导入数据至盘古平台。

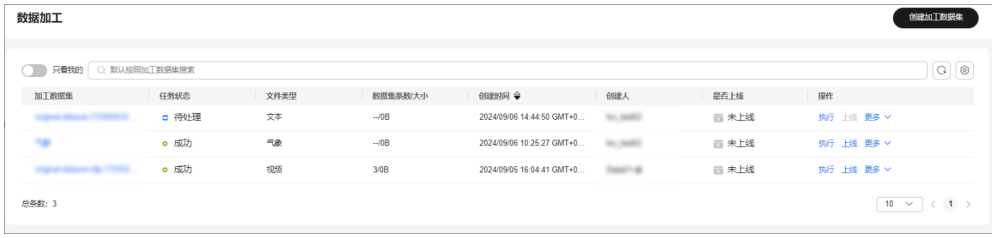
创建文本类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-9 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，单击界面右上角“创建加工数据集”。

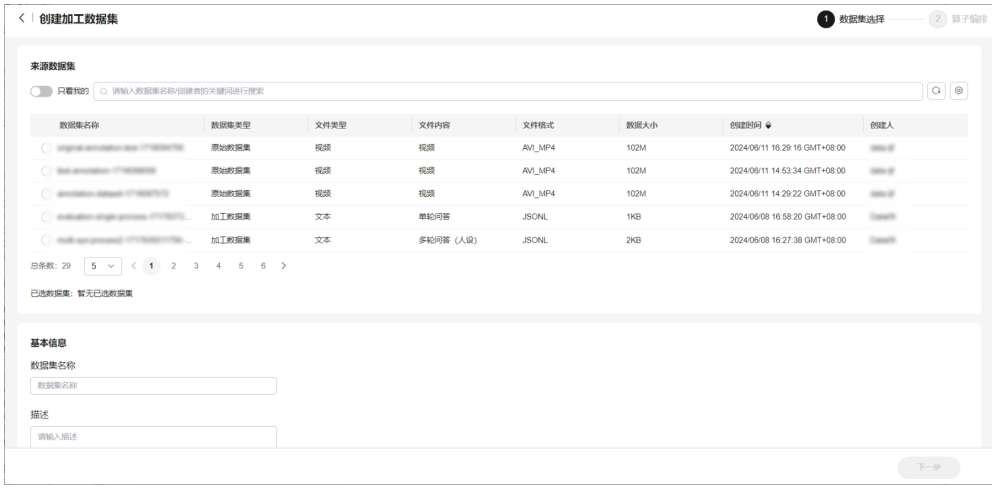
图 3-10 数据加工



3. 在“创建加工数据集”页面，选择需要加工的文本类数据集，并设置数据集的名称和描述。

   选择数据集时，默认选择当前空间的数据集。如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。

图 3-11 创建加工数据集



4. 单击“下一步”进入“算子编排”页面。对于文本类数据集，可选择预置加工算子，请参见**文本类加工算子能力清单**。

   a. 在左侧“添加算子”模块勾选所需算子。

   b. 在右侧“加工步骤编排”页面配置各算子的参数，可通过右侧 ⠿ 按钮，拖拽算子的上下顺序来调整算子在加工任务流中的执行顺序。

   c. 算子编排过程中，可以单击右上角“保存为新模板”将当前算子编排流程保存为模板，后续创建新的数据加工任务时，可以直接单击“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 3-12 算子编排

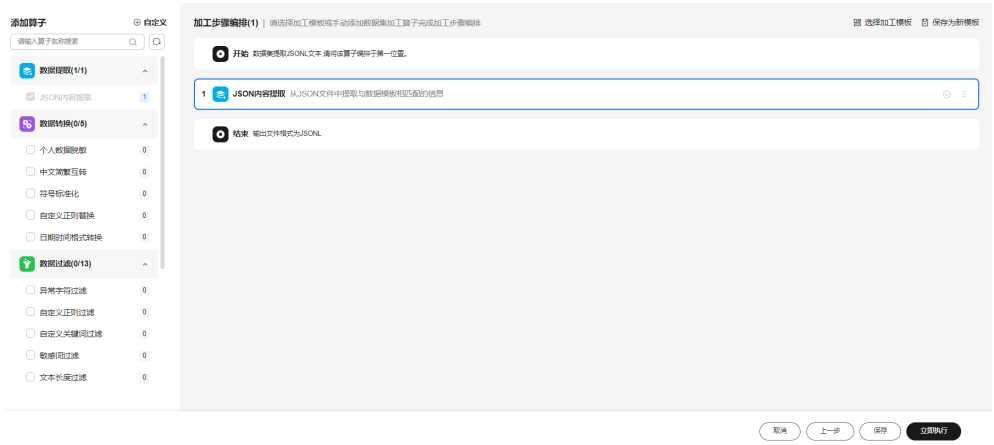


图 3-13 选择加工模板

5. 算子编排完成后，单击“立即执行”，平台会直接启动数据加工任务。若单击“保存”，数据集列表页中将新增一个任务状态为“待处理”的数据加工任务，可单击操作列“执行”启动加工。

图 3-14 数据加工

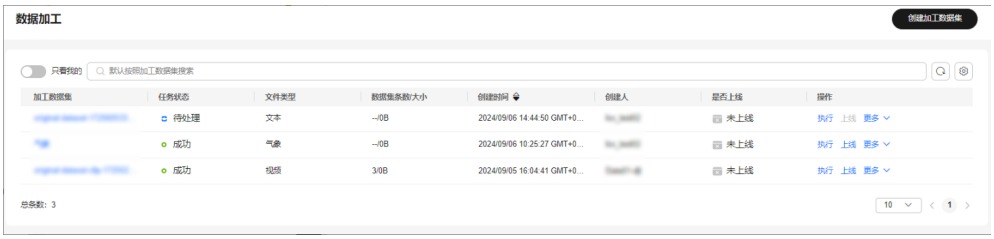
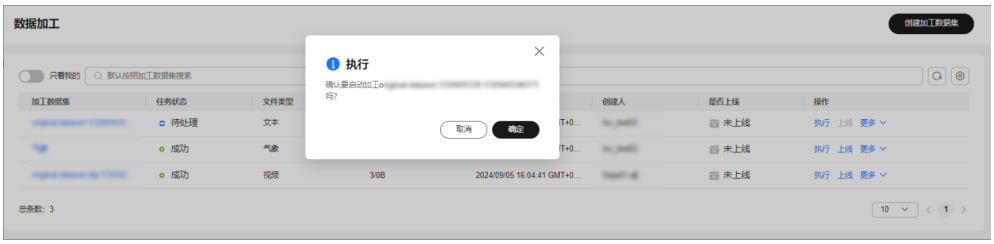


图 3-15 执行加工



6. 当加工数据集任务运行成功后，状态将从“处理中”变为“成功”，表示数据已经完成加工，加工完成的数据集支持上线、编辑与删除操作。
7. 平台支持查看加工后的数据集。单击加工完成的数据集名称，在“数据文件”页签的文件操作列单击“下载”，再单击“确定”，下载完成后即可查看。

3.5.3.2 上线加工后的文本类数据集

加工后的文本类数据集需要执行上线操作，用于后续的数据标注、评估、发布任务，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-16 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，在数据集操作列单击“上线”，执行上线操作。
3. 单击数据集名称查看加工任务的基本信息、加工详情、加工后的数据文件以及数据血缘。
   - 在“基本信息”页签可查看数据集的详细信息及操作概览。
   - 在“加工详情”页签可以查看数据集的加工步骤和运行日志。
   - 在“数据文件”页签可下载加工后的数据文件，可以与原始数据进行比对，查看加工前后的差异。
   - 在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

说明

- 上线后的加工数据集不支持编辑和删除操作。若执行该操作，需将数据集下线。
- 若上线后的加工数据集已执行发布操作**发布数据集**，则不可将该加工数据集下线。

3.5.4 加工视频类数据集

3.5.4.1 创建视频类数据集加工任务

创建视频类数据集加工任务前，请先完成“原始数据集”的创建与上线，具体步骤请参见**导入数据至盘古平台**。

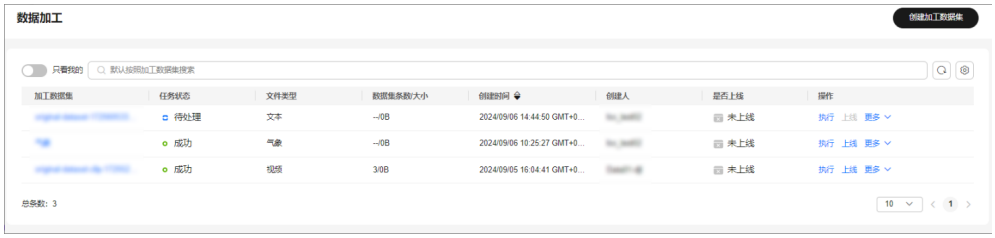
创建视频类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-17 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，单击界面右上角“创建加工数据集”。

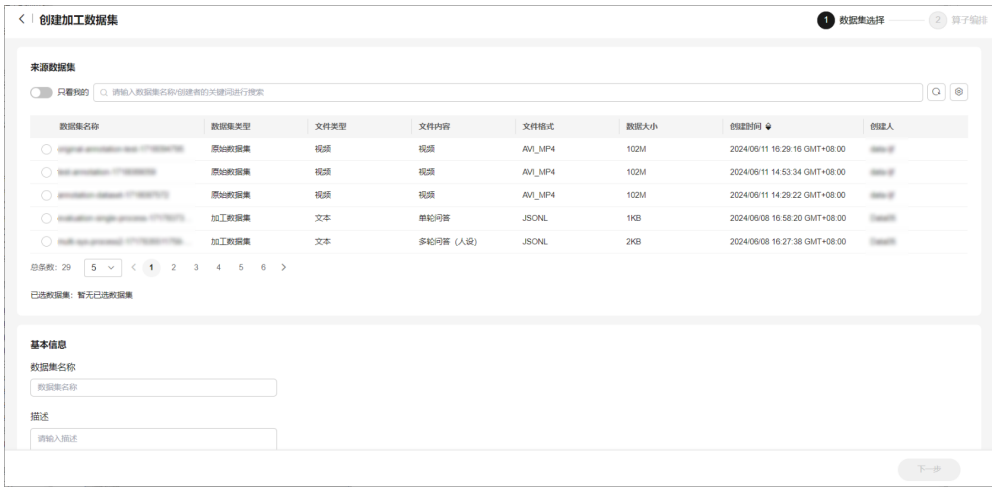
图 3-18 数据加工



3. 在“创建加工数据集”页面，选择需要加工的视频类数据集，并设置数据集的名称和描述信息。

   选择数据集时，默认选择当前空间的数据集。如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。

图 3-19 创建加工数据集



4. 单击“下一步”进入“算子编排”页面。对于视频类数据集，可选择的加工算子及参数配置请参见**表3-16**。

   a. 在左侧“添加算子”模块勾选所需算子。

   b. 在右侧“加工步骤编排”页面配置各算子的参数，可通过右侧 ⠿ 按钮，拖拽算子的上下顺序来调整算子在加工任务流中的执行顺序。

   c. 算子编排过程中，可以单击右上角“保存为新模板”将当前算子编排流程保存为模板，后续创建新的数据加工任务时，可以直接单击“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 3-20 算子编排

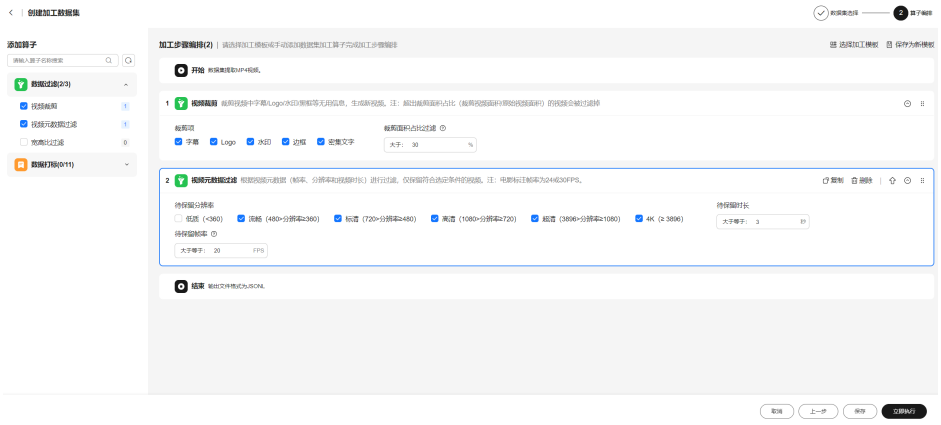


图 3-21 选择加工模板

5. 算子编排完成后，单击“立即执行”，平台会直接启动数据加工任务。若单击“保存”，数据集列表页中将新增一个任务状态为“待处理”的数据加工任务，可单击操作列“执行”启动加工。

图 3-22 数据加工

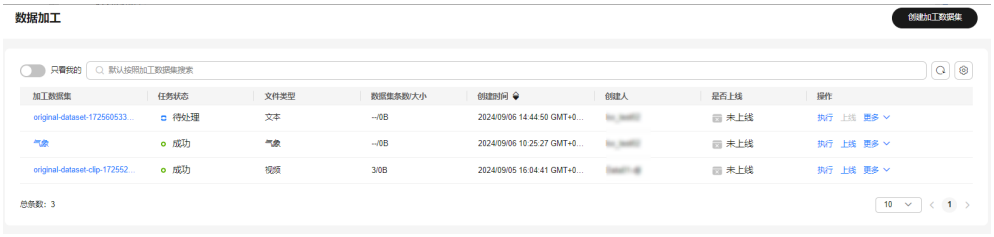
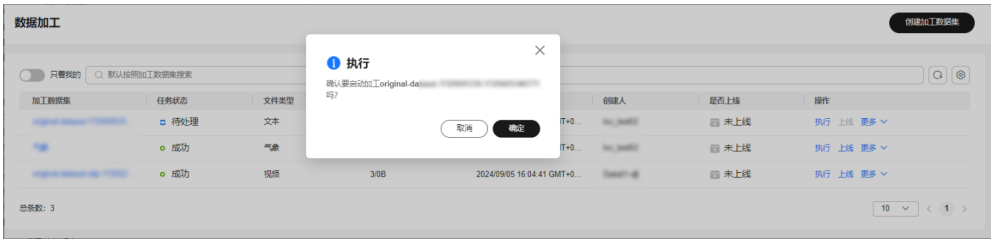


图 3-23 执行加工



6. 当加工数据集任务运行成功后，状态将从“处理中”变为“成功”，表示数据已经完成加工，加工完成的数据集支持上线、编辑与删除操作。
7. 平台支持查看加工后的数据集。单击加工完成的数据集名称，在“数据文件”页签的文件操作列单击“下载”，再单击“确定”，下载完成后即可查看。

3.5.4.2 上线加工后的视频类数据集

加工后的视频类数据集需要执行上线操作，用于后续的数据标注、评估、发布任务，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-24 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，在数据集操作列单击“上线”，执行上线操作。

3. 单击数据集名称查看加工任务的基本信息、加工详情、加工后的数据文件以及数据血缘。
   - 在“基本信息”页签可查看数据集的详细信息及操作概览。
   - 在“加工详情”页签可以查看数据集的加工步骤和运行日志。
   - 在“数据文件”页签可下载加工后的数据文件，可以与原始数据进行比对，查看加工前后的差异。
   - 在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

说明

- 上线后的加工数据集不支持编辑和删除操作。若执行该操作，需将数据集下线。
- 若上线后的加工数据集已执行发布操作**发布数据集**，则不可将该加工数据集下线。

3.5.5 加工图片类数据集

3.5.5.1 创建图片类数据集加工任务

创建图片类数据集加工任务前，请先完成“原始数据集”的创建与上线，具体步骤请参见**导入数据至盘古平台**。

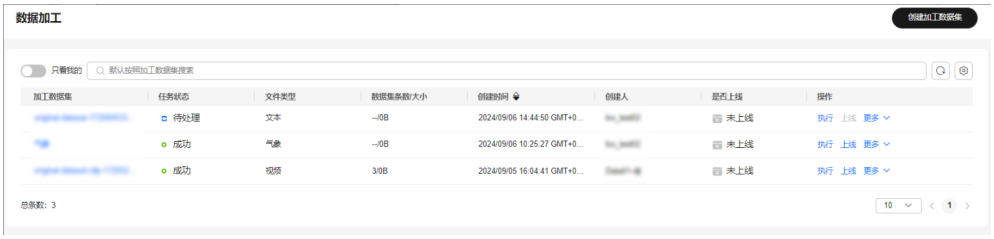
创建图片类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-25 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，单击界面右上角“创建加工数据集”。

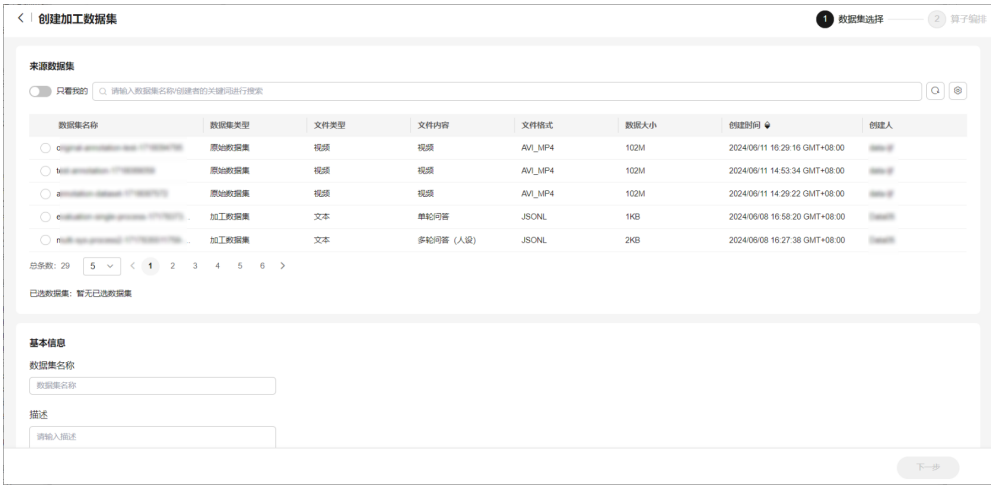
图 3-26 数据加工



3. 在“创建加工数据集”页面，选择需要加工的图片类数据集，并设置数据集的名称和描述信息。

   选择数据集时，默认选择当前空间的数据集。如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。

图 3-27 创建加工数据集



4. 单击“下一步”进入“算子编排”页面。对于图片类数据集，可选择的加工算子及参数配置请参见**表3-17**、**表3-18**。

   a. 在左侧“添加算子”模块勾选所需算子。

   b. 在右侧“加工步骤编排”页面配置各算子的参数，可通过右侧 ⠿ 按钮，拖拽算子的上下顺序来调整算子在加工任务流中的执行顺序。

   c. 算子编排过程中，可以单击右上角“保存为新模板”将当前算子编排流程保存为模板，后续创建新的数据加工任务时，可以直接单击“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 3-28 算子编排

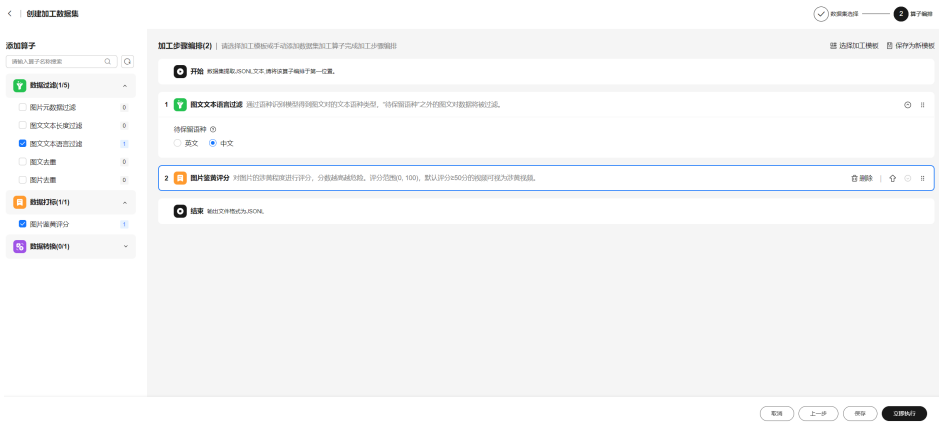


图 3-29 选择加工模板

5. 算子编排完成后，单击“立即执行”，平台会直接启动数据加工任务。若单击“保存”，数据集列表页中将新增一个任务状态为“待处理”的数据加工任务，可单击操作列“执行”启动加工。

图 3-30 数据加工

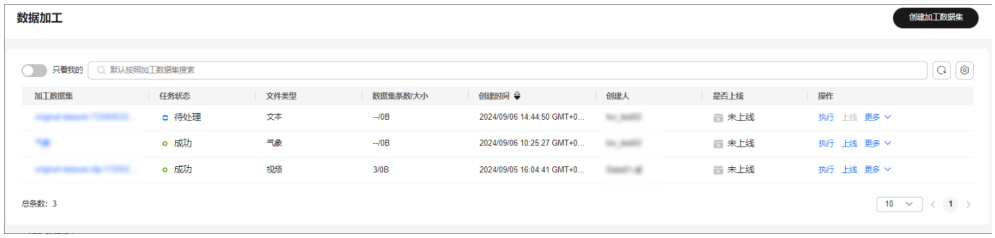
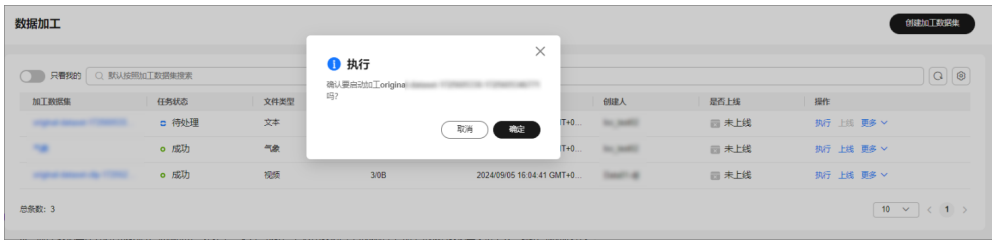


图 3-31 执行加工



6. 当加工数据集任务运行成功后，状态将从“处理中”变为“成功”，表示数据已经完成加工，加工完成的数据集支持上线、编辑与删除操作。
7. 平台支持查看加工后的数据集。单击加工完成的数据集名称，在“数据文件”页签的文件操作列单击“下载”，再单击“确定”，下载完成后即可查看。

3.5.5.2 上线加工后的图片类数据集

加工后的图片类数据集需要执行上线操作，用于后续的数据标注、评估、发布任务，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-32 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，在数据集操作列单击“上线”，执行上线操作。
3. 单击数据集名称查看加工任务的基本信息、加工详情、加工后的数据文件以及数据血缘。
   - 在“基本信息”页签可查看数据集的详细信息及操作概览。
   - 在“加工详情”页签可以查看数据集的加工步骤和运行日志。
   - 在“数据文件”页签可下载加工后的数据文件，可以与原始数据进行比对，查看加工前后的差异。
   - 在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

说明

- 上线后的加工数据集不支持编辑和删除操作。若执行该操作，需将数据集下线。
- 若上线后的加工数据集已执行发布操作**发布数据集**，则不可将该加工数据集下线。

3.5.6 加工气象类数据集

3.5.6.1 创建气象类数据集加工任务

创建气象类数据集加工任务前，请先完成“原始数据集”的创建与上线，具体步骤请参见**导入数据至盘古平台**。

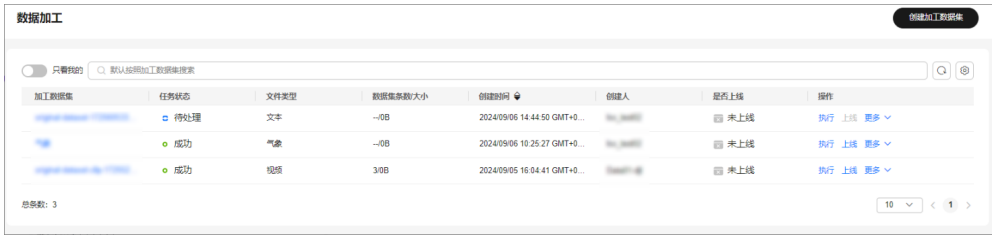
创建气象类数据集加工任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-33 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，单击界面右上角“创建加工数据集”。

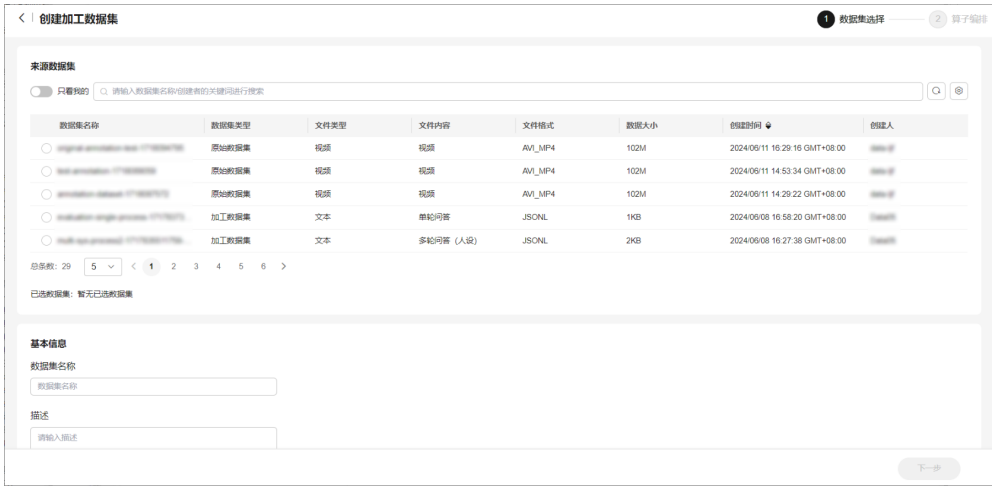
图 3-34 数据加工



3. 在“创建加工数据集”页面，选择需要加工的气象类数据集，并设置数据集的名称和描述信息。

   选择数据集时，默认选择当前空间的数据集。如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。

图 3-35 创建加工数据集



4. 单击“下一步”进入“算子编排”页面。对于气象类数据集，可选择的加工算子及参数配置请参见**表3-19**。

   a. 在左侧“添加算子”模块勾选所需算子。

   b. 在右侧“加工步骤编排”页面配置各算子的参数，可通过右侧 ⠿ 按钮，拖拽算子的上下顺序来调整算子在加工任务流中的执行顺序。

   c. 算子编排过程中，可以单击右上角“保存为新模板”将当前算子编排流程保存为模板，后续创建新的数据加工任务时，可以直接单击“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 3-36 算子编排

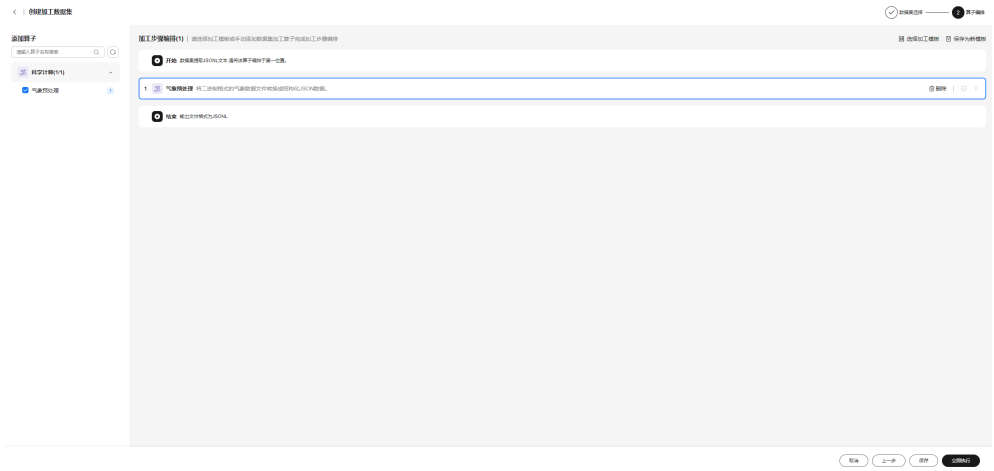


图 3-37 选择加工模板

5. 算子编排完成后，单击“立即执行”，平台会直接启动数据加工任务。若单击“保存”，数据集列表页中将新增一个任务状态为“待处理”的数据加工任务，可单击操作列“执行”启动加工。

图 3-38 数据加工

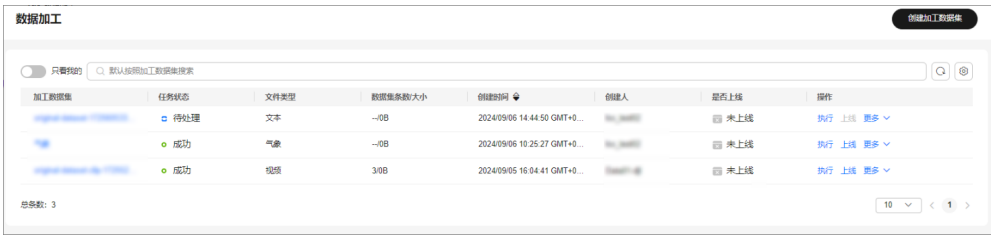
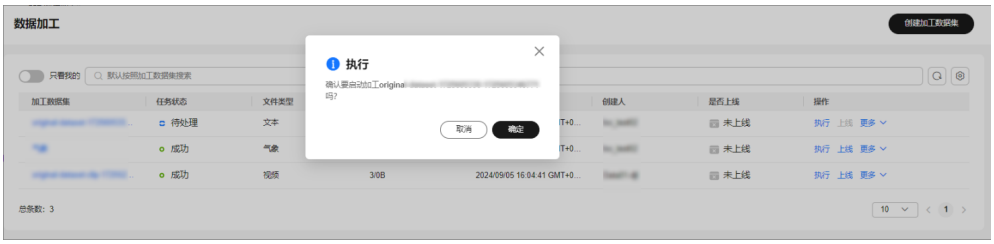


图 3-39 执行加工



6. 当加工数据集任务运行成功后，状态将从“处理中”变为“成功”，表示数据已经完成加工，加工完成的数据集支持上线、编辑与删除操作。
7. 平台支持查看加工后的数据集。单击加工完成的数据集名称，在“数据文件”页签的文件操作列单击“下载”，再单击“确定”，下载完成后即可查看。

3.5.6.2 上线加工后的气象类数据集

加工后的气象类数据集需要执行上线操作，用于后续的数据发布操作，具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-40 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据加工”，在数据集操作列单击“上线”，执行上线操作。
3. 单击数据集名称查看加工任务的基本信息、加工详情、加工后的数据文件以及数据血缘。
   - 在“基本信息”页签可查看数据集的详细信息及操作概览。
   - 在“加工详情”页签可以查看数据集的加工步骤和运行日志。
   - 在“数据文件”页签可下载加工后的数据文件，可以与原始数据进行比对，查看加工前后的差异。
   - 在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

说明

- 上线后的加工数据集不支持编辑和删除操作。若执行该操作，需将数据集下线。
- 若上线后的加工数据集已执行发布操作**发布数据集**，则不可将该加工数据集下线。

3.6 标注数据集

3.6.1 数据集标注场景介绍

数据标注概念

数据标注是数据工程中的关键步骤，旨在为无标签的数据集添加准确的标签，从而为模型训练提供有效的监督信号。标注数据的质量直接影响模型的训练效果和精度，因此高效、准确的标注过程至关重要。数据标注不仅仅是人工输入，它还涉及对数据内容的理解和分类，以确保标签精准地反映数据的特征和用途。

为了帮助用户高效、准确地完成数据标注任务，ModelArts Studio大模型开发平台提供了标注审核功能（即对标注后的数据集进行审核），确保标注结果经过验证和质量控制，提升数据的可靠性和可用性。同时，平台支持对视频类和图片类数据集进行AI预标注，标注员可以在此基础上进行审核和修正，从而有效减少人工标注的工作量，并保证原始数据集内容的完整性。

通过这些功能，平台不仅降低了标注成本，还为用户提供了灵活的定制化服务，满足不同业务场景的标注需求，确保为后续模型训练和优化提供高质量的数据支持。

数据标注意义

数据标注在数据工程中的作用是不可忽视的。它不仅是模型训练的基础，还直接影响到训练结果的准确性与有效性。通过标注，平台帮助用户提高数据的可用性，确保数据集与业务需求高度契合。数据标注的意义主要体现在以下几个方面：

- **提升训练数据的质量**：通过高质量的标注，用户能够获得准确、可靠的标签数据，为后续模型训练提供更有价值的输入数据，提升训练模型的准确性和表现。
- **满足不同业务需求**：ModelArts Studio大模型开发平台支持不同类型的数据标注，包括文本、图片、视频等，可以针对不同的数据和业务场景提供定制化的标注方案，满足多样化的需求。
- **增强模型的准确性与鲁棒性**：准确的标注数据能够帮助模型更好地学习数据的潜在模式和规律，进而提高模型的性能、准确性和鲁棒性。
- **节省时间与成本**：AI预标注可以显著减少人工干预，提高标注的效率和一致性，帮助用户节省标注成本和时间，尤其是在大规模数据集的处理过程中。

总的来说，数据标注是数据工程中不可或缺的一环，通过高效、准确的标注过程，ModelArts Studio大模型开发平台为用户提供了灵活、定制化的解决方案，确保数据质量，助力后续模型训练和优化，推动AI技术的成功应用。

支持数据标注的数据集类型

ModelArts Studio大模型开发平台支持标注操作的数据集类型如下：

- 文本类数据集，详见**创建文本类数据集标注任务**。
- 视频类数据集，详见**创建视频类数据集标注任务**。
- 图片类数据集，详见**创建图片类数据集标注任务**。

3.6.2 标注文本类数据集

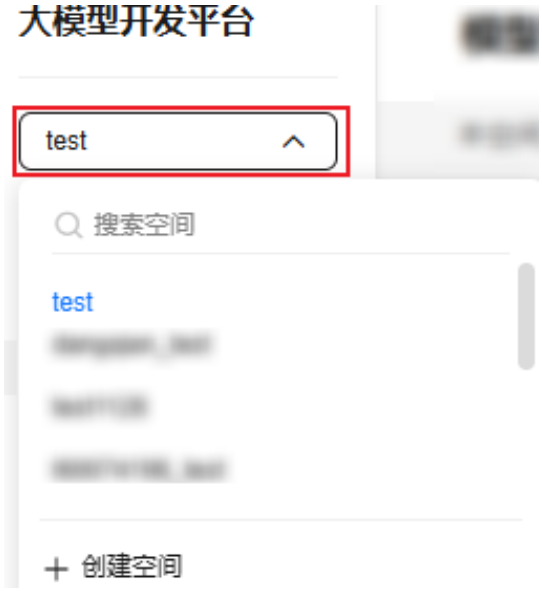
3.6.2.1 创建文本类数据集标注任务

创建文本类数据集标注任务前，请先完成**创建文本类数据集加工任务**。

创建文本类数据集标注任务步骤如下：

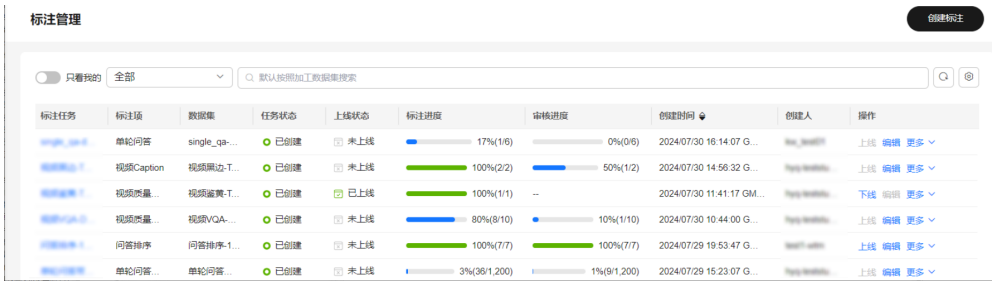
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-41 进入操作空间



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

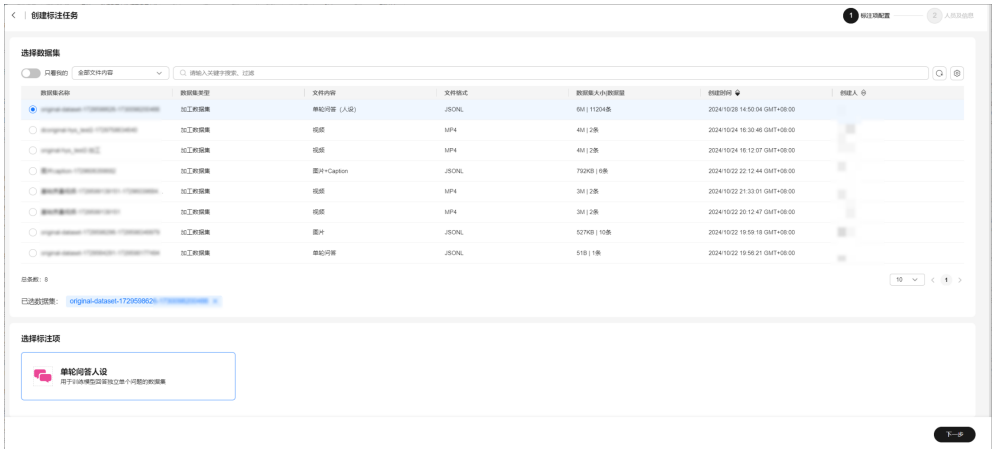
图 3-42 标注管理



3. 在“创建标注任务”页面选择需要标注的加工后的文本类数据集，并设置标注项。

   设置标注项时，不同类型的数据文件对应的标注项也有所差异，可基于页面提示进行设置。

图 3-43 创建标注任务



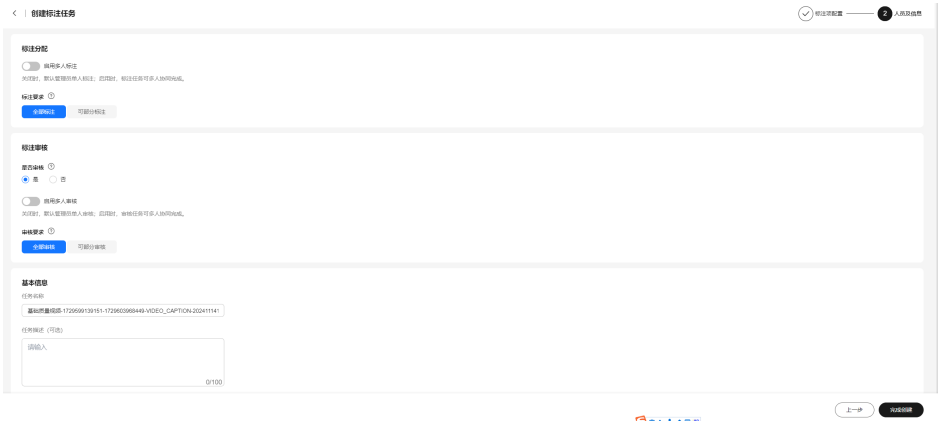
4. 单击“下一步”设置标注人员及信息，单击“完成创建”。

   分配标注任务时，可以选择是否启用多人标注。启用多人标注后，可以指定参与标注的人员。

   标注任务可选择是否启用标注审核，可设置多人审核，详见**审核文本类数据集标注结果**。审核要求可以选择以下两种方式：

   - 选择“可部分审核”：审核人员确认部分数据达到标注要求后，可以一键通过所有的标注。
   - 选择“全部审核”：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。

图 3-44 设置标注人员、标注信息示例



5. 在“标注管理”页面，单击操作列“上线”，可执行后续标注操作。对于未上线的标注任务，可执行编辑和删除操作。
6. 在“标注作业”页面，单击操作列“标注”可进行数据标注。如果需要将该标注任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-45 标注作业

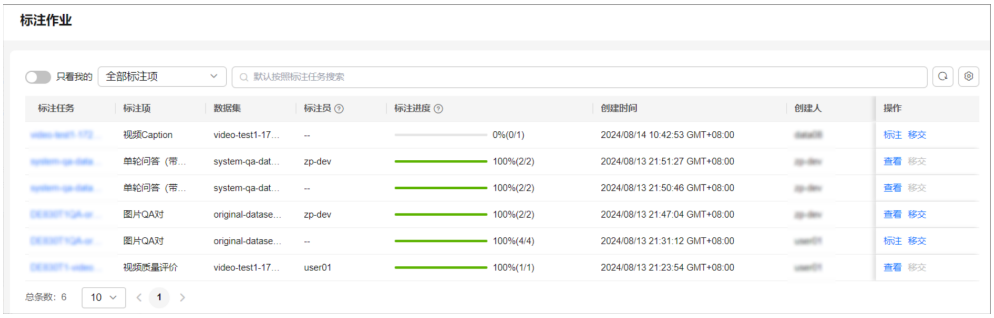
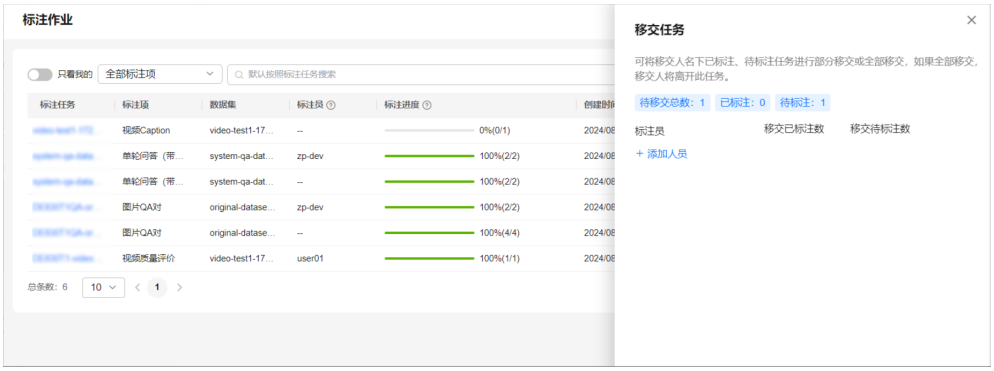


图 3-46 移交标注任务



7. 进入标注页面后，逐一对数据进行标注。

   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，如图3-47。

图 3-47 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。

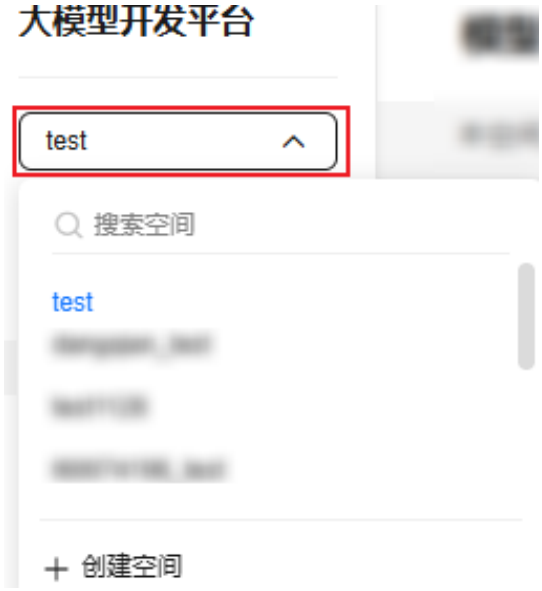
3.6.2.2 审核文本类数据集标注结果

创建数据集标注任务时，如果设置了启用标注审核，在完成标注后可以在“标注审核”页面审核标注结果。

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。创建标注任务时如果指定了审核人员，则审核人员可以审核数据集，管理员（主账号）可以对所有数据集进行审核。

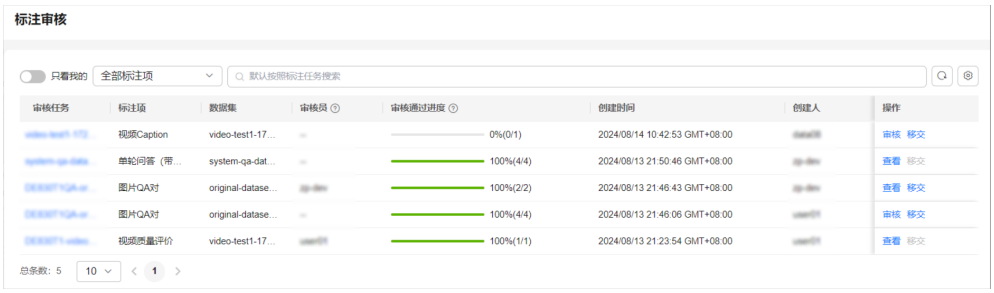
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-48 进入操作空间



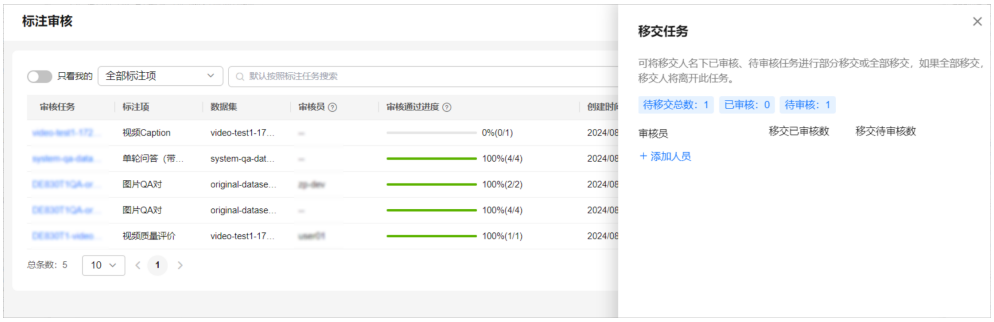
2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注审核”
3. 在“标注审核”页面，单击操作列“审核”可进入审核页面审核数据。

图 3-49 标注审核



如果需要将该审核任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-50 移交审核任务



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，期间可对不满足要求的数据进行驳回，驳回后将分给标注人员重新标注。

3.6.2.3 上线标注后的文本类数据集

数据集标注完成并且审核无问题后，需要对该数据集执行上线操作。上线后的数据集可以用于后续的数据评估、发布任务。

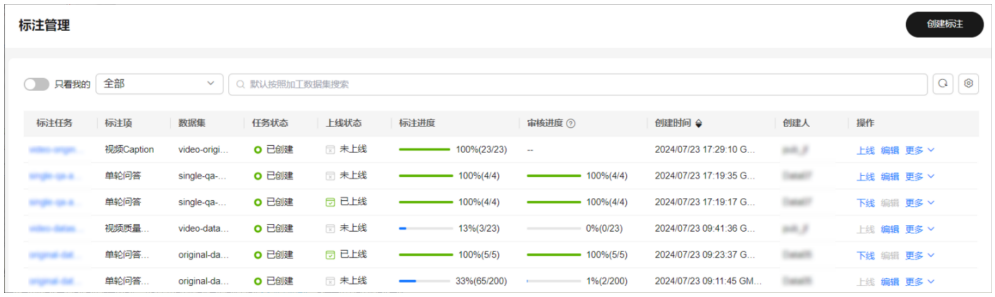
上线标注后的数据集步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-51 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注管理”
3. 在“标注管理”页面，单击操作列的“上线”对数据集进行上线。

图 3-52 上线标注后的数据集



说明

对不再使用的数据集可在操作列执行下线操作。若对当前标注数据集已执行发布操作**发布文本类数据集**，则不可将该标注数据集下线。

3.6.3 标注视频类数据集

3.6.3.1 创建视频类数据集标注任务

创建视频类数据集标注任务前，请先完成**创建视频类数据集加工任务**。

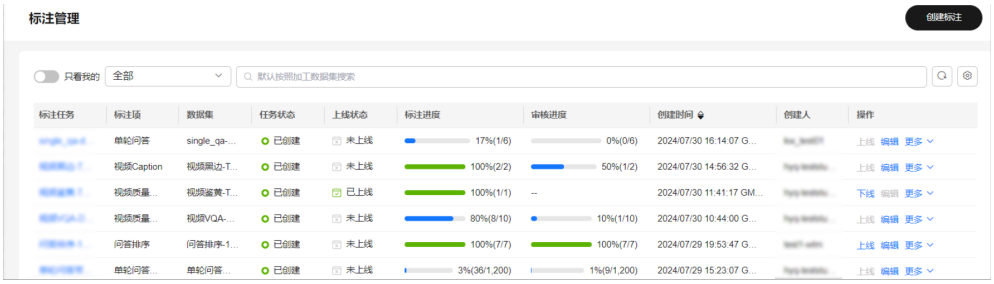
创建视频类数据集标注任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-53 进入操作空间

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

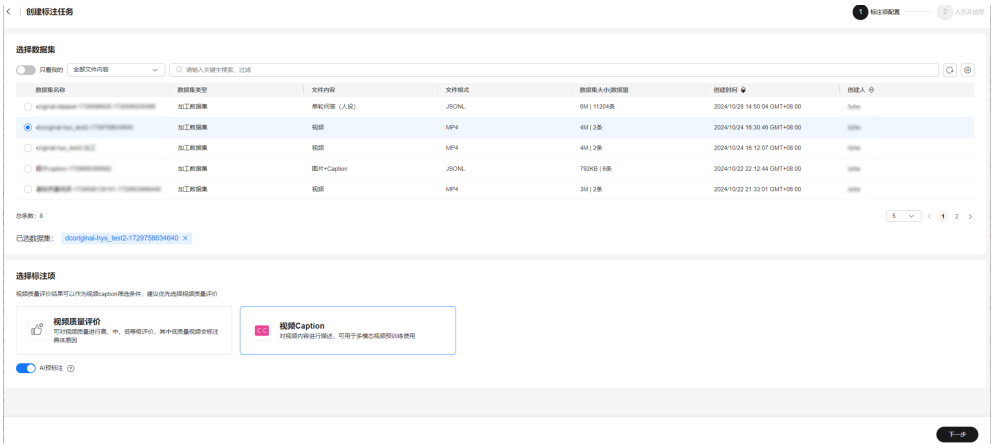
图 3-54 标注管理



2. 在“创建标注任务”页面选择需要标注的加工后的视频类数据集，并设置标注项。

   当选择“视频Caption”标注项时，可以设置使用AI大模型对数据集进行预标注。启动预标注将会借助AI模型生成标注内容，这些内容不会覆盖原始数据集，仅作为标注人员的参考，以提高标注效率。

图 3-55 创建标注任务



3. 单击“下一步”设置标注人员及信息，单击“完成创建”。

   分配标注任务时，可以选择是否启用多人标注。启用多人标注后，可以指定参与标注的人员。

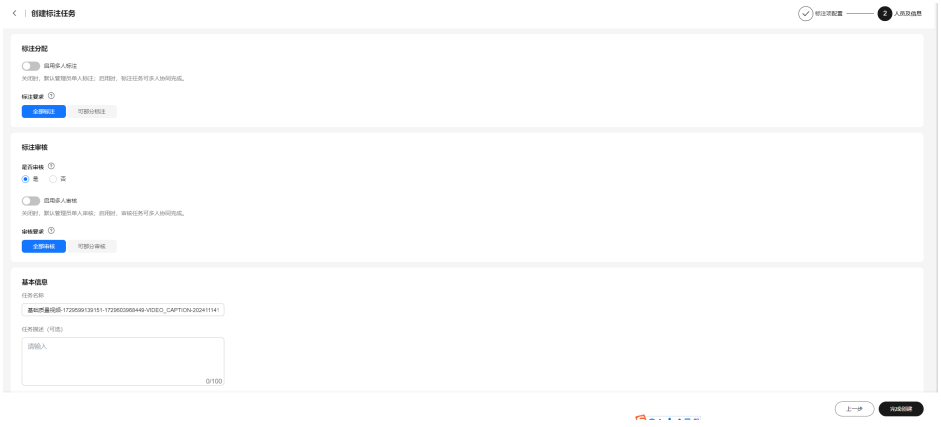
   对于使用AI进行预标注的视频Caption任务可设置以下两种方式的“标注要求”：

   - 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。

   标注任务可选择是否启用标注审核，可设置多人审核，详见**审核文本类数据集标注结果**。审核要求可以选择以下两种方式：

   - 选择“可部分审核”：审核人员确认部分数据达到标注要求后，可以一键通过所有的标注。
   - 选择“全部审核”：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。

图 3-56 设置标注人员、标注信息示例



4. 在“标注管理”页面，单击操作列“上线”，可执行后续标注操作。对于未上线的标注任务，可执行编辑和删除操作。
5. 在“标注作业”页面，单击操作列“标注”可进行数据标注。如果需要将该标注任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-57 标注作业

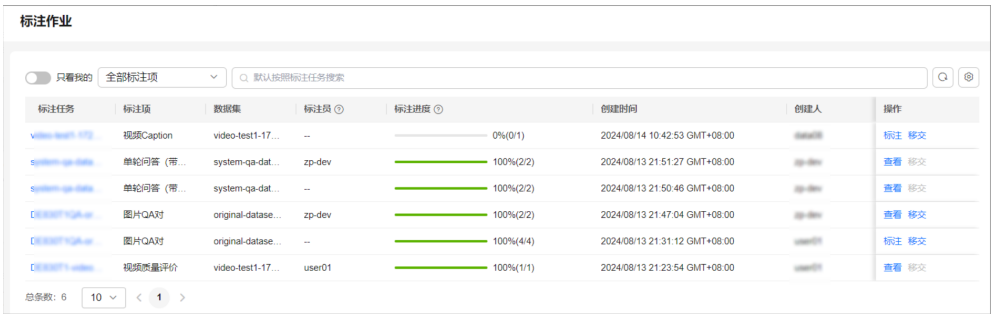
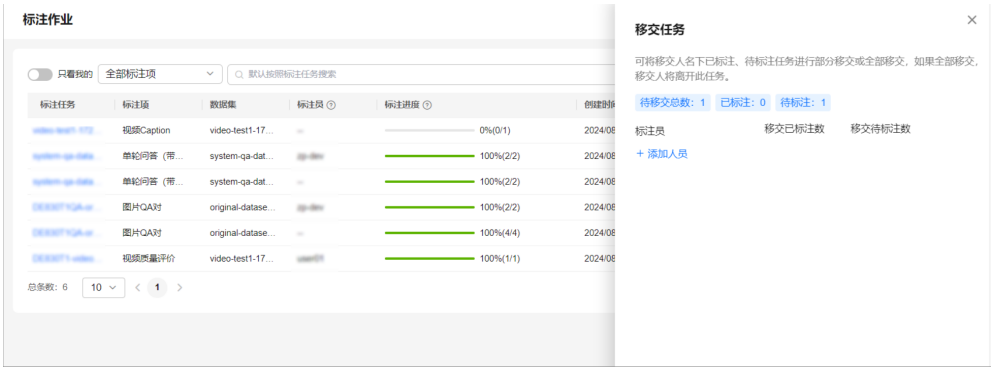


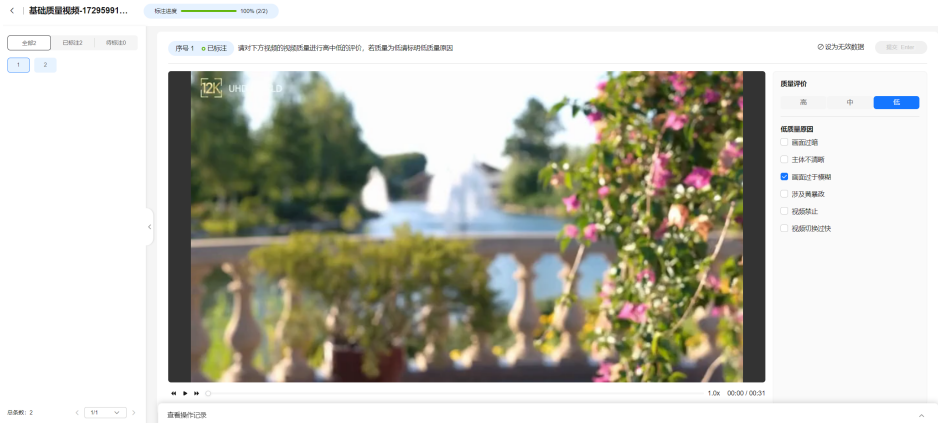
图 3-58 移交标注任务



6. 进入标注页面后，逐一对数据进行标注。

   以标注视频Caption数据为例，需要逐一标注视频的质量，如图3-59。

图 3-59 视频类数据集标注示例



7. 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使用AI大模型进行辅助标注，并且将标注要求设置为“可部分标注”，则可以在标注部分数据后，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。

3.6.3.2 审核视频类数据集标注结果

创建数据集标注任务时，如果设置了启用标注审核，在完成标注后可以在“标注审核”页面审核标注结果。

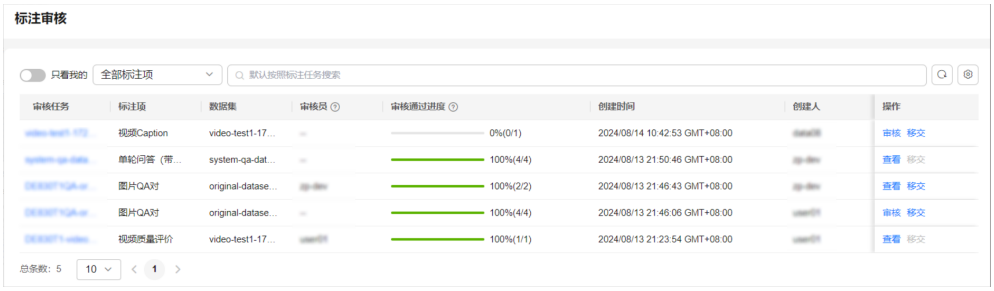
对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。创建标注任务时如果指定了审核人员，则审核人员可以审核数据集，管理员（主账号）可以对所有数据集进行审核。

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-60 进入操作空间

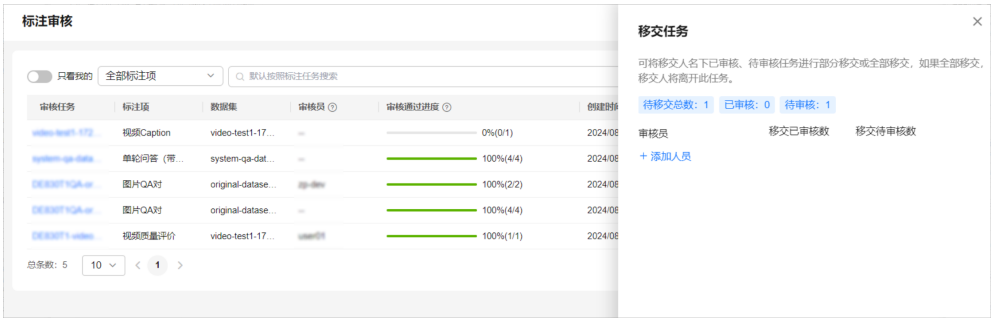
2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注审核”
3. 在“标注审核”页面，单击操作列“审核”可进入审核页面审核数据。

图 3-61 标注审核



如果需要将该审核任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-62 移交审核任务



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，期间可对不满足要求的数据进行驳回，驳回后将分给标注人员重新标注。

3.6.3.3 上线标注后的视频类数据集

数据集标注完成并且审核无问题后，需要对该数据集执行上线操作。上线后的数据集可以用于后续的数据评估、发布任务。

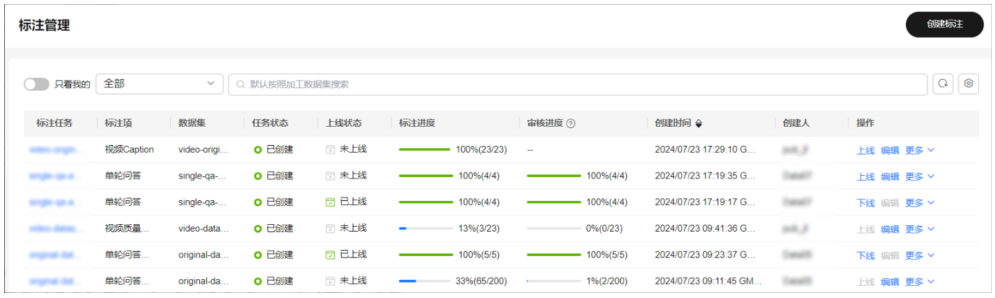
上线标注后的数据集步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-63 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注管理”
3. 在“标注管理”页面，单击操作列的“上线”对数据集进行上线。

图 3-64 上线标注后的数据集



说明

对不再使用的数据集可在操作列执行下线操作。若对当前标注数据集已执行发布操作**发布视频类数据集**，则不可将该标注数据集下线。

3.6.4 标注图片类数据集

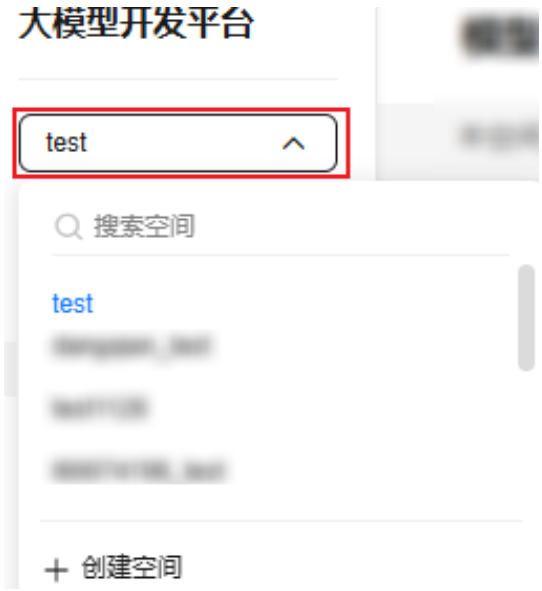
3.6.4.1 创建图片类数据集标注任务

创建图片类数据集标注任务前，请先完成**创建图片类数据集加工任务**。

创建图片类数据集标注任务步骤如下：

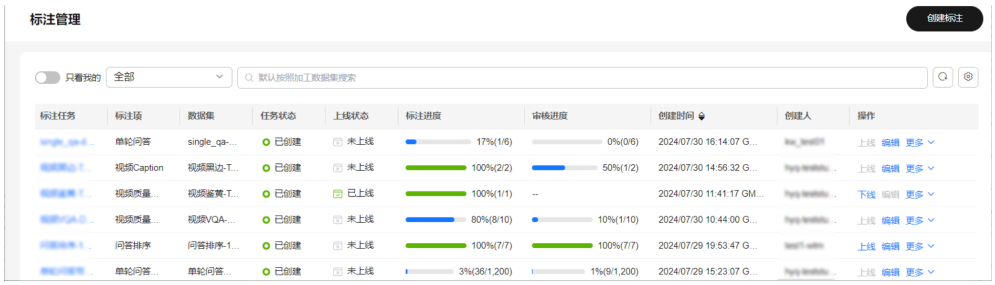
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-65 进入操作空间



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

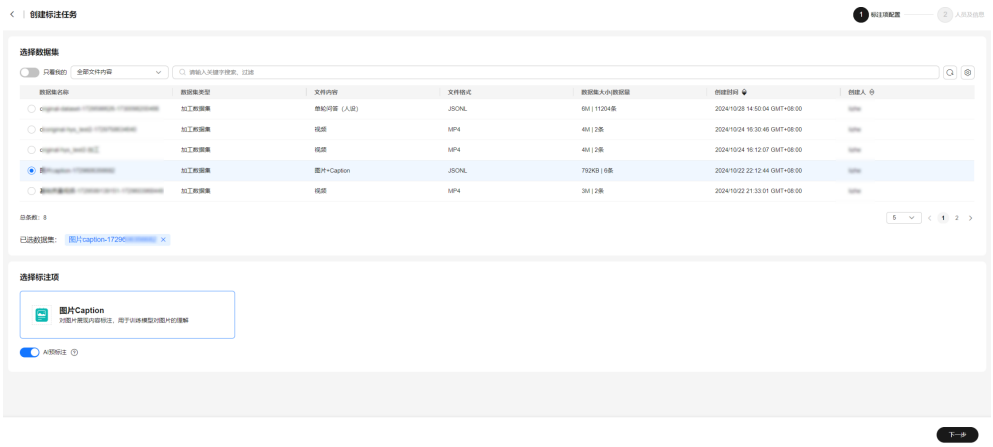
图 3-66 标注管理



3. 在“创建标注任务”页面选择需要标注的加工后的图片类数据集，并设置标注项。

   当选择“图片Caption”标注项时，可以设置使用AI大模型对数据集进行预标注。启动预标注将会借助AI模型生成标注内容，这些内容不会覆盖原始数据集，仅作为标注人员的参考，以提高标注效率。

图 3-67 创建标注任务



4. 单击“下一步”设置标注人员及信息，单击“完成创建”。

   分配标注任务时，可以选择是否启用多人标注。启用多人标注后，可以指定参与标注的人员。

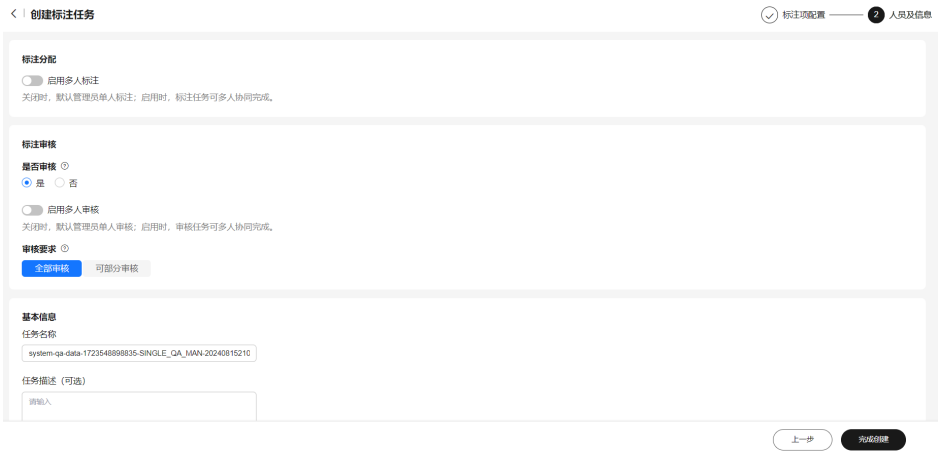
   对于使用AI进行预标注的图片Caption任务可设置以下两种方式的“标注要求”：

   - 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。

   标注任务可选择是否启用标注审核，可设置多人审核，详见**审核文本类数据集标注结果**。审核要求可以选择以下两种方式：

   - 选择“可部分审核”：审核人员确认部分数据达到标注要求后，可以一键通过所有的标注。
   - 选择“全部审核”：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。

图 3-68 设置标注人员、标注信息示例



5. 在“标注管理”页面，单击操作列“上线”，可执行后续标注操作。对于未上线的标注任务，可执行编辑和删除操作。
6. 在“标注作业”页面，单击操作列“标注”可进行数据标注。如果需要将该标注任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-69 标注作业

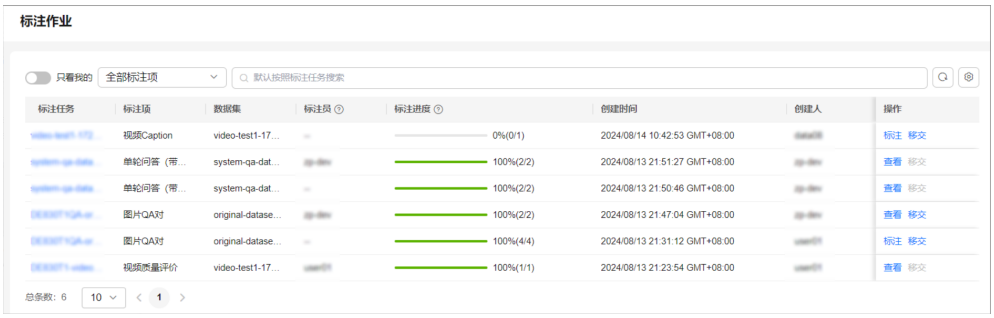
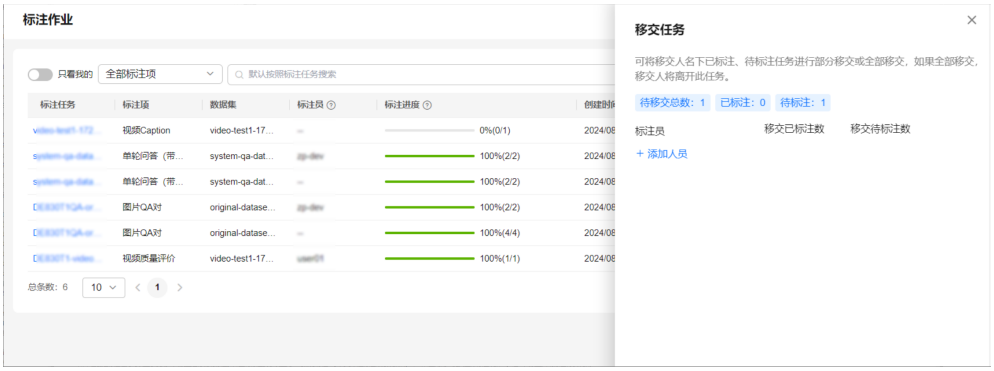


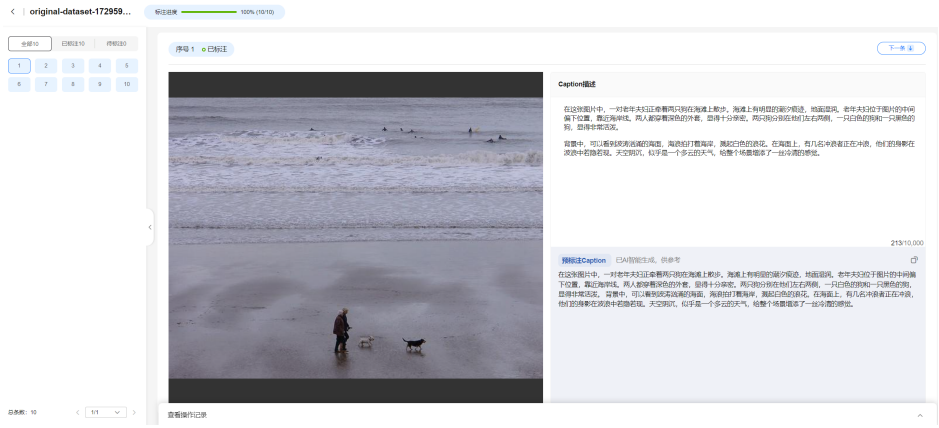
图 3-70 移交标注任务



7. 进入标注页面后，逐一对数据进行标注。

   以标注图片Caption数据为例，逐一标注图片的Caption描述，如图3-71，右下角展示了AI预标注的Caption。

图 3-71 图片类数据集标注示例



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。

   如果在创建标注任务时设置了使用AI大模型进行辅助标注，并且将标注要求设置为“可部分标注”，则可以在标注部分数据后，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。

3.6.4.2 审核图片类数据集标注结果

创建数据集标注任务时，如果设置了启用标注审核，在完成标注后可以在“标注审核”页面审核标注结果。

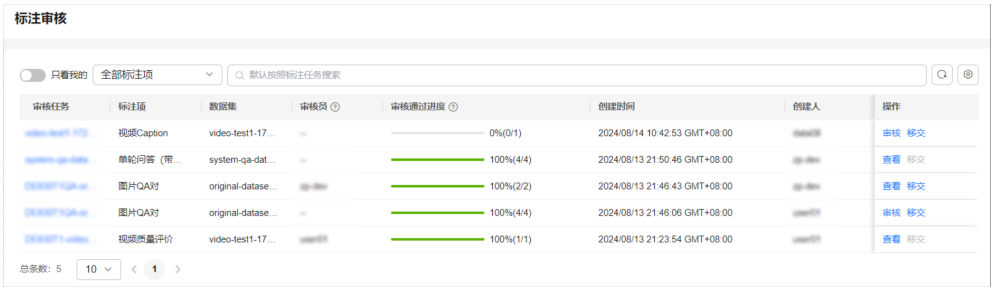
对于审核不合格的数据可以填写不合格原因并驳回给标注员重新标注。创建标注任务时如果指定了审核人员，则审核人员可以审核数据集，管理员（主账号）可以对所有数据集进行审核。

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-72 进入操作空间

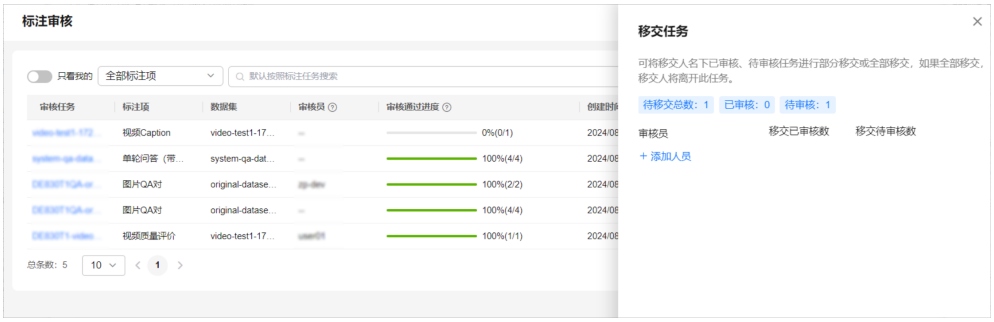
2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注审核”
3. 在“标注审核”页面，单击操作列“审核”可进入审核页面审核数据。

图 3-73 标注审核



如果需要将该审核任务移交给其他人员，可以单击操作列“移交”设置移交人员以及移交的数量。

图 3-74 移交审核任务



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，期间可对不满足要求的数据进行驳回，驳回后将分给标注人员重新标注。

3.6.4.3 上线标注后的图片类数据集

数据集标注完成并且审核无问题后，需要对该数据集执行上线操作。上线后的数据集可以用于后续的数据评估、发布任务。

上线标注后的数据集步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-75 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据标注 > 标注管理”
3. 在“标注管理”页面，单击操作列的“上线”对数据集进行上线。

图 3-76 上线标注后的数据集

说明

对不再使用的数据集可在操作列执行下线操作。若对当前标注数据集已执行发布操作**发布图片类数据集**，则不可将该标注数据集下线。

3.7 评估数据集

3.7.1 数据集评估场景介绍

数据评估概念

数据评估旨在通过对数据集进行系统的质量检查，评估其准确性、完整性、一致性和代表性等多个维度，发现潜在问题并加以解决。

在构建和使用数据集的过程中，数据评估是确保数据质量的关键步骤，直接影响模型的性能和应用效果。高质量的数据集能够显著提升模型的准确性，并增强模型在实际应用中的可靠性与稳定性。因此，数据评估是数据工程中不可或缺的一环，帮助用户在数据准备阶段识别并解决数据中的问题，为后续的模型训练和优化奠定坚实基础。

ModelArts Studio大模型开发平台提供了全面的数据集质量评估工具，能够帮助用户从多个维度检测和优化数据集的质量。平台预设了多种数据类型的基础评估标准，用户可以直接使用这些标准，也可以根据具体的业务需求创建自定义的评估标准。通过这种灵活的配置方式，用户能够根据不同的应用场景和目标，精确地评估和优化数据质量，确保数据在进入模型训练阶段之前达到高标准，进而提升模型的性能和效果。

数据集评估标准介绍

平台预置了多种数据类型的基础评估标准，用户可以直接使用这些标准，也可以根据具体的业务需求创建自定义的评估标准。

- NLP数据质量标准 V1.0：ModelArts Studio大模型开发平台针对文本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。该标准旨在帮助用户高效评估和优化文本数据的质量，确保数据符合模型训练的要求，提升模型的性能和可靠性。用户可以直接使用该标准进行评估，也可以根据特定业务需求进行自定义调整，确保评估标准与应用场景高度契合，从而为后续的模型训练和优化提供高质量的数据支持。
- 视频数据质量标准 V1.0：ModelArts Studio大模型开发平台针对视频数据集预设了一套评估标准，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。该标准帮助用户评估和优化视频数据的质量，确保数据符合大模型训练的要求，提升模型的精度与可靠性。用户可以直接使用该标准进行评估，也可根据具体的业务需求自定义评估标准，确保视频数据满足不同应用场景的要求，为后续的模型训练和优化提供高质量的视频数据支持。
- 图片数据质量标准 V1.0：ModelArts Studio大模型开发平台针对图片数据集预设的一套评估标准，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。该标准帮助用户系统地评估和优化图片数据的质量，确保数据符合模型训练的要求，从而提升模型的准确性和应用效果。用户可以直接采用该标准进行评估，或根据具体业务需求自定义评估标准，以确保图片数据符合特定场景的需求，为后续的模型训练和优化提供可靠的数据支持。

数据评估意义

数据评估在数据工程中的作用非常重要，它帮助用户确保数据在进入模型训练阶段之前具备高质量，从而提升模型的效果和可靠性。数据评估的主要意义体现在以下几个方面：

- **确保数据质量**：通过评估数据集的准确性、完整性和一致性，用户可以及时发现并修复数据中的问题，确保数据符合训练标准。
- **提升模型性能**：高质量的数据集直接影响模型的训练效果。通过准确的评估，用户能够确保数据集的高质量，进而提升模型的性能和精度。
- **减少数据问题带来的风险**：数据中潜在的错误和缺陷可能导致模型训练不充分或效果不理想。通过数据评估，用户能够提前发现和解决这些问题，避免模型训练阶段出现数据问题。
- **灵活的评估标准**：ModelArts Studio大模型开发平台不仅提供预设的标准，还允许用户根据不同的数据类型和业务需求创建自定义的评估标准，使评估过程更加灵活和精准。
- **节省时间和成本**：通过自动化的数据评估功能，用户能够迅速了解数据的质量问题，减少手动检查的工作量和时间成本，为后续的数据优化和模型训练节省资源。

总的来说，数据评估为用户提供了一种高效、可靠的数据质量检测机制，使得在数据准备阶段就能够确保数据的高标准，从而为后续的模型训练和优化打下坚实基础，帮助提升大模型的精度和可靠性。

支持数据评估的数据集类型

ModelArts Studio大模型开发平台支持评估操作的数据集类型如下：

- 文本类数据集，详见**创建文本类数据集评估任务**。
- 视频类数据集，详见**创建视频类数据集评估任务**。
- 图片类数据集，详见**创建图片类数据集评估任务**。

3.7.2 评估文本类数据集

3.7.2.1 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**。

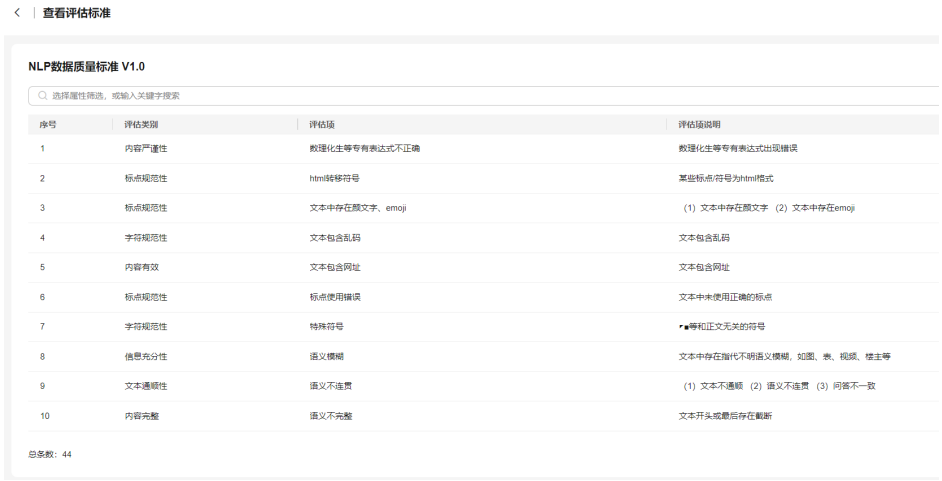
创建文本类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-77 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估标准”，平台预置的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 3-78 预置文本类数据集评估标准



3. 在“评估标准”页面单击右上角“创建评估标准”，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”创建评估标准。评估标准创建完成后可以在“评估标准”页面查看创建的评估标准，并支持编辑、删除操作。

3.7.2.2 创建文本类数据集评估任务

创建文本类数据集评估任务前，请先完成**创建文本类数据集加工任务**。

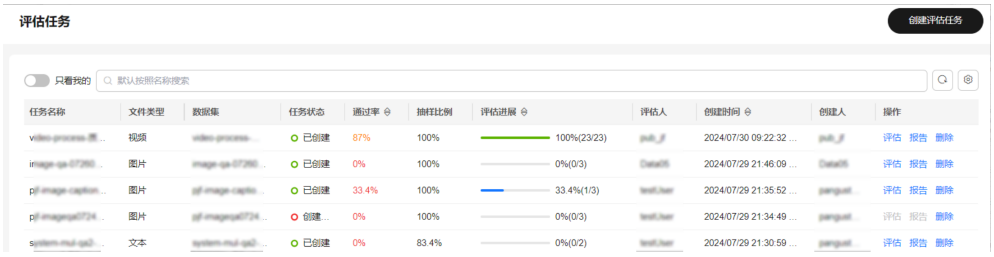
创建文本类数据集评估任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-79 进入操作空间

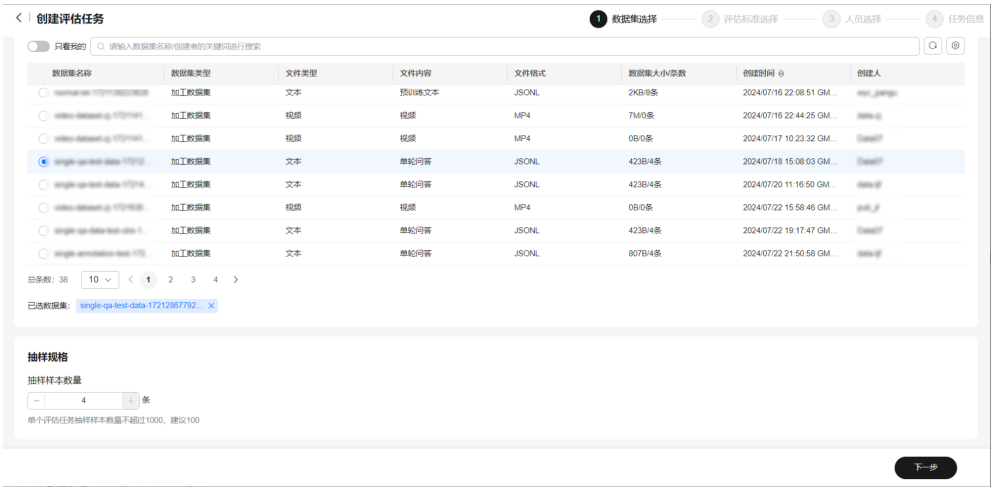
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”，单击界面右上角“创建评估任务”。

图 3-80 创建评估任务



3. 在“数据集选择”页签选择需要进行评估的加工数据集，并设置抽样规格，即从数据集中抽取一定比例数据用于评估。

图 3-81 选择数据集



4. 单击“下一步”选择需要使用的评估标准。标准选择完成后，单击“下一步”设置评估人员。

图 3-82 选择评估标注

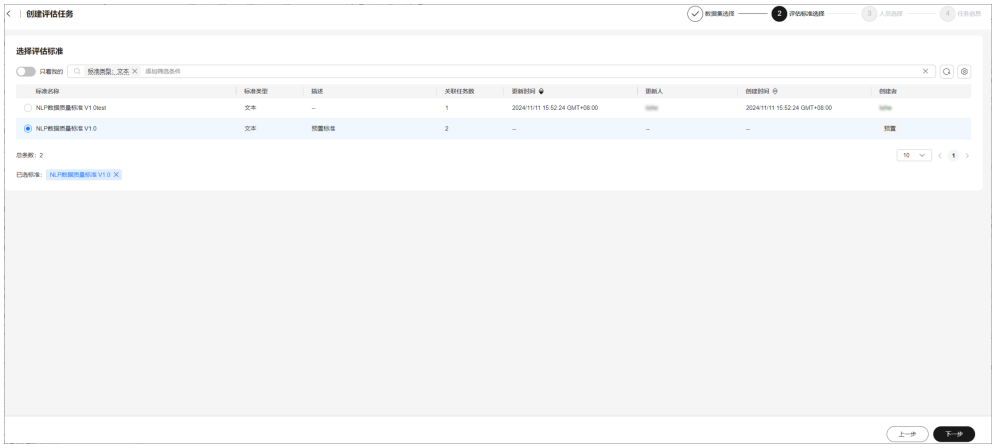


图 3-83 选择评估人员

5. 评估人员设置完成后，单击“下一步”填写任务名称。单击“完成创建”，将返回“评估任务”页面，创建成功后状态将显示为“已创建”状态。
6. 评估任务创建成功后，单击操作列“评估”进入评估页面。

图 3-84 评估数据集质量

7. 在评估页面，可参考评估项对当前数据的问题进行标注，且不满足时需要单击“不通过”，满足则单击“通过”。对于文本类数据集而言，可选择问题内容后，单击鼠标右键进行数据问题的标注。

图 3-85 标记数据集问题

8. 全部数据评估完成后，评估状态显示为“100%”，表示当前数据集已经评估完成，可以回退到“评估任务”页面，查看，单击操作列“报告”，获取数据集质量评估报告。

3.7.2.3 获取文本类数据集评估报告

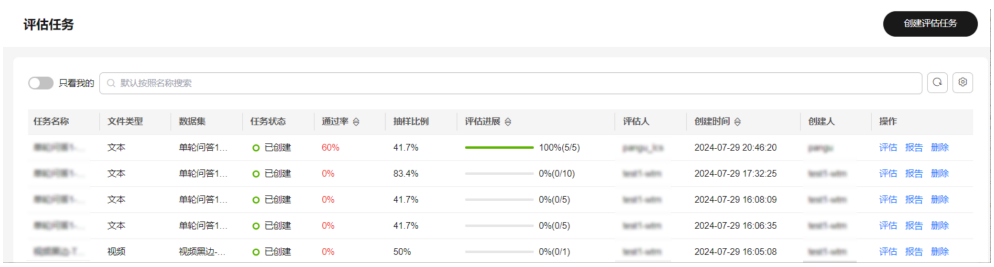
ModelArts Studio大模型开发平台提供了详细的质量评估报告，帮助用户全面了解数据集的质量情况。获取数据集评估报告步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-86 进入操作空间

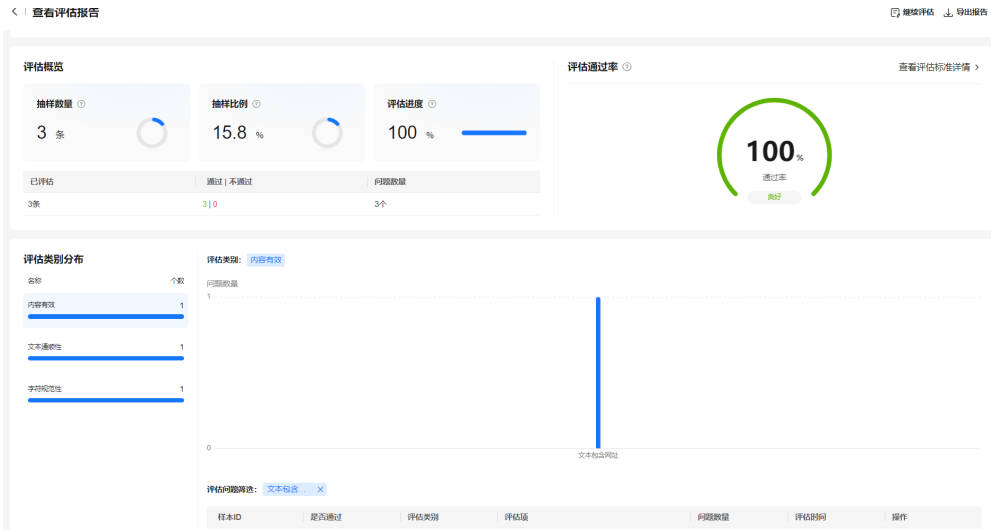
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”。
3. 单击操作列“报告”可以查看详细的质量评估报告。

图 3-87 查看数据集评估报告



4. 在“查看评估报告”页面，可以查看评估概览、通过率、评估类别分布等信息。如果数据集未完成全部评估，可以单击右上角“继续评估”，评估剩余的数据。

图 3-88 查看评估报告详情



3.7.3 评估视频类数据集

3.7.3.1 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数据集预设了一套评估标准，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**。

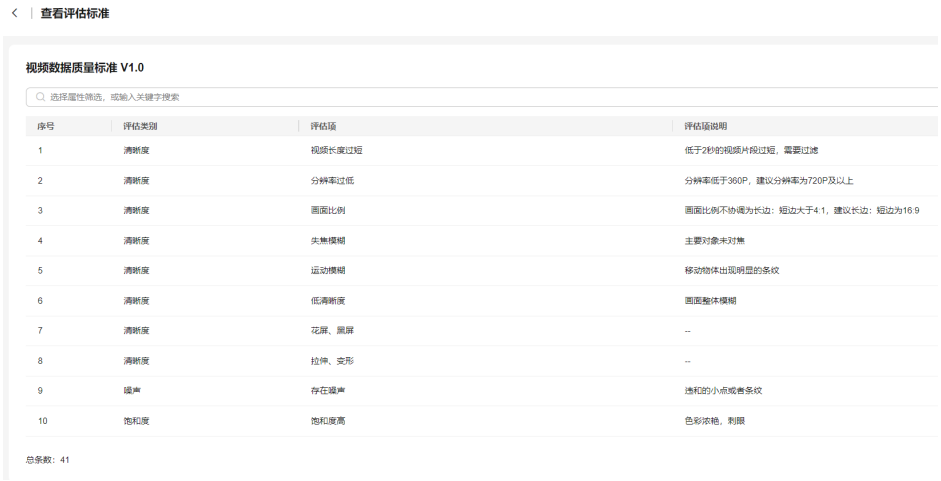
创建视频类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-89 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估标准”，平台预置的文本类数据集评估标准“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 3-90 预置视频类数据集评估标准



3. 在“评估标准”页面单击右上角“创建评估标准”，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”创建评估标准。评估标准创建完成后可以在“评估标准”页面查看创建的评估标准，并支持编辑、删除操作。

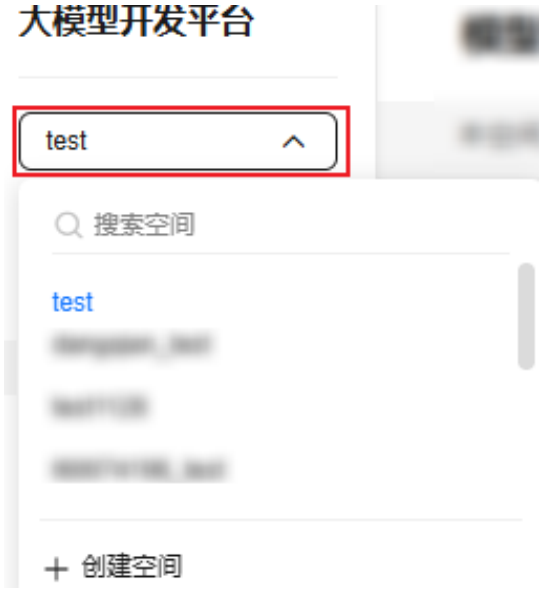
3.7.3.2 创建视频类数据集评估任务

创建视频类数据集评估任务前，请先完成**创建视频类数据集加工任务**。

创建视频类数据集评估任务步骤如下：

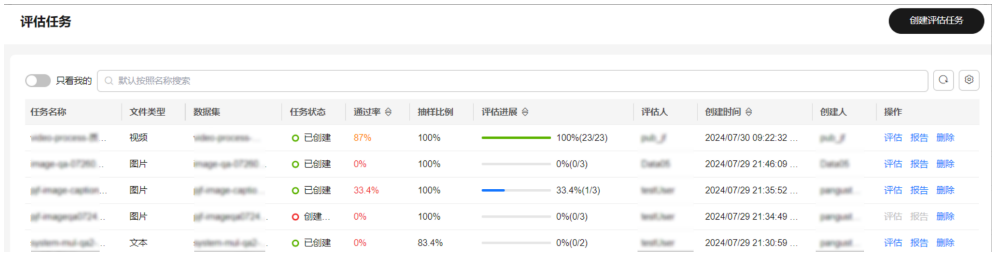
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-91 进入操作空间



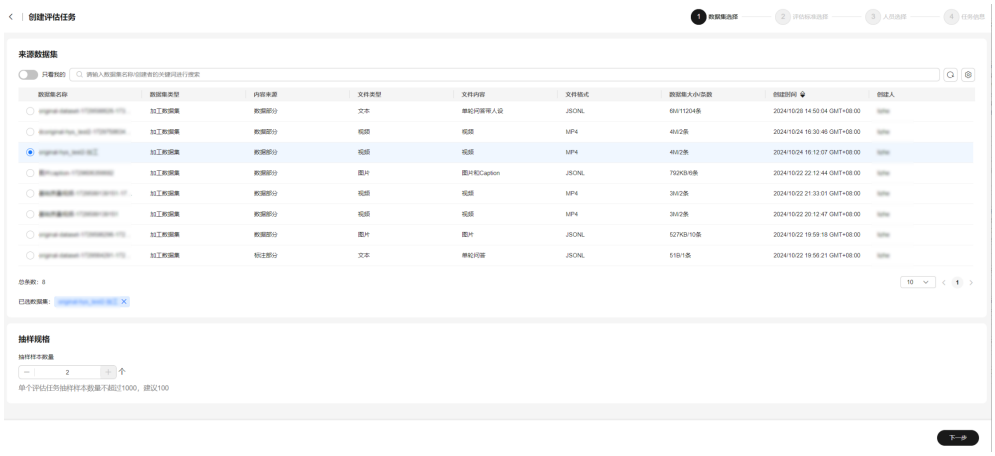
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”，单击界面右上角“创建评估任务”。

图 3-92 创建评估任务



3. 在“数据集选择”页签选择需要进行评估的加工数据集，并设置抽样规格，即从数据集中抽取一定比例数据用于评估。

图 3-93 选择数据集



4. 单击“下一步”选择需要使用的评估标准。标准选择完成后，单击“下一步”设置评估人员。

图 3-94 选择评估标注

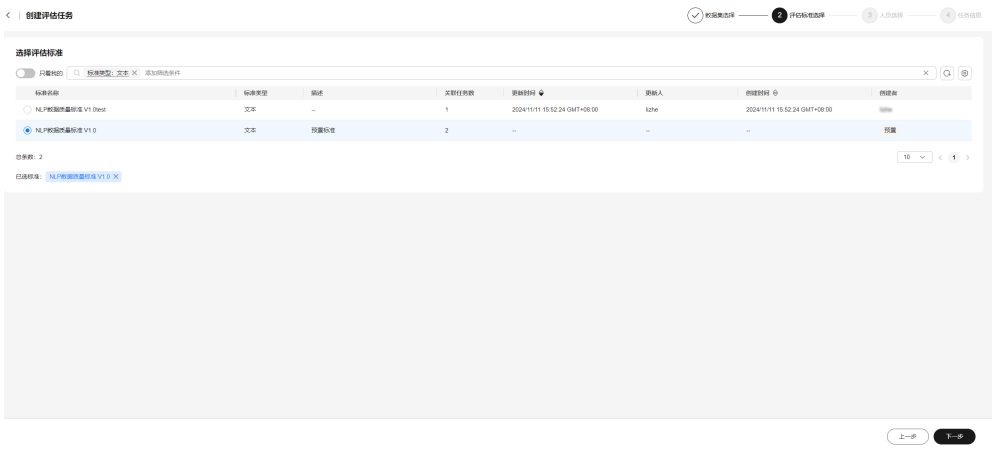
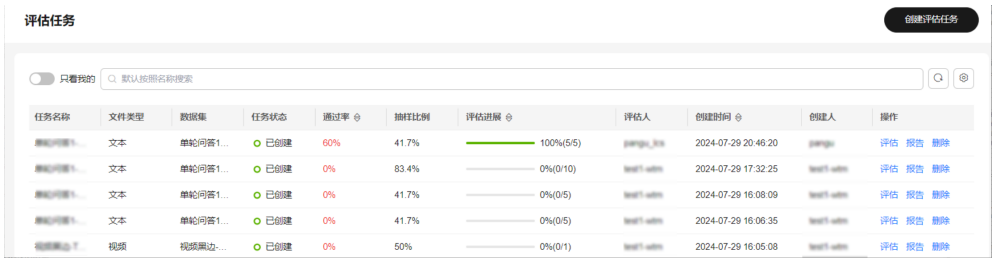


图 3-95 选择评估人员

5. 评估人员设置完成后，单击“下一步”填写任务名称。单击“完成创建”，将返回“评估任务”页面，创建成功后状态将显示为“已创建”状态。
6. 评估任务创建成功后，单击操作列“评估”进入评估页面。

图 3-96 评估数据集质量



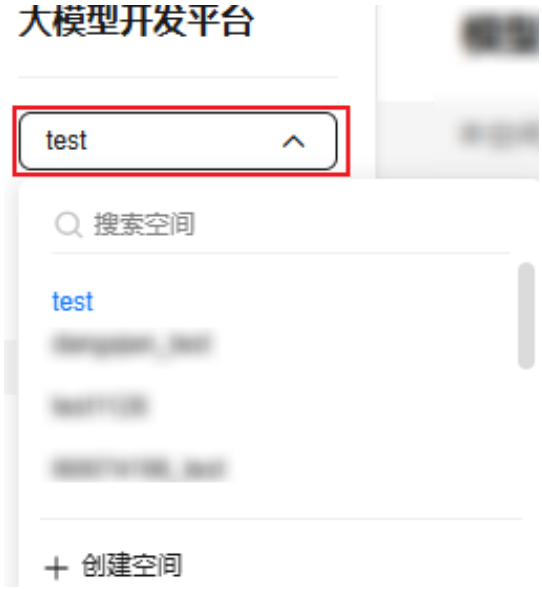
7. 在评估页面，可参考评估项对当前数据的问题进行标注，且不满足时需要单击“不通过”，满足则单击“通过”。
8. 全部数据评估完成后，评估状态显示为“100%”，表示当前数据集已经评估完成，可以回退到“评估任务”页面，查看，单击操作列“报告”，获取数据集质量评估报告。

3.7.3.3 获取视频类数据集评估报告

ModelArts Studio大模型开发平台提供了详细的质量评估报告，帮助用户全面了解数据集的质量情况。获取数据集评估报告步骤如下：

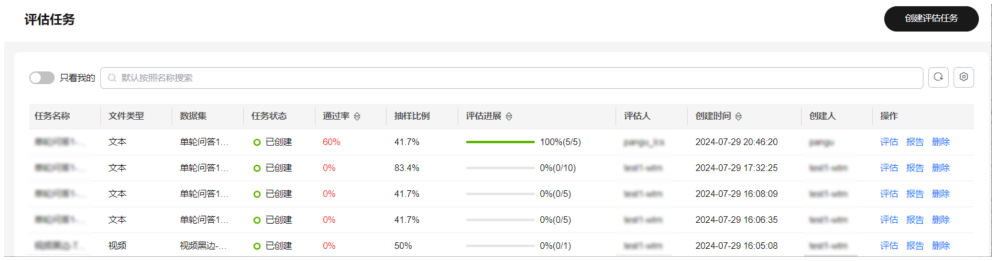
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-97 进入操作空间



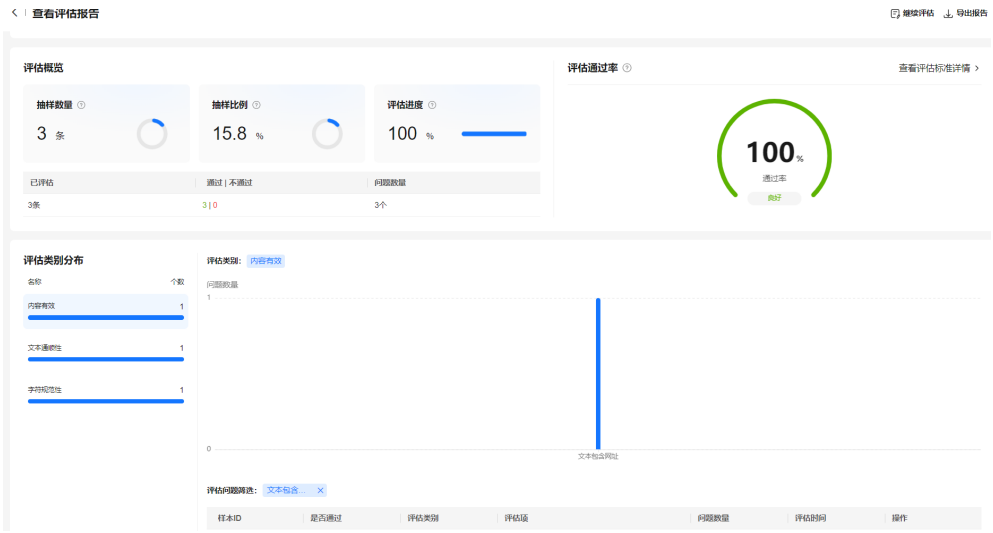
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”。
3. 单击操作列“报告”可以查看详细的质量评估报告。

图 3-98 查看数据集评估报告



4. 在“查看评估报告”页面，可以查看评估概览、通过率、评估类别分布等信息。如果数据集未完成全部评估，可以单击右上角“继续评估”，评估剩余的数据。

图 3-99 查看评估报告详情



3.7.4 评估图片类数据集

3.7.4.1 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数据集预设的一套评估标准，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**。

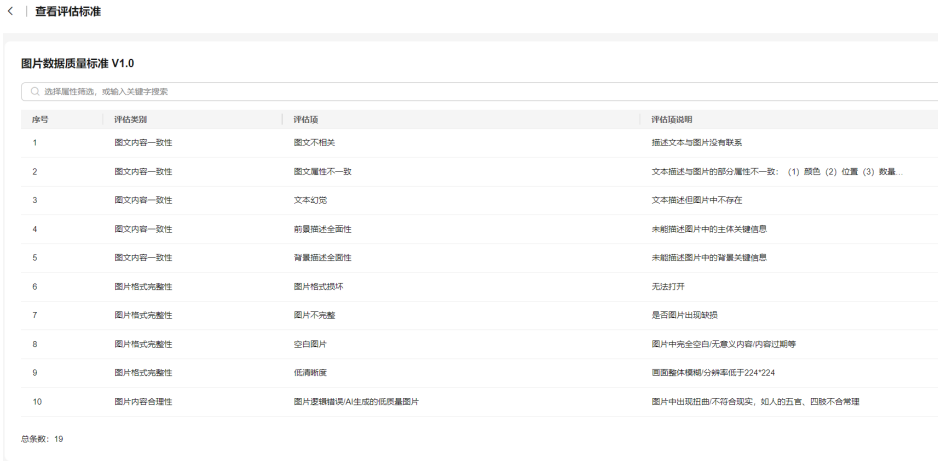
创建图片类数据集评估标准步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-100 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估标准”，平台预置的文本类数据集评估标准“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。

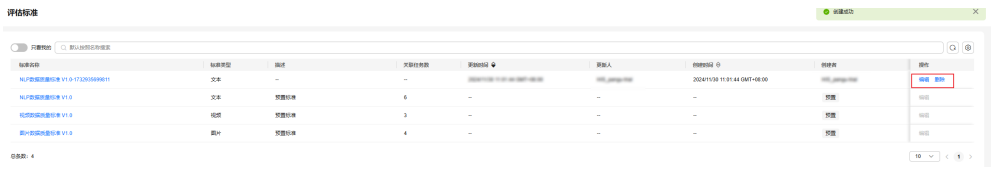
图 3-101 预置图片类数据集评估标准



3. 在“评估标准”页面单击右上角“创建评估标准”，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”创建评估标准。评估标准创建完成后可以在“评估标准”页面查看创建的评估标准，并支持编辑、删除操作，如图3-102。

图 3-102 评估标准列表



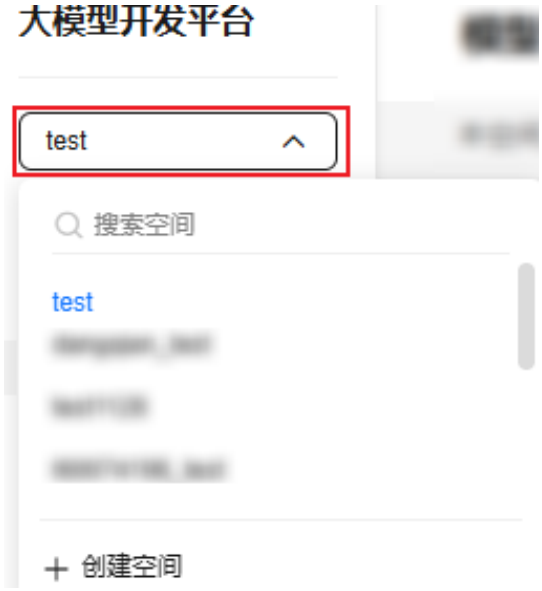
3.7.4.2 创建图片类数据集评估任务

创建图片类数据集评估任务前，请先完成**创建图片类数据集评估标准**。

创建图片类数据集评估任务步骤如下：

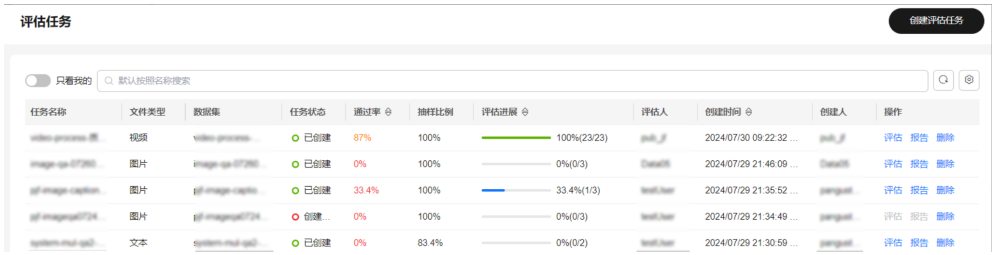
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-103 进入操作空间



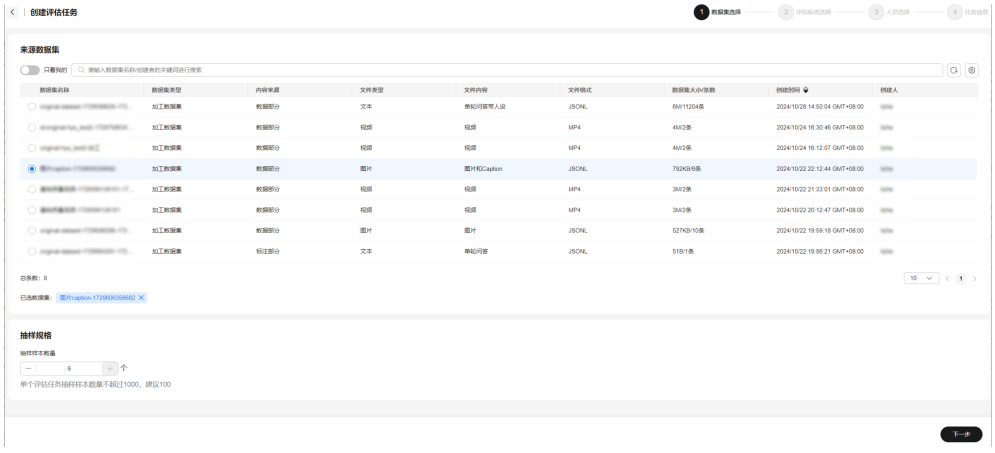
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”，单击界面右上角“创建评估任务”。

图 3-104 创建评估任务



3. 在“数据集选择”页签选择需要进行评估的加工数据集，并设置抽样规格，即从数据集中抽取一定比例数据用于评估。

图 3-105 选择数据集



4. 单击“下一步”选择需要使用的评估标准。标准选择完成后，单击“下一步”设置评估人员。

图 3-106 选择评估标注

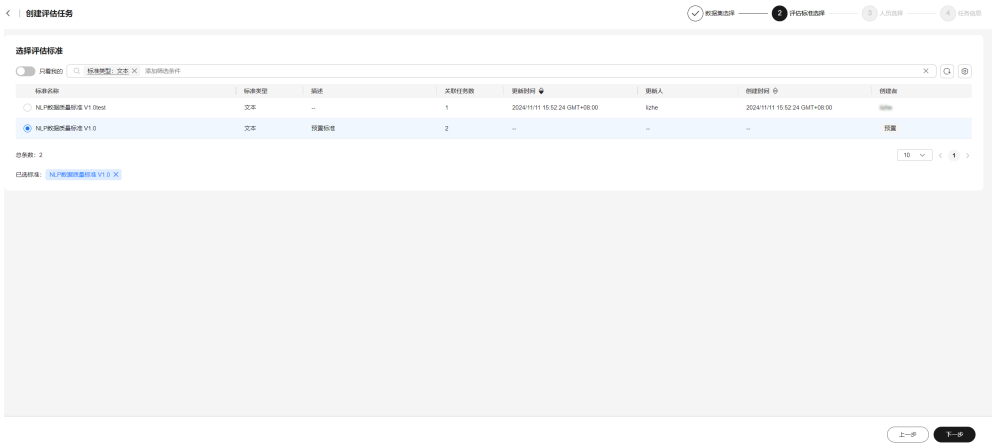
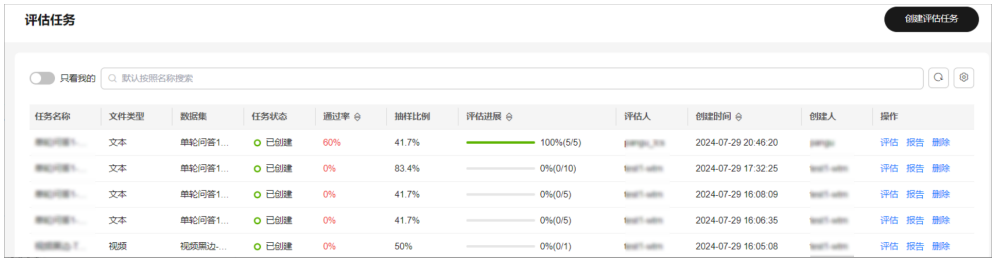


图 3-107 选择评估人员

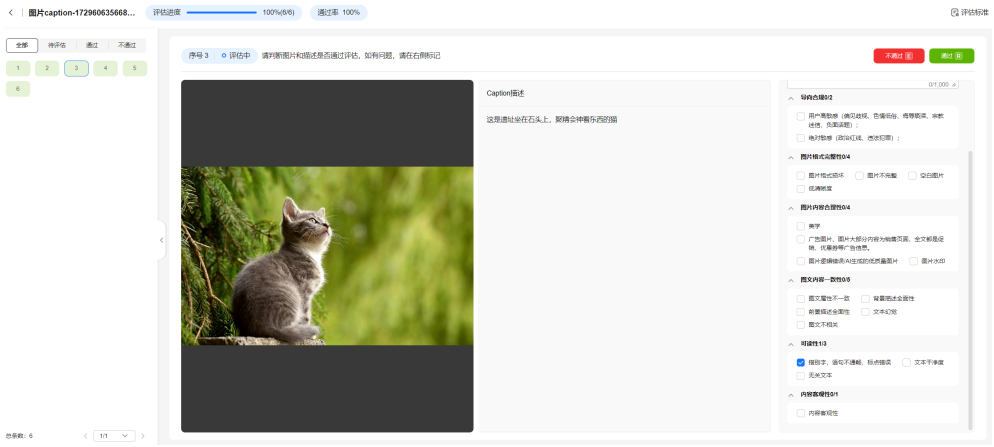
5. 评估人员设置完成后，单击“下一步”填写任务名称。单击“完成创建”，将返回“评估任务”页面，创建成功后状态将显示为“已创建”状态。
6. 评估任务创建成功后，单击操作列“评估”进入评估页面。

图 3-108 评估数据集质量



7. 在评估页面，可参考评估项对当前数据的问题进行标注，且不满足时需要单击“不通过”，满足则单击“通过”。

图 3-109 标记数据集问题



8. 全部数据评估完成后，评估状态显示为“100%”，表示当前数据集已经评估完成，可以回退到“评估任务”页面，查看，单击操作列“报告”，获取数据集质量评估报告。

3.7.4.3 获取图片类数据集评估报告

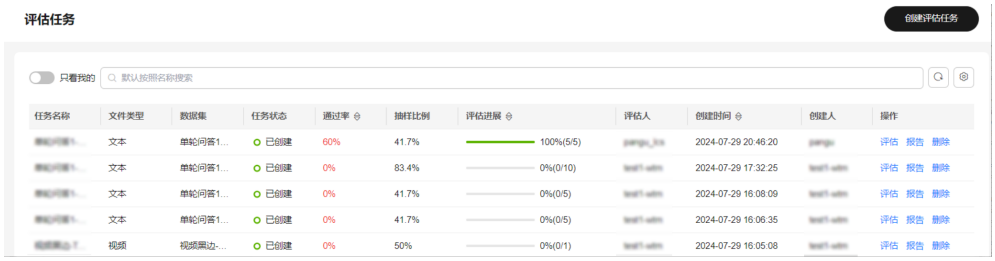
ModelArts Studio大模型开发平台提供了详细的质量评估报告，帮助用户全面了解数据集的质量情况。获取数据集评估报告步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-110 进入操作空间

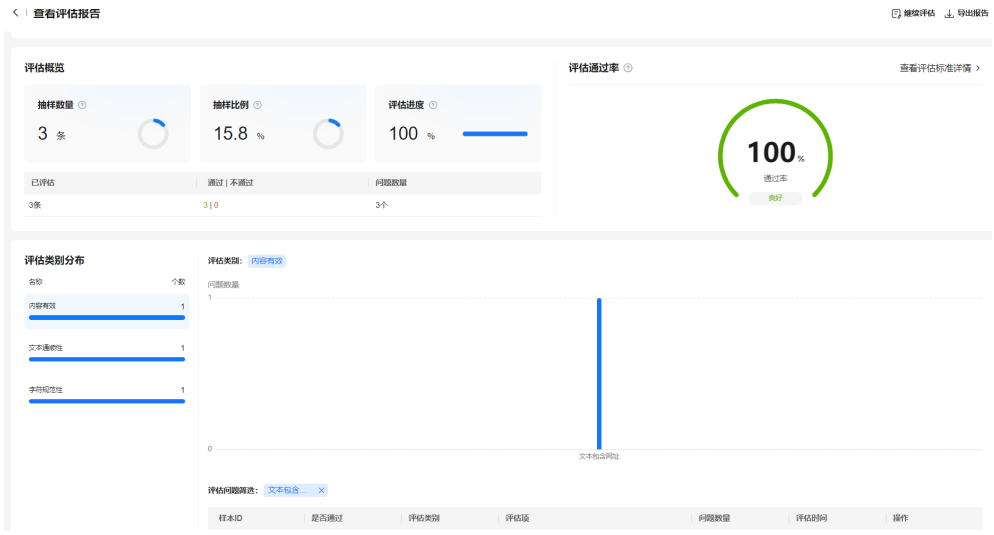
2. 在左侧导航栏中选择“数据工程 > 数据评估 > 评估任务”。
3. 单击操作列“报告”可以查看详细的质量评估报告。

图 3-111 查看数据集评估报告



4. 在“查看评估报告”页面，可以查看评估概览、通过率、评估类别分布等信息。如果数据集未完成全部评估，可以单击右上角“继续评估”，评估剩余的数据。

图 3-112 查看评估报告详情



3.8 发布数据集

3.8.1 数据集发布场景介绍

数据发布概念

数据发布是指将经过加工、标注、评估的数据集导出并生成符合特定任务或模型训练需求的正式数据集。数据发布是数据处理流程中的关键步骤，也是数据集构建的最终环节。

数据发布过程不仅包括将数据转化为适合使用的格式，还要求根据任务需求对数据集的比例进行科学调整，确保数据集在规模、质量和内容上满足模型训练的标准。

通过灵活调整数据集的比例配比，用户能够保证数据的均衡性，避免因数据分布不均可能引发的问题，从而构建高质量、适应性强的数据集，为后续的模型训练、验证和应用提供坚实的数据支持。

数据发布意义

数据发布不仅包括数据的格式转换，还涉及数据比例的调整，以确保数据在规模、质量和内容上满足训练标准。具体而言，数据集发布具有以下重要意义：

- **数据比例和结构调整**：平台提供灵活的数据比例调整功能，用户可以按需调整数据集的各类数据比例，确保数据集在训练时的代表性和均衡性，从而避免数据分布不均导致的训练问题。
- **多种数据格式支持**：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“默认格式”、“盘古格式”和“自定义格式”，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。
- **灵活的定制化服务**：对于文本类、图片类数据集，用户自定义数据格式，用户可以使用脚本灵活调整数据格式，以满足特定业务场景的需求。
- **提高训练效率**：通过发布符合标准的数据集，用户可以大幅提升数据的处理效率，减少后续的调整工作，快速进入模型训练阶段。

数据集发布是数据工程中的重要环节，它通过科学的数据比例调整和格式转换，确保数据集能够满足模型训练的要求。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，灵活选择和定制数据发布格式，保证数据的兼容性与一致性，从而为后续的模型训练和应用部署奠定坚实基础。

支持数据发布的数据集类型

ModelArts Studio大模型开发平台支持发布操作的数据集类型如下：

- 文本类数据集，详见**发布文本类数据集**。
- 视频类数据集，详见**发布视频类数据集**。
- 图片类数据集，详见**发布图片类数据集**。
- 气象类数据集，详见**发布气象类数据集**。
- 预测类数据集，详见**发布预测类数据集**。
- 其他类数据集，详见**发布其他类数据集**。

支持发布的数据格式

ModelArts Studio大模型开发平台支持将**文本类、图片类数据集**发布为三种格式：

- **默认格式**：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。
- **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。
- **自定义格式**：适用于文本类、图片类数据集，用户可以根据需求自定义数据格式，支持自定义脚本进行格式转换，灵活满足特定的业务需求。

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默认格式。

3.8.2 发布文本类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

文本类数据集支持发布的格式为：

- 默认格式：平台默认的格式。
- 盘古格式：训练盘古大模型时，需要将数据集格式发布为“盘古格式”。
- 自定义格式：文本类数据集可以使用自定义脚本进行数据格式转换。

发布文本类数据集操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

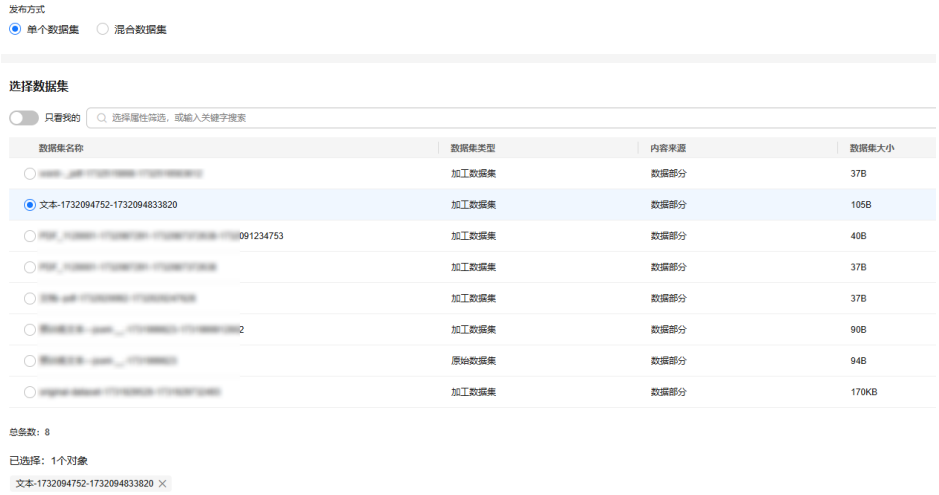
图 3-113 进入操作空间

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

图 3-114 创建文本数据集发布任务

4. 设置发布方式。除“问答排序”类型外，其余数据类型可选两种发布方式：“单个数据集”、“混合数据集”。选择数据集时，默认选择当前空间数据集，如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。
   - 若选择发布方式为“单个数据集”，选择数据集后，单击“下一步”。

图 3-115 发布方式 1



- 若选择发布方式为“混合数据集”，勾选多个数据集后，单击“下一步”。在“已选择数据集配比”中，用户可以设置从数据集中抽取指定数量的数据用于训练。进行数据配比的目的是为了确保模型能够更全面地学习和理解数据的多样性，提升模型的泛化能力和性能。

图 3-116 发布方式 2

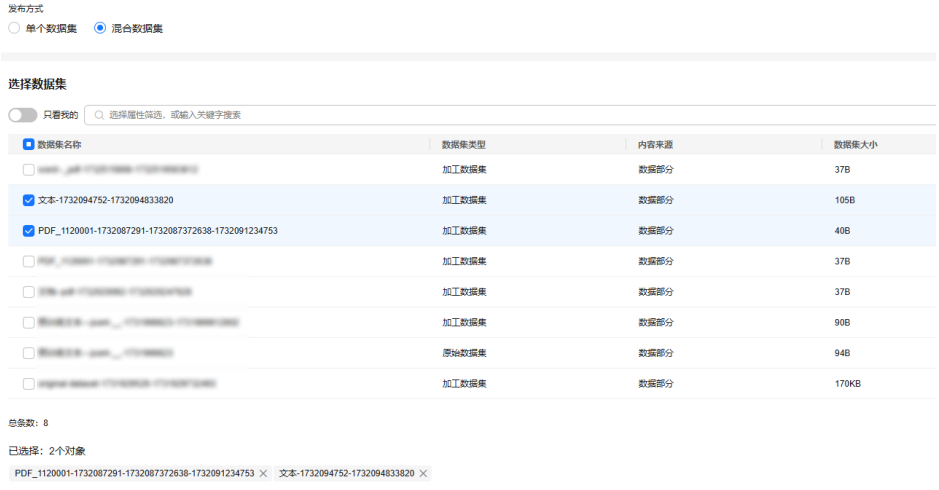


图 3-117 数据集配比

5. 设置发布格式。由于数据工程需要支持对接盘古大模型或三方大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

   在“格式配置”分页，选择发布格式，单击“下一步”。当前支持默认格式、盘古格式、自定义格式：

   - “默认格式”为数据工程功能支持的原始格式。
   - “盘古格式”为使用盘古大模型训练或评测时所需要使用的数据格式。
   - “自定义格式”可以通过自定义格式转换脚本，将数据集转化为适用于其他模型的格式。例如盘古数据集中，context、target字段分别表示问题和答

案。对于Alpaca格式的数据集，instruction对应问题，input对应上下文或者背景信息，output对应答案，用户可以上传自定义的python脚本实现数据集格式的转换。平台页面中会提供脚本示例，可下载作为参考。

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. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。

   发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.8.3 发布视频类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

视频类数据集当前仅支持发布为“默认格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

   图 3-118 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据发布”，单击界面右上角“创建发布数据集”。
3. 在“创建发布数据集”页面，选择“视频”类型的数据集。

图 3-119 创建视频数据集发布任务

4. 勾选所需要的数据集后，单击“下一步”进入数据过滤步骤。

   数据过滤阶段可以设置多种过滤属性，对视频数据集进行筛选。例如，过滤掉数据集中低于360分辨率的视频。

5. 如不需要进行数据过滤可直接单击“下一步”跳过该操作。

图 3-120 数据过滤

6. 当前视频类数据集仅支持发布默认格式，选择好数据集的发布格式后，单击“下一步”。

7. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。发布后的数据集支持重新发布和删除操作。

   发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.8.4 发布图片类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

图片类数据集支持发布的格式为：

- 默认格式：平台默认的格式。
- 盘古格式：训练盘古大模型时，需要将数据集格式发布为“盘古格式”。
- 自定义格式：文本类数据集可以使用自定义脚本进行数据格式转换。

发布图片类数据集操作步骤如下：

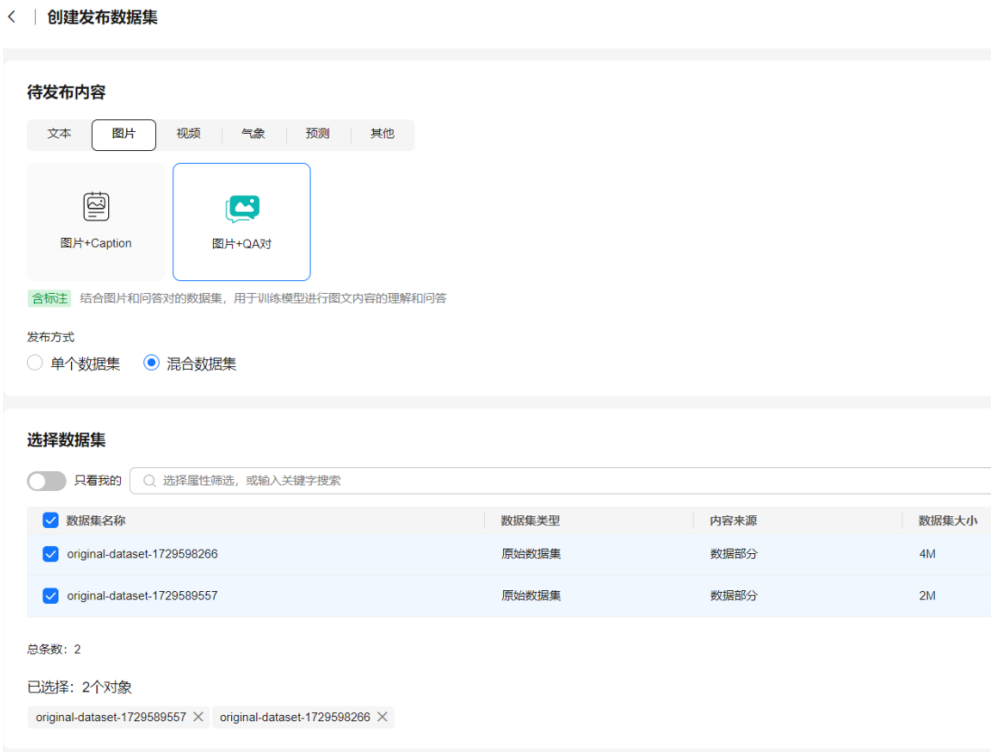
1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

   图 3-121 进入操作空间

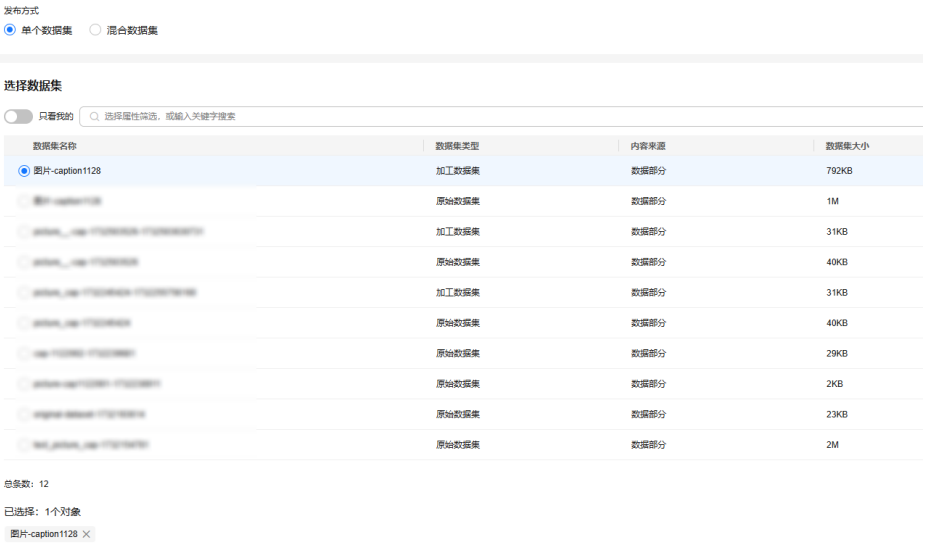
2. 在左侧导航栏中选择“数据工程 > 数据发布”，单击界面右上角“创建发布数据集”。
3. 在“创建发布数据集”页面，选择“图片”类型的数据集，并根据训练任务场景选择“图片+Caption”、“图片+QA对”类型的数据。

图 3-122 创建图片类数据集发布任务



4. 设置发布方式。图片类数据集可选两种发布方式：“单个数据集”、“混合数据集”。选择数据集时，默认选择当前空间数据集，如果用户具备其他空间的访问权限，可以选择来自其他空间的数据集。
   - 若选择发布方式为“单个数据集”，选择数据集后，单击“下一步”。

     图 3-123 发布方式 1

   - 若选择发布方式为“混合数据集”，勾选多个数据集后，单击“下一步”。在“已选择数据集配比”中，用户可以设置从数据集中抽取指定数量的数据用于训练。进行数据配比的目的是为了确保模型能够更全面地学习和理解数据的多样性，提升模型的泛化能力和性能。

图 3-124 发布方式 2

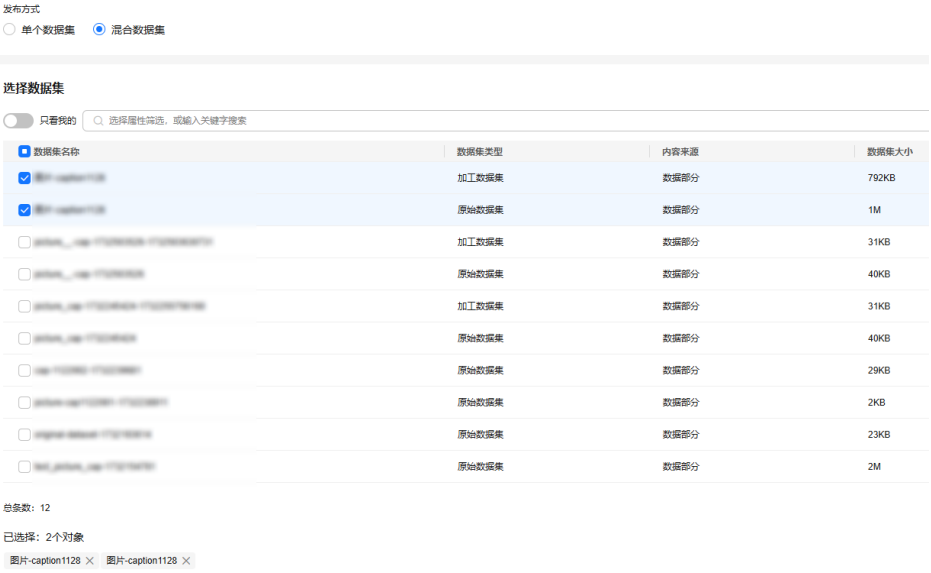


图 3-125 数据集配比

5. 设置发布格式。由于数据工程需要支持对接盘古大模型或三方大模型，为了使这些数据集能够被这些大模型正常训练，平台支持发布不同格式的数据集。

   在“格式配置”分页，选择发布格式，单击“下一步”。当前支持默认格式、盘古格式、自定义格式：

   - “默认格式”为数据工程功能支持的原始格式。
   - “盘古格式”为使用盘古大模型训练或评测时所需要使用的数据格式。
   - “自定义格式”可以通过自定义格式转换脚本，将数据集转化为适用于其他模型的格式。例如盘古数据集中，context、target字段分别表示问题和答案。对于Alpaca格式的数据集，instruction对应问题，input对应上下文或者背景信息，output对应答案，用户可以上传自定义的python脚本实现数据集格式的转换。平台页面中会提供脚本示例，可下载作为参考。

   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. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。

   发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.8.5 发布气象类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

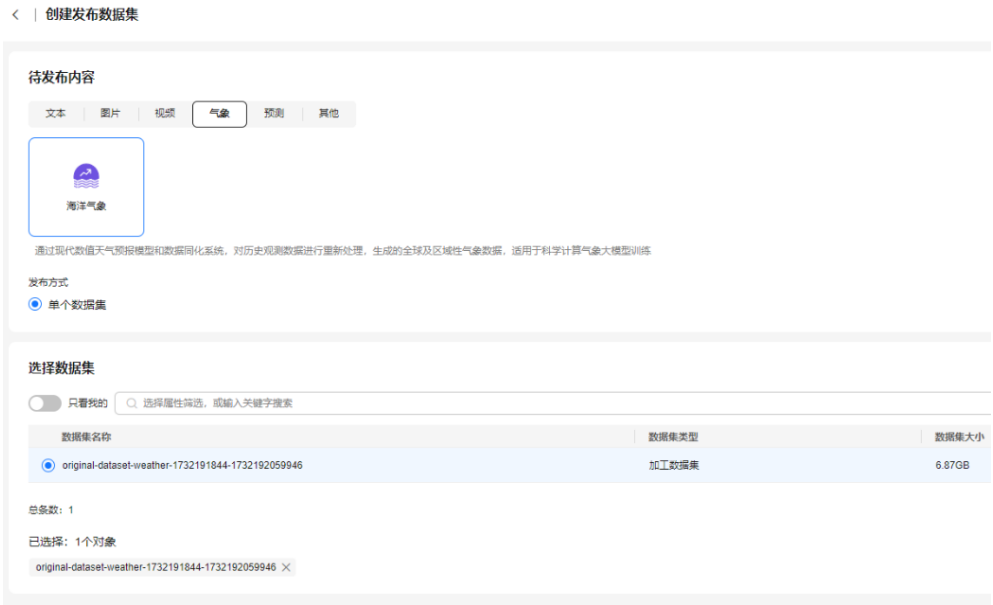
气象类数据集当前仅支持发布为“默认格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-126 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据发布”，单击界面右上角“创建发布数据集”。
3. 在“创建发布数据集”页面，选择“气象”类型的数据集，当前可选“海洋气象”类型的数据。

图 3-127 创建气象类数据集发布任务



4. 当前气象类数据集仅支持发布默认格式，选择好数据集的发布格式后，单击“下一步”。
5. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。发布后的数据集支持重新发布和删除操作。

发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.8.6 发布预测类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

预测类数据集当前仅支持发布为“默认格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-128 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据发布”，单击界面右上角“创建发布数据集”。
3. 在“创建发布数据集”页面，选择“预测”类型的数据集。并根据训练任务场景选择“时序”、“回归分类”类型的数据。

图 3-129 创建预测类数据集发布任务

4. 当前预测类数据集仅支持发布默认格式，选择好数据集的发布格式后，单击“下一步”。
5. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。发布后的数据集支持重新发布和删除操作。

   发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.8.7 发布其他类数据集

原始数据集和加工后的数据集不可以直接用于模型训练，需要独立创建一个“发布数据集”。

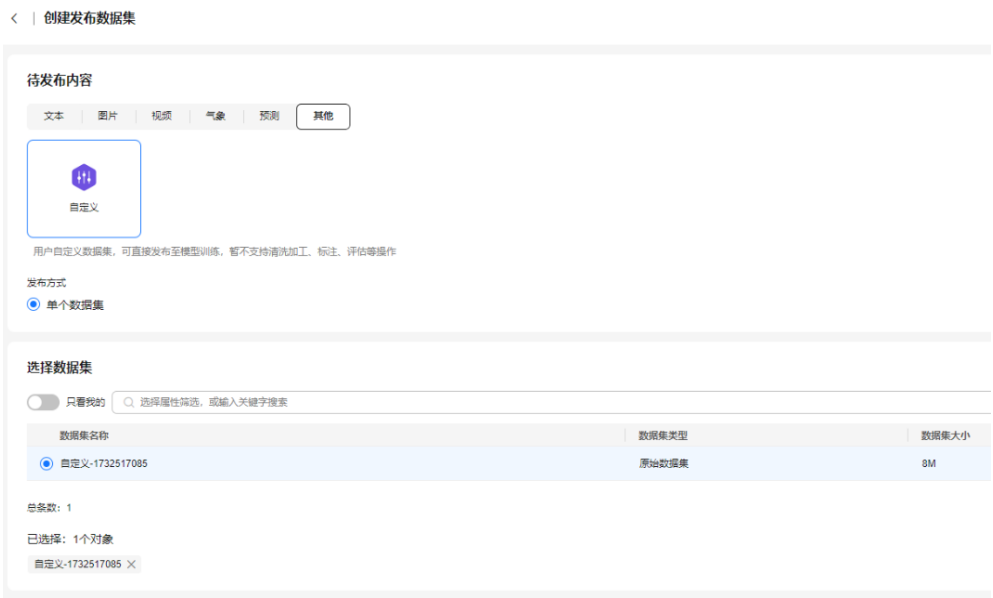
其他类数据集当前仅支持发布为“默认格式”，操作步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 3-130 进入操作空间

2. 在左侧导航栏中选择“数据工程 > 数据发布”，单击界面右上角“创建发布数据集”。
3. 在“创建发布数据集”页面，选择“其他”类型的数据集，当前可选“自定义”类型的数据。

图 3-131 创建其他类数据集发布任务



4. 当前其他类数据集仅支持发布默认格式，选择好数据集的发布格式后，单击“下一步”。
5. 设置数据集的“资产可见性”，填写数据集名称、描述，设置扩展信息后，单击“确认发布”进行数据集发布操作。发布后的数据集支持重新发布和删除操作。

   发布后的数据集会作为当前空间的数据资产同步显示在“空间资产 > 数据”页面。单击数据集名称，可以在“数据血缘”页签查看该数据集所经历的操作，如加工、发布操作。

3.9 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3-20。

表 3-20 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	The task operator not exist.	执行数据加工使用的算子出现异常，请联系服务技术支持解决。
	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。
	executor MRS job failed.	MRS服务状态异常，请联系服务技术支持解决。
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。
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。

功能模块	常见报错	解决方案
	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 status offline.	请检查数据集是否被下线。
	dataset obs file empty.	检查数据集文件是否还存在于原先的OBS桶中。
数据发布	Mrs job is failed, job id: [%s].	MRS服务状态异常，请联系服务技术支持解决。
	download obs file [%s] failed.	下载OBS桶中的数据失败，请检查OBS中数据是否被删除。
	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上线。

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，该数据集格式要求请参见**文本类数据集格式要求**。

构建 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4-1**。

表 4-1 构建 NLP 大模型所需数据量

模型规格	训练类型	推荐数据量	最小数据量（数据条数）	单场景推荐训练数据量	单条数据Token长度限制
N1	微调	-	1000条/每场景	≥ 1万条/每场景	32K
N2	微调	-	1000条/每场景	≥ 1万条/每场景	32K

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4-2**。

表 4-2 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	创建原始数据集	数据集是指用于模型训练或评测的一组相关数据样本，上传至平台的数据将被创建为原始数据集进行统一管理。	创建原始数据集
	上线原始数据集	在正式发布数据集前，需要执行上线操作。	上线原始数据集
加工数据集	创建文本类数据集加工任务	数据集中若存在异常数据，可通过数据集加工功能去除异常字符、表情符号、个人敏感内容等。 **说明** 盘古NLP大模型仅支持接入文本类数据集。 若数据类型为文档、网页，则加工数据集为**必选项**，否则为**可选项**。	创建文本类数据集加工任务
	上线加工后的数据集	对加工后的数据集执行上线操作。	上线加工后的文本类数据集
标注数据集（可选）	创建文本类数据集标注任务	创建数据集标注任务，对数据集执行标注操作，标注后的数据可以用于模型训练或评测。	创建文本类数据集标注任务
	审核数据集标注结果	对数据集的标注结果进行审核	审核文本类数据集标注结果
	上线标注后的数据集	对标注后的数据集执行上线操作。	上线标注后的文本类数据集
评估数据集（可选）	创建文本类数据集评估标准	创建数据集评估标准。评估文本通顺性、信息充分性、内容有效性等。	创建文本类数据集评估标准
	创建文本类数据集评估任务	创建数据集质量评估任务，基于评估标注对数据逐一评估其质量。	创建文本类数据集评估任务
	获取数据集质量评估报告	查看数据集评估任务的进展和数据集质量。	获取文本类数据集评估报告
发布数据集	创建文本类数据集发布任务	创建发布数据集，并进行正式的发布操作，用于后续的训练任务。 平台支持发布的数据集格式为默认格式、盘古格式。 **训练盘古NLP大模型需选择发布格式为盘古格式。**	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大模型专门用于处理和理解人类语言。它能够执行多种任务，如对话问答、文案生成和阅读理解，同时具备逻辑推理、代码生成和插件调用等高级功能。

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。
- **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4-3**。

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4-4**。

表 4-3 不同系列 NLP 大模型对处理文本的长度差异

模型支持区域	模型名称	可处理最大Token长度	说明
西南-贵阳一	Pangu-NLP-N1-Chat-32K-20241030	32K	盘古NLP大模型，此版本是2024年10月发布的十亿级模型版本，支持8K训练，4K/32K推理。基于Snt9B3卡可单卡推理部署，此模型版本支持全量微调、LoRA微调、INT8量化、断点续训、在线推理和能力调测特性。单卡部署4K模型版本支持64并发，单卡部署32K模型版本支持32并发。
	Pangu-NLP-N1-Chat-128K-20241030	128K	此版本是2024年10月发布的十亿级模型版本，支持128K在线推理。基于Snt9B3卡支持8卡推理部署，此模型版本仅支持预置模型版本，不支持SFT后模型版本做128K推理部署。

模型支持区域	模型名称	可处理最大Token长度	说明
	Pangu-NLP-N2-Base-20241030	-	此版本是2024年10月发布的百亿级模型版本，支持模型增量预训练。基于Snt9B3卡支持32卡起训，预训练后的模型版本需要通过SFT之后，才可支持推理部署。
	Pangu-NLP-N2-Chat-32K-20241030	32K	此版本是2024年10月发布的百亿级模型版本，支持8K训练，4K/32K推理。基于Snt9B3卡可支持32卡起训，支持4卡推理部署，此模型版本支持全量微调、LoRA微调、INT8量化、断点续训、在线推理、能力调测、边缘部署特性。2卡部署4K模型版本支持4并发，4卡部署32K模型版本支持4并发。

表 4-4 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

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，包括预训练和微调，二者区别详见表4-5。

表 4-5 预训练和微调训练类型区别

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。

须知

当前平台提供的NLP大模型的训练方式仅支持微调，不支持预训练。

4.2.2 创建 NLP 大模型训练任务

创建 NLP 大模型微调任务

步骤1 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 4-1 进入操作空间

步骤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”。

步骤3 在“创建训练任务”页面，模型类型选择“NLP大模型”，训练类型选择“微调”。模型选择完成后，参考表4-6完成训练参数设置。

表 4-6 NLP 大模型微调参数说明

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	• 全量微调：在模型有监督微调过程中，对大模型的全部参数进行更新。这种方法通常会带来最优的模型性能，但需要大量的计算资源和时间，计算开销较高。 • 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。

参数分类	训练参数	参数说明
	基础模型	选择微调训练所用的基础模型， 可在“从资产选模型”或者“从人物选模型”中进行选择。
	高级设置	plog日志。plog日志是一种用来记录模型运行情况的信息。开启plog日志，能帮助开发者了解模型执行的状态、捕捉错误、分析问题。不同的日志级别表示日志的重要性和详细程度，从低到高依次是：DEBUG、INFO、WARNING、ERROR。
训练参数	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。
	训练轮数	指完成全部训练数据集训练的次数。
	学习率	学习率决定了每次训练时模型参数更新的幅度。选择合适的学习率非常重要：如果学习率太大，模型可能会无法收敛；如果学习率太小，模型的收敛速度会变得非常慢。
	优化器	优化器参数指的是用于更新模型权重的优化算法的相关参数，可以选择adamw。 ● adamw是一种改进的Adam优化器，它在原有的基础上加入了权重衰减（weight decay）的机制，可以有效地防止过拟合（overfitting）的问题。
	学习率衰减比率	学习率衰减后的比率，用于控制训练过程中学习率的下降幅度。经过衰减后，学习率的最低值由初始学习率和衰减比率决定。其计算公式为：最低学习率 = 初始学习率 * 学习率衰减比率。也就是说，学习率在每次衰减后不会低于这个计算出来的最低值。

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过程中逐渐增加学习率的过程。在训练的初始阶段，模型的权重通常是随机初始化的，此时模型的预测能力较弱。如果直接使用较大的学习率进行训练，可能会导致模型在初始阶段更新过快，从而影响模型的收敛。 为了解决这个问题，可以在训练的初始阶段使用较小的学习率，然后逐渐增加学习率，直到达到预设的最大学习率。这个过程就叫做热身比例。通过使用热身比例，可以避免模型在初始阶段更新过快，从而有助于模型更好地收敛。
	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。
	模型保存步数	指每训练一定数量的步骤（或批次）后，模型的状态就会被保存下来。 可以通过token_num = step * batch_size * sequence公式进行预估。其中： ● token_num：已训练的数据量。 ● step：已完成的训练步数。 ● batch_size：每个训练步骤中使用的样本数据量。 ● sequence：每个数据样本中的Token数量。 数据量以Token为单位。
	流水线并行微批次大小	在流水线并行处理中，通过合理设置并行程度，可以减少各阶段之间的空闲等待时间，从而提升整个流水线的效率。
	每个数据并行下的批处理大小	设置在并行训练中，每个微批次包含的数据批量大小，适当的数据批量大小能够确保训练各个阶段都能充分利用计算资源，提升并行效率。
数据配置	训练数据	选择训练模型所需的数据集。要求数据集经过发布操作，发布数据集操作方法请参见**发布数据集**。
资源配置	计费模式	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。
基本信息	名称	训练任务名称。

参数分类	训练参数	参数说明
	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。

步骤4 参数填写完成后，单击“立即创建”。

步骤5 创建好训练任务后，返回“模型训练”页面，单击操作列“启动”，并在任务确认弹窗中单击“确定”启动训练任务。

----结束

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指标、训练任务详情和训练日志。

查看模型训练状态

在模型训练列表中查看训练任务的状态，各状态说明详见表4-7。

表 4-7 训练状态说明

训练状态	训练状态含义
已发布	模型已经训练完成并进行发布，用户可以使用模型进行部署、推理操作。
训练完成	模型训练已经成功完成。
训练中	模型正在训练中，训练过程尚未结束。
训练失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已停止	模型训练已被用户手动停止。
停止中	模型训练正在停止中。
训练异常	模型训练过程中出现了非预期的异常情况，需查看日志定位训练异常原因。
待启动	模型训练任务已经创建，但尚未启动训练过程。
初始化	模型训练任务正在进行初始化配置，准备开始训练。

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，模型的训练指标介绍请参见表4-8。

图 4-2 查看训练指标

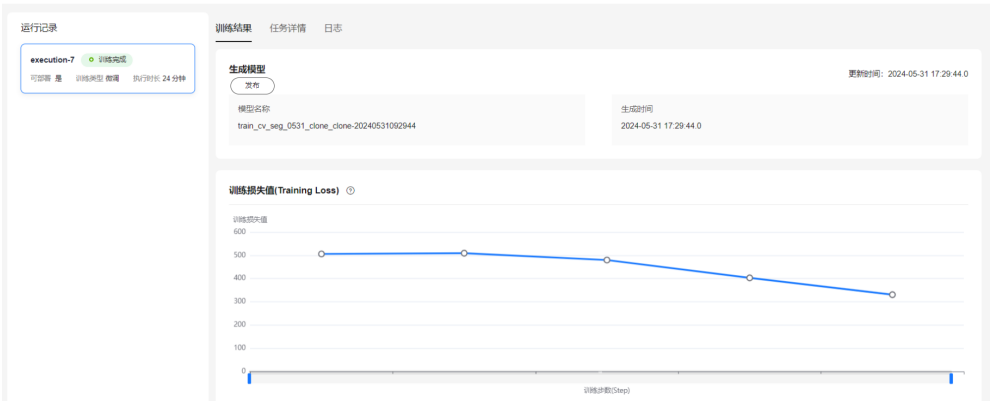


表 4-8 训练指标说明

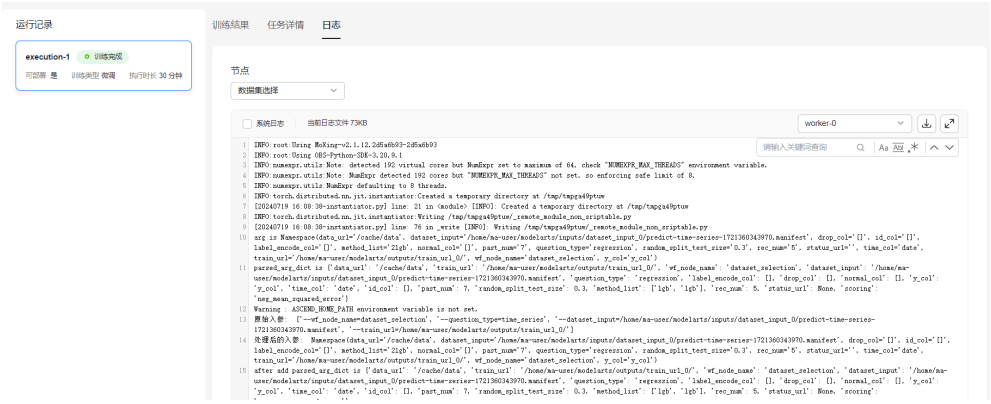
模型	训练指标	指标说明
NLP大模型	训练损失值	训练损失值是一种衡量模型预测结果和真实结果之间的差距的指标，通常情况下越小越好。 一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。
	困惑度	用来衡量大语言模型预测一个语言样本的能力，数值越低，准确率也就越高，表明模型性能越好。
	指标看板	• bleu-1：模型生成句子与实际句子在单字层面的匹配度，数值越高，表明模型性能越好。 • bleu-2：模型生成句子与实际句子在词组层面的匹配度，数值越高，表明模型性能越好。 • bleu-3：模型生成结果和实际句子的加权平均精确率，数值越高，表明模型性能越好。

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**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 4-3 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. 在模型训练列表页面选择训练完成的任务，单击训练任务名称进去详情页。
2. 在“训练结果”页面，单击“发布”。

图 4-4 训练结果页面

3. 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资产同步显示在“空间资产 > 模型”页面。

图 4-5 发布模型

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，进入所需操作空间。

图 4-6 进入操作空间

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checkpoints、训练参数、训练数据以及基本信息等。
   - 启动。单击操作列的“启动”，再单击弹窗的“确定”，可以启动训练任务。
   - 克隆。单击操作列的“更多 > 克隆”，可以复制当前训练任务。
   - 重试。单击操作列的“更多 > 重试”，可以编辑运行失败的节点，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。

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4-9**。

表 4-9 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。

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。

常见报错	问题现象	原因分析	解决方案
训练日志提示“ValueError: label_map not match”	训练日志中提示“ValueError: label_map not match”，并打印出标签数据，例如： ValueError: label_map not match. {1:'apple', 2:'orange', 3:'banana', 4:'pear'} & {1:'apple', 2:'orange', 3:'banana'}	训练集中的标签个数与验证集中的个数不一致，导致该错误发生。 例如，训练集中的标签共有4个，验证集中的标签只有3个。	请保持数据中训练集和验证集的标签数量一致。

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。当前仅支持对NLP大模型进行压缩。采用的压缩方式是INT8，INT8量化压缩可以显著减小模型的存储大小，降低功耗，并提高计算速度。

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 4-7 进入操作空间

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”。参考表4-10创建模型压缩任务。

表 4-10 模型压缩任务参数说明

参数类别	参数名称	说明
压缩配置	压缩模型	选择需要进行压缩的模型，可使用来自资产的模型或任务的模型。
	压缩策略	例如，可使用INT8压缩策略，同等QPS目标下，INT8可以降低推理显存占用。
基本信息	任务名称	模型压缩任务的名称。
	描述	模型压缩任务的描述。

3. 参数填写完成后单击“立即创建”创建模型压缩任务。

4.4 部署 NLP 大模型

4.4.1 创建 NLP 大模型部署任务

模型训练完成后，可以启动模型的部署操作。

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 4-8 进入操作空间

2. 在左侧导航栏中选择“模型开发 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，模型类型选择“NLP大模型”，参考表4-11完成部署参数设置，启动模型部署。

表 4-11 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。
	部署方式	云上部署：算法部署至平台提供的资源池中。
	最大TOKEN长度	模型可最大请求的上下文TOKEN数。
	架构类型	算法所支持的结构类型，模型选择完成后，会自动适配架构类型。
安全护栏	选择模式	安全护栏保障模型调用安全。若关闭，推理服务可能会有违规风险，建议开启。
	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，不可调整。
资源配置	实例数	设置部署模型时所需的实例数，单次部署服务时，部署实例个数建议不大于10，否则可能触发限流导致部署失败。
基本信息	名称	设置部署任务的名称。

参数分类	部署参数	参数说明
	描述（可选）	设置部署任务的描述。

4. 参数填写完成后，单击“立即部署”。

说明

您可以选择预置模型进行部署，部署时默认开通安全护栏权限。

4.4.2 查看 NLP 大模型部署任务详情

部署任务创建成功后，可以在“模型开发 > 模型部署”页面查看模型的部署状态。

当状态依次显示为“初始化 > 部署中 > 运行中”时，表示模型已成功部署，可以进行调用。

此过程可能需要较长时间，请耐心等待。在此过程中，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 4-9 部署详情

4.4.3 管理 NLP 大模型部署任务

模型更新、修改部署

成功创建部署任务后，如需修改已部署的模型或配置信息，可以在详情页面单击右上角的“模型更新”或“修改部署”进行调整。更新模型时可以替换模型，但在修改部署时模型不可替换。

在“模型更新”或“修改部署”后进行升级操作时，可选择全量升级或滚动升级两种方式：

- **全量升级**：新旧版本的服务同时运行，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。
- **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 4-10 模型更新

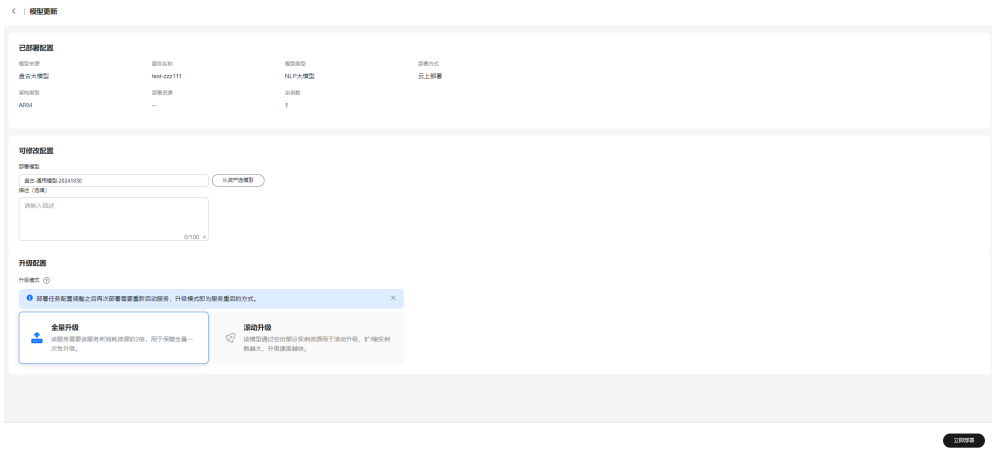


图 4-11 修改部署

4.5 调用 NLP 大模型

4.5.1 使用“能力调测”调用 NLP 大模型

平台提供的“能力调测”功能支持用户直接调用预置模型或经过训练的模型。使用该功能前，需完成模型的部署操作，详见**创建NLP大模型部署任务**。

NLP大模型支持文本对话能力，在输入框中输入问题，模型就会返回对应的答案内容。

图 4-12 调测 NLP 大模型

表 4-12 NLP 大模型能力调测参数说明

参数	说明
温度	用于控制生成文本的多样性和创造力。调高温度会使得模型的输出更多样性和创新性。
核采样	控制生成文本多样性和质量。调高核采样可以使输出结果更加多样化。
最大口令限制	用于控制聊天回复的长度和质量。
话题重复度控制	用于控制生成文本中的重复程度。调高参数模型会更频繁地切换话题，从而避免生成重复内容。
词汇重复度控制	用于调整模型对频繁出现的词汇的处理方式。调高参数会使模型减少相同词汇的重复使用，促使模型使用更多样化的词汇进行表达。
历史对话保留轮数	选择“多轮对话”功能时具备此参数。表示系统能够记忆的历史对话数。

4.5.2 使用 API 调用 NLP 大模型

模型部署成功后，可以通过“文本对话”API调用NLP大模型。

表 4-13 NLP 大模型 API 清单

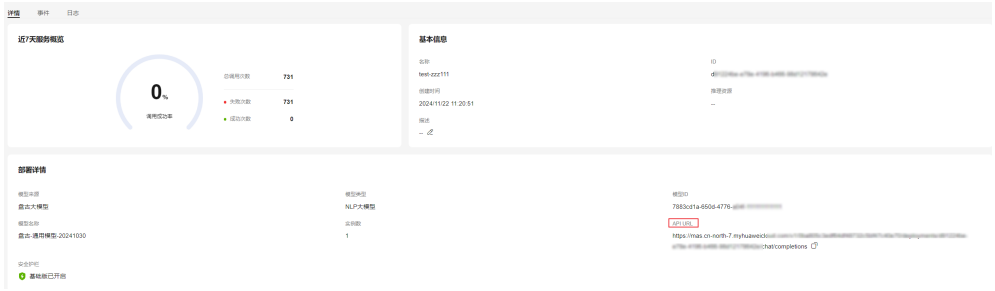
API分类	API访问路径（URI）
文本对话	/v1/{project_id}/deployments/{deployment_id}/chat/completions

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 4-13 进入操作空间

2. 单击左侧“模型开发 > 模型部署”。
   - 调用已部署的模型。单击状态为“运行中”的模型名称，在“详情”页签，可获取API的URL。

     图 4-14 获取已部署模型的调用路径

   - 调用预置服务。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”弹窗获取调用路径。

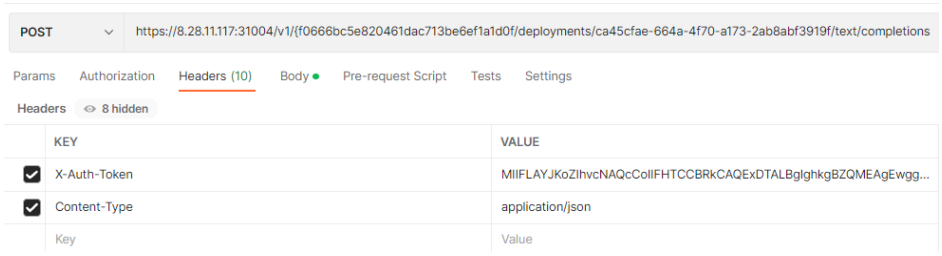
     图 4-15 获取预置服务调用路径

3. 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。

4. 在Postman中新建POST请求，并填入API请求地址。
5. 参考图4-16填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**步骤3**中获取的Token值。

   图 4-16 填写 NLP 大模型 API



   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
}
```

6. 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。

4.5.3 统计模型调用信息

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，进入所需操作空间。

   图 4-17 进入操作空间

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**。

构建科学计算大模型训练数据要求

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：

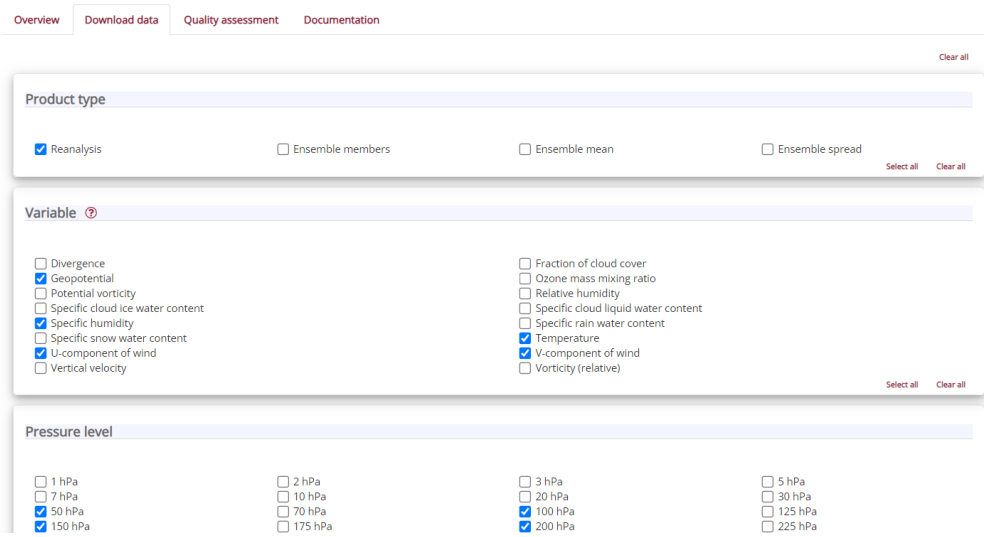
   a. 注册并登录数据下载平台，在高空变量数据下载链接中：

      - 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



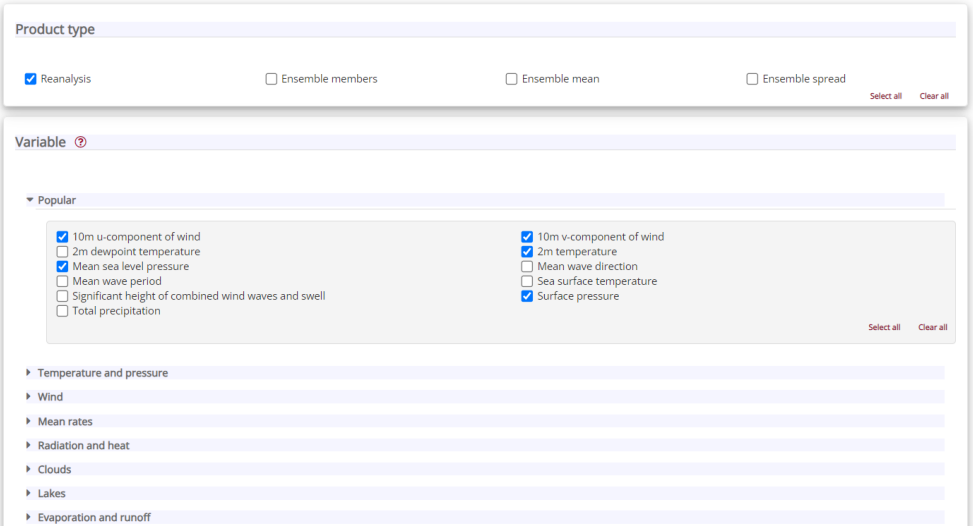
2. 示例二：以下载2021年7月16日表面变量数据为例，下载内容为表面变量（10m u风、10m v风、2米温度、海平面气压）0点、6点、12点、18点时刻的数据文件，下载步骤示例如下：

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	创建原始数据集	数据集是指用于模型训练或评测的一组相关数据样本，上传至平台的数据将被创建为原始数据集进行统一管理。	创建原始数据集
	上线原始数据集	在正式发布数据集前，需要执行上线操作。	上线原始数据集
加工数据集（可选）	创建气象类数据集加工任务	数据集中若存在异常数据，可通过数据集加工功能去除异常字符、表情符号、个人敏感内容等。	创建气象类数据集加工任务
	上线加工后的数据集	对加工后的数据集执行上线操作。	上线加工后的文本类数据集
发布数据集	创建气象类数据集发布任务	创建发布数据集，并进行正式的发布操作，用于后续的训练、评测任务。	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

科学计算大模型主要用于。

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：中期天气要素预测模型、区域中期海洋智能预测模型。

- **中期天气要素预测模型选择建议**：

  科学计算大模型的中期天气要素预测模型，可以对未来一段时间的天气进行预测，具备以下优势：

  - 高时间精度：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。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：中期天气要素预测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：中期天气要素预测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。

  中期天气要素预测模型信息见**表5-3**。

表 5-3 中期天气要素预测模型信息

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
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

该模型类型主要用于天气基础要素预测，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-20241030**模型。

表 5-4 中期天气要素预测模型的类型

模型支持区域	模型名称	使用场景	说明
西南-贵阳一	Pangu-AI4S-Weather_1h-20241030	用于天气基础要素预测，时间分辨率为1小时。	支持预训练、微调、在线推理、能力调测特性，基于Snt9B33，支持1个训练单元训练及1个推理单元部署。

模型支持区域	模型名称	使用场景	说明
	Pangu-AI4S-Weather_3h-20241030	用于天气基础要素预测，时间分辨率为3小时。	支持预训练、微调、在线推理、能力调测特性，基于Snt9B3，支持1个训练单元训练及1个推理单元部署。
	Pangu-AI4S-Weather_6h-20241030	用于天气基础要素预测，时间分辨率为6小时。	支持预训练、微调、在线推理、能力调测特性，基于Snt9B3，支持1个训练单元训练及1个推理单元部署。
	Pangu-AI4S-Weather_24h-20241030	用于天气基础要素预测，时间分辨率为24小时。	支持预训练、微调、在线推理、能力调测特性，基于Snt9B3，支持1个训练单元训练及1个推理单元部署。

- **区域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。区域中期海洋智能预测模型当前主要包括区域海洋要素模型，信息见表5-5。

  表 5-5 区域中期海洋智能预测模型信息

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
区域海洋要素模型	0m, 6m, 10m, 20m, 30m, 50m, 70m, 100m, 125m, 150m, 200m, 250m, 300m, 400m, 500m	T：海温(℃) S：海盐(PSU) U：海流经向速率(ms-1) V：海流纬向速率(ms-1)	SSH：海表高度(m)	24h	1/12°	特定区域

该模型类型主要用于区域海洋基础要素预测，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

模型支持区域	模型名称	使用场景	说明
西南-贵阳一	Pangu-AI4S-Ocean_Regional_24h-20241030	用于区域海洋基础要素预测	支持预训练、微调、在线推理、能力调测特性，基于Snt9B3支持1个训练单元训练及1个推理单元部署。

科学计算大模型训练类型选择建议

- **中期天气要素预测模型的训练类型选择建议：**

  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。

- **区域中期海洋智能预测模型的训练类型选择建议**：

  区域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训练任务步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需操作空间。

图 5-3 进入操作空间

2. 在左侧导航栏中选择“模型开发 > 模型训练”，单击界面右上角“创建训练任务”。
3. 在“创建训练任务”页面，模型类型选择“科学计算大模型”。模型选择完成后，参考**表5-7**、**表5-8**完成训练参数设置，启动模型训练。

表 5-7 科学计算大模型（中期天气要素预测）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	可选择“预训练”和“微调”。
	基础模型	可以选择“从资产选模型”和“从任务选模型”，模型会自带时间分辨率，会根据预设的时间间隔处理和生成预测结果。 ● 若训练类型为“预训练”，训练任务使用训练数据重新训练出与基础模型分辨率相同的模型。 ● 若训练类型为“微调”，训练任务会使用训练数据在基础模型的基础上进行训练。

参数分类	参数名称	参数说明
	plog日志	plog日志。plog日志是一种用来记录模型运行情况的信息。开启plog日志，能帮助开发者了解模型执行的状态、捕捉错误、分析问题。不同的日志级别表示日志的重要性和详细程度，从低到高依次是：DEBUG、INFO、WARNING、ERROR。
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]。

参数分类	参数名称	参数说明
	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。
	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	用于定义权重衰减的系数。权重衰减是一种正则化技术，可以防止模型过拟合。取值需≥0。
	学习率	用于定义学习率的大小。学习率决定了模型参数在每次更新时变化的幅度。如果学习率过大，模型可能会在最优解附近震荡而无法收敛。如果学习率过小，模型收敛的速度可能会非常慢。当batch_size减小时，学习率也应相应地线性减小。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。

参数分类	参数名称	参数说明
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。
	高空层次	设置训练数据的高空层次信息，在“预训练”的场景中也支持您添加或去除新的高空层次，训练任务会根据您配置的高空层次对模型重新进行训练。
	高空变量	设置训练数据的高空变量信息，在“预训练”的场景中也支持您添加或去除新的高空变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的高空变量对模型重新进行训练。
	表面变量	设置训练数据的表面变量信息，同时在“预训练”的场景中也支持您添加或去除新的表面变量，选择后会在变量权重中增加或去除该变量权重，训练任务会根据您配置的表面变量对模型重新进行训练。
	表面静态量	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。
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。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

表 5-8 科学计算大模型（区域中期海洋智能预测）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	可根据科学计算大模型适用场景和建议选择“预训练”和“微调”。
	基础模型	可以选择“预置模型”和“我的模型”，模型会自带时间分辨率，会根据预设的时间间隔处理和生成预测结果。 ● 若训练类型为“预训练”，训练任务使用训练数据重新训练出与基础模型分辨率相同的模型。 ● 若训练类型为“微调”，训练任务会使用训练数据在基础模型的基础上进行训练。
	plog日志	plog日志。plog日志是一种用来记录模型运行情况的信息。开启plog日志，能帮助开发者了解模型执行的状态、捕捉错误、分析问题。不同的日志级别表示日志的重要性和详细程度，从低到高依次是：DEBUG、INFO、WARNING、ERROR。
	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。
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)
区域范围	/	在图中设置训练模型的经纬度范围，即区域模型的经纬度范围。该范围需要在上传区域数据的范围之内。
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)。

参数分类	参数名称	参数说明
正则化参数	路径删除概率	用于定义路径删除机制中的删除概率。路径删除是一种正则化技术，它在训练过程中随机删除一部分的网络连接，以防止模型过拟合。这个值越大，删除的路径越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0，1）。
	特征删除概率	用于定义特征删除机制中的删除概率。特征删除（也称为特征丢弃）是另一种正则化技术，它在训练过程中随机删除一部分的输入特征，以防止模型过拟合。这个值越大，删除的特征越多，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1)。
	给输入数据加噪音的概率	定义了给输入数据加噪音的概率，定义了给输入数据加噪音的概率。加噪音是一种正则化技术，它通过在输入数据中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输入数据加噪音的尺度	给输入数据加噪音的尺度，定义了给输入数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
	给输出数据加噪音的概率	给输出数据加噪音的概率，定义了给输出数据加噪音的概率。加噪音是一种正则化技术，它通过在模型的输出中添加随机噪音来增强模型的泛化能力。取值范围：[0,1]。
	给输出数据加噪音的尺度	给输出数据加噪音的尺度，定义了给输出数据加噪音的尺度。这个值越大，添加的噪音越强烈，模型的正则化效果越强，但同时也可能会降低模型的拟合能力。取值范围：[0,1]。
优化器参数	优化器种类	优化器种类。优化器是用于更新模型参数的算法，目前支持ADAM优化器。
	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]。

参数分类	参数名称	参数说明
	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	用于定义权重衰减的系数。权重衰减是一种正则化技术，可以防止模型过拟合。取值需≥0。
	学习率	用于定义学习率的大小。学习率决定了模型参数在每次更新时变化的幅度。如果学习率过大，模型可能会在最优解附近震荡而无法收敛。如果学习率过小，模型收敛的速度可能会非常慢。当batch_size减小时，学习率也应相应地线性减小。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	T	海表面2m温度 (℃)的权重设置。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。
	U	海表面10m经向风速(ms-1)的权重设置。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。
	V	海表面10m纬向风速(ms-1)的权重设置。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。
	P	平均海平面气压(Pa)的权重设置。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。
	SSH	海表面高度(m)的权重设置。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。
	SP	海表面气压 (Pa)的权重设置。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。

参数分类	参数名称	参数说明
	WT	深海层海温(℃)的权重设置。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。
	WU	深海层海流经向速率 (ms-1)的权重设置。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。
	WV	深海层海流经纬向速率 (ms-1)的权重设置。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。
	WS	深海层海盐(PSU)的权重设置。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。
基本信息	名称	训练任务名称。
	描述	训练任务描述。

4. 填写训练任务“名称”、“描述”，单击“立即创建”创建科学计算大模型训练任务。
5. 创建好训练任务后，返回“模型训练”页面，单击操作列“启动”，并在任务确认弹窗中单击“确定”启动训练任务。

5.2.3 查看科学计算大模型训练状态与指标

查看模型训练状态

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指标、训练任务详情和训练日志。

表 5-9 训练状态说明

训练状态	训练状态含义
已发布	模型已经训练完成并进行发布，用户可以使用模型进行部署、推理操作。
训练完成	模型训练已经成功完成。
训练中	模型正在训练中，训练过程尚未结束。

训练状态	训练状态含义
训练失败	模型训练过程中出现错误，需查看日志定位训练失败原因。
已停止	模型训练已被用户手动停止。
停止中	模型训练正在停止中。
训练异常	模型训练过程中出现了非预期的异常情况，需查看日志定位训练异常原因。
待启动	模型训练任务已经创建，但尚未启动训练过程。
初始化	模型训练任务正在进行初始化配置，准备开始训练。

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5-10**。

图 5-4 查看训练指标

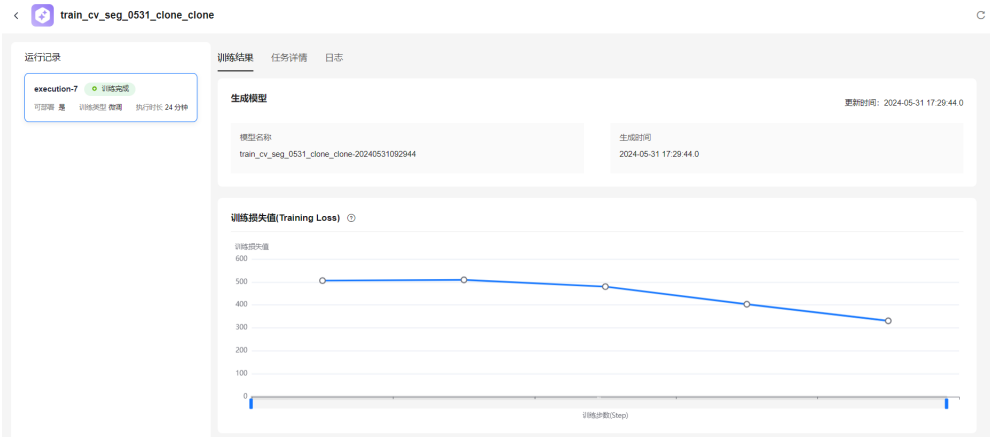


表 5-10 训练指标说明

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。

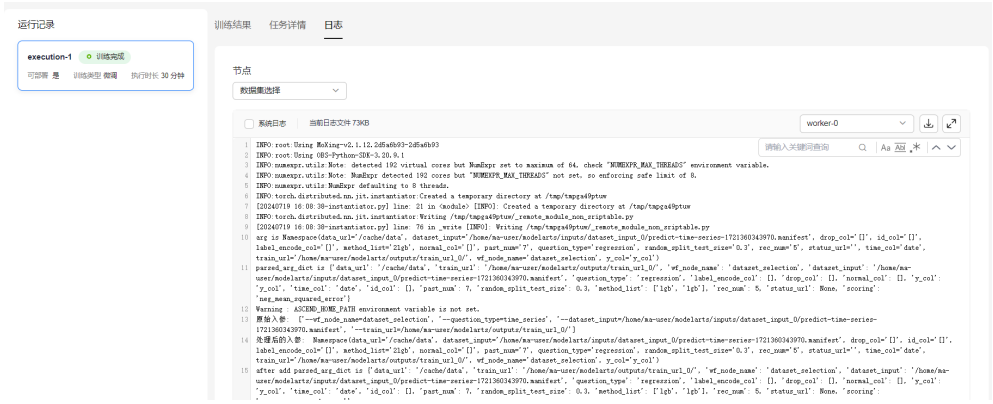
模型	训练指标	指标说明
	高空Loss（深海Loss）	高空Loss（深海Loss）是衡量模型在高空层次变量或在深海变量预测结果与真实结果之间差距的指标。该值越小，表示模型在高空（深海）变量的预测精度越高。
	表面Loss（海表Loss）	表面Loss（海表Loss）是衡量模型在表面层次变量或在海表变量预测结果与真实结果之间差距的指标。该值越小，表示模型在表面（海表）变量的预测精度越高。
	RMSE	均方根误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的平方误差的平均值的平方根。该值越小，代表模型性能越好。
	MAE	平均绝对误差，衡量预测值与真实值之间差距的指标。它是所有单个观测的绝对误差的平均值。该值越小，代表模型性能越好。
	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 5-5 获取训练日志

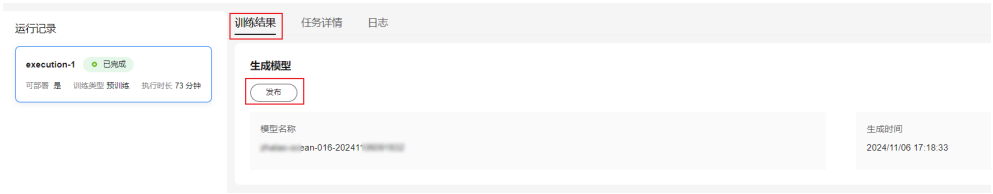


5.2.4 发布训练后的科学计算大模型

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. 在模型训练列表页面选择训练完成的任务，单击训练任务名称进去详情页。
2. 在“训练结果”页面，单击“发布”。

图 5-6 训练结果



3. 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资产同步显示在“空间资产 > 模型”页面。

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checkpoints、训练参数、训练数据以及基本信息等。
   - 启动。单击操作列的“启动”，再单击弹窗的“确定”，可以启动训练任务。
   - 克隆。单击操作列的“更多 > 克隆”，可以复制当前训练任务。
   - 重试。单击操作列的“更多 > 重试”，可以编辑运行失败的节点，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5-11。

表 5-11 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。
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

模型训练完成后，可以启动模型的部署操作。

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 > 模型部署”，单击界面右上角“创建部署”。
3. 在“创建部署”页面，模型类型选择“科学计算大模型”，参考表5-12完成部署参数设置，启动模型部署。

表 5-12 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”。
	场景	选择模型场景，分为“全球天气要素预测”、“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”、“全球中期海量智能预测”。 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。 如果使用全球中期降水预测模型，需要选择1个平台预置好的全球中期降水预测模型，并选择对应的全球中期天气要素预测模型。并且至少有一个中期天气要素模型时间分辨率要小于等于降水模型时间分辨率。
	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。 边缘部署：算法部署至客户的边缘设备中。
	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。
	架构类型	算法所支持的结构类型，模型选择完成后，会自动适配架构类型。
资源配置	实例数	设置部署模型是所需的实例数，单次部署服务时，部署实例个数建议不大于10，否则可能触发限流导致部署失败。
基本信息	名称	设置部署任务的名称。
	描述（可选）	设置部署任务的描述。

4. 参数填写完成后，单击“立即部署”。

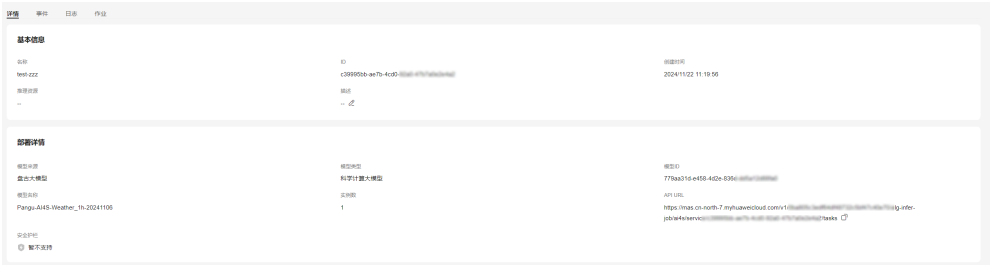
5.3.2 查看科学计算大模型部署任务详情

部署任务创建成功后，可以在“模型开发 > 模型部署”页面查看模型的部署状态。

当状态依次显示为“初始化 > 部署中 > 运行中”时，表示模型已成功部署，可以进行调用。

此过程可能需要较长时间，请耐心等待。在此过程中，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 5-7 部署详情



5.3.3 管理科学计算大模型部署任务

模型更新、修改部署

成功创建部署任务后，如需修改已部署的模型或配置信息，可以在详情页面单击右上角的“模型更新”或“修改部署”进行调整。更新模型时可以替换模型和修改作业配置参数，但在修改部署时模型不可替换或修改作业配置参数。

在“模型更新”或“修改部署”后进行升级配置操作。平台支持全量升级方式：新旧版本的服务同时运行，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。

图 5-8 模型更新

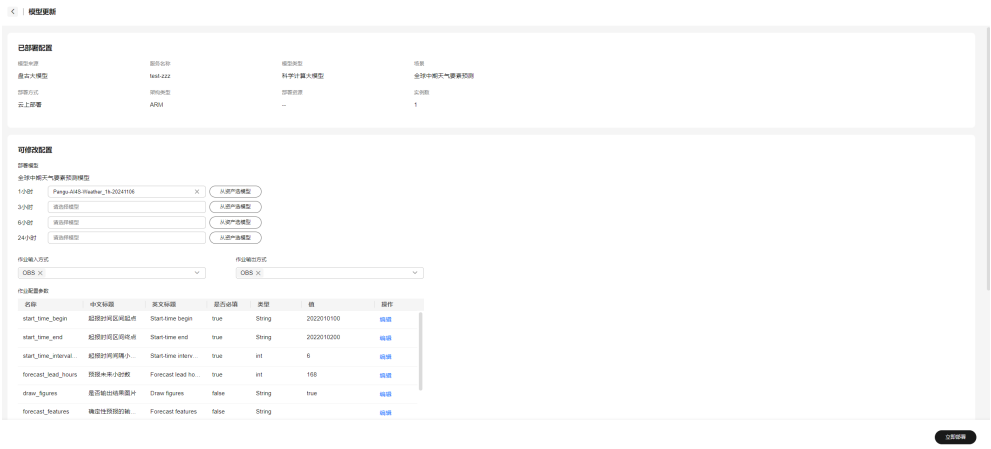
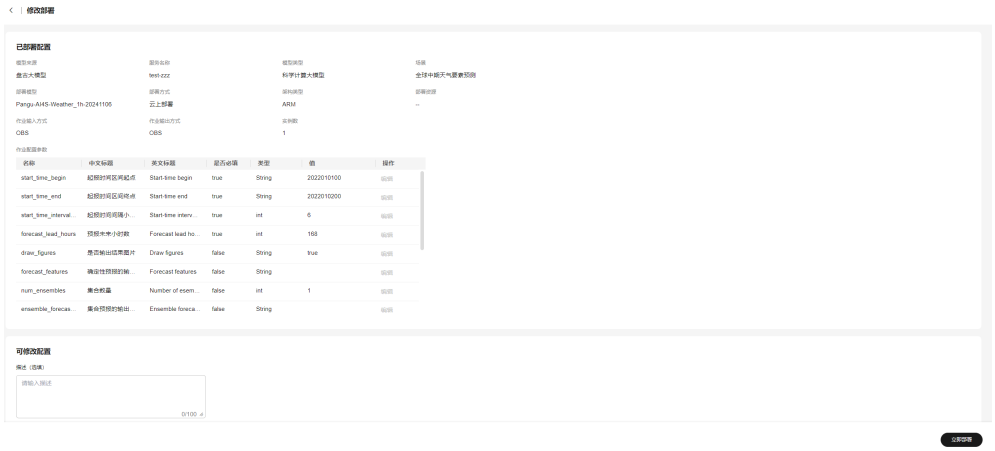


图 5-9 修改部署



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

平台提供的“能力调测”功能支持用户直接调用预置模型或经过训练的模型。使用该功能前，需完成模型的部署操作，详见**创建科学计算大模型部署任务**。

科学计算大模型支持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、全球海浪高度预测能力，在选择好模型后，根据需求选择相应的数据和模型配置信息，模型就会返回相应的预测结果。

表 5-13 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。
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。

参数	说明
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。
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。

参数	说明
输出设置	用于选择是否输出图片结果。

表 5-14 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。不同模型的指标已页面展示为准。
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。

图 5-10 调测科学计算大模型-1（天气/降水预测）

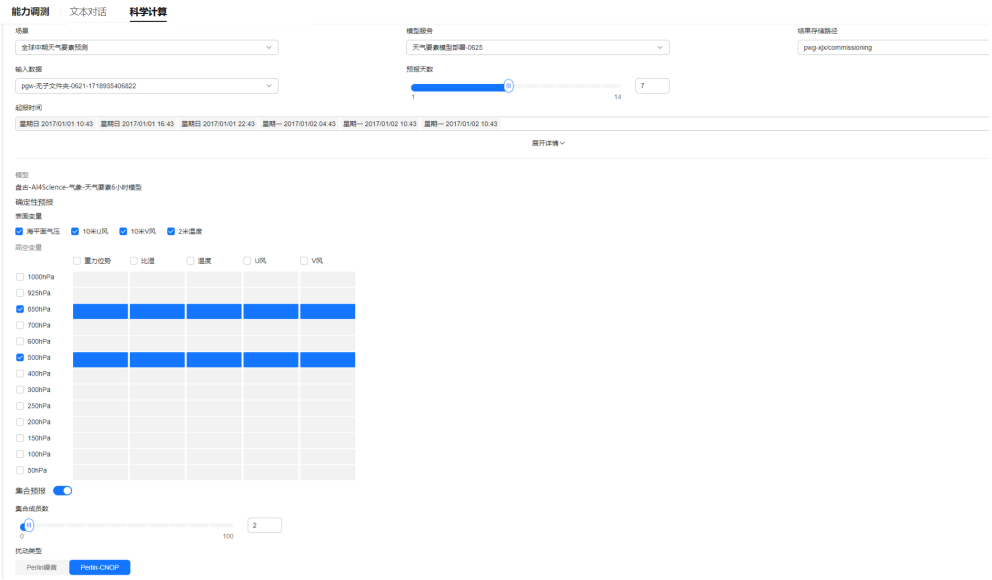


图 5-11 调测科学计算大模型-2（天气/降水预测）

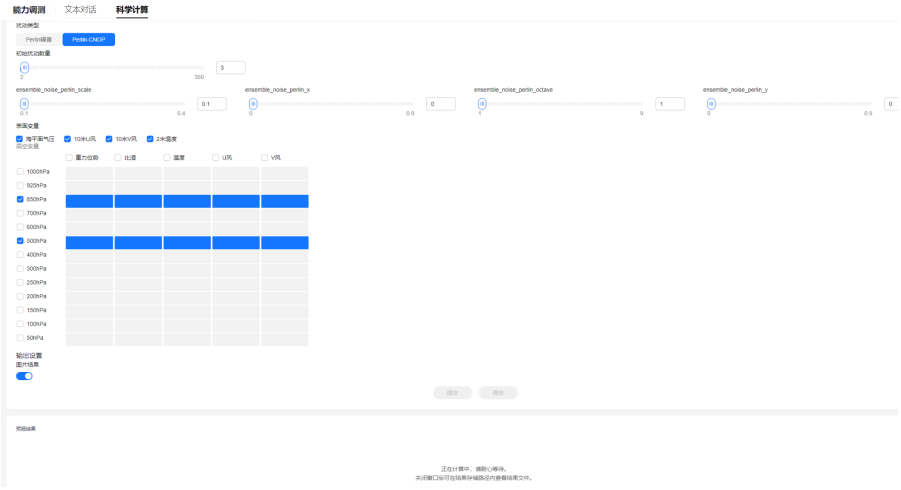
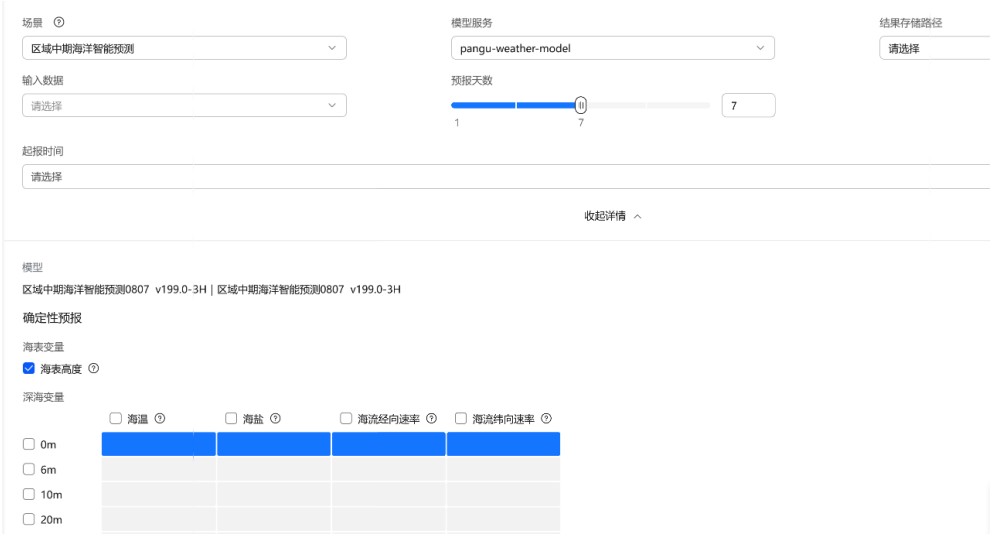


图 5-12 调测科学计算大模型（海洋类预测）

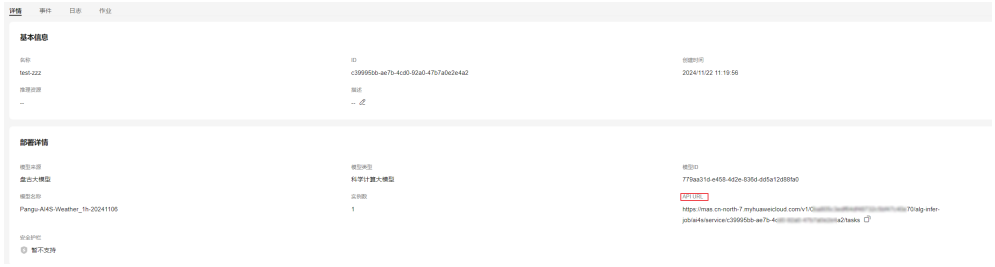


5.4.2 使用 API 调用科学计算大模型

使用API调用科学计算大模型步骤如下：

1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 单击左侧“模型开发 > 模型部署”。
   - 若调用已部署的模型，单击状态为“运行中”的模型名称，在“详情”页签，可获取API的URL。

     图 5-13 获取已部署模型的调用路径

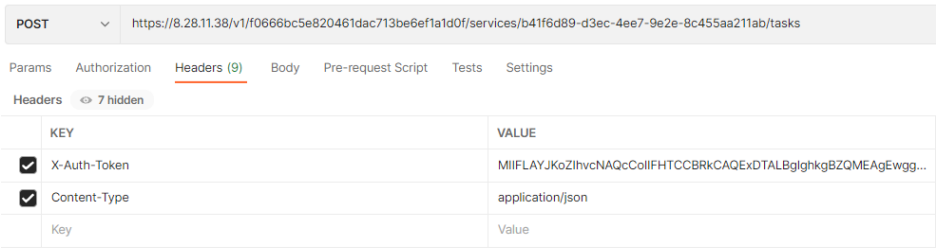
   - 若调用预置服务，在“预置服务”页签中，选择所需调用的科学计算大模型，单击“调用路径”，在“调用路径”弹窗获取调用路径。

     图 5-14 获取预置服务调用路径

3. 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。
4. 在Postman中新建POST请求，并填入步骤2的API请求地址。
5. 参考图5-15填写2个请求Header参数。
   - 参数名为Content-Type，参数值为application/json。
   - 参数名为X-Auth-Token，参数值为**步骤3**中获取的Token值。

   图 5-15 填写科学计算大模型 API

6. 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
   }
}
```

7. 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。
8. 在Postman中新建一个GET请求，填入域名（将步骤2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 开发盘古大模型提示词工程

6.1 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。

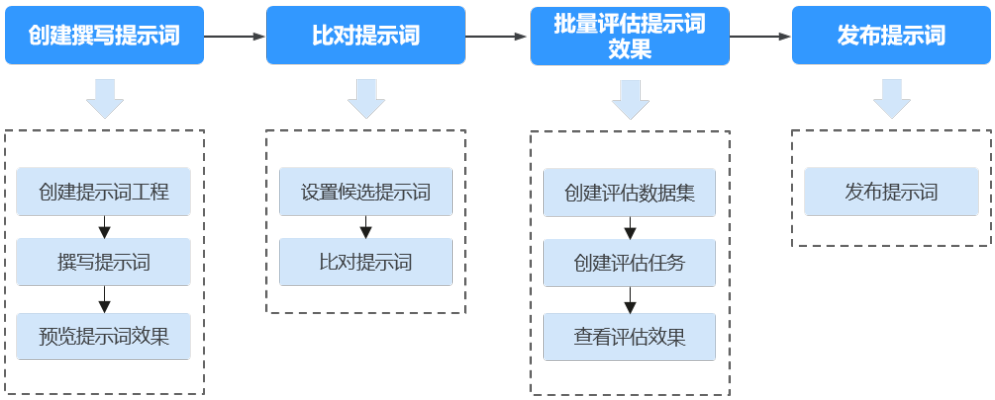
提示词工程使用流程

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 6-1 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 6-1 提示词工程使用流程



6.2 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。

6.3 撰写提示词

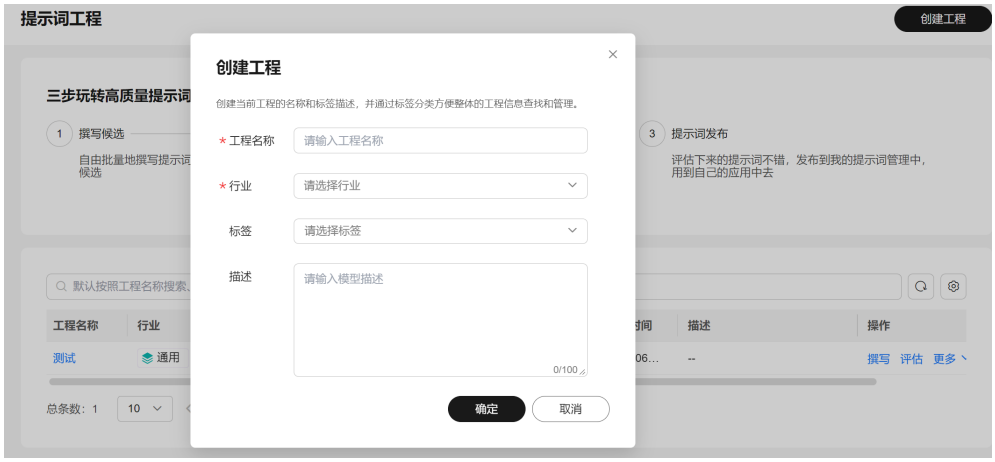
6.3.1 创建提示词工程

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 6-2 创建提示词工程



6.3.2 撰写所需提示词

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。

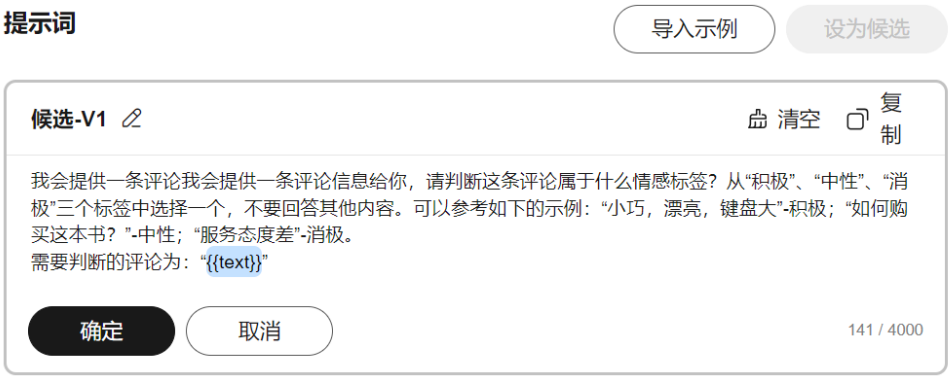
同时，撰写提示词过程中，可以通过设置模型参数来控制模型的生成行为，如调整温度、核采样、最大Token限制等参数。模型参数的设置会影响模型的生成质量和多样性，因此需要根据不同的场景进行选择。

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 6-3 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 6-4 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 6-5 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 6-6 模型设置

6.3.3 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 6-7 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。

6.4 横向比较提示词效果

6.4.1 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。

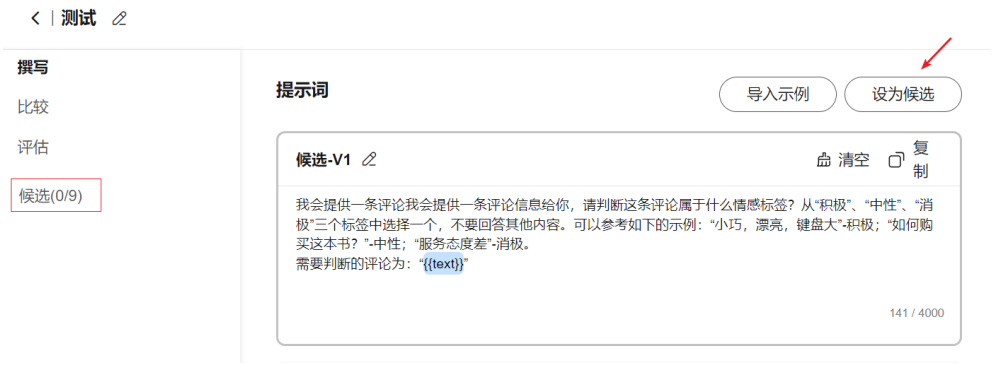
1. 登录ModelArts Studio大模型开发平台，进入所需空间。
2. 在左侧导航栏中选择“Agent 开发 > 提示词工程 > 提示词开发”。
3. 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 6-8 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 6-9 设为候选



6.4.2 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”。

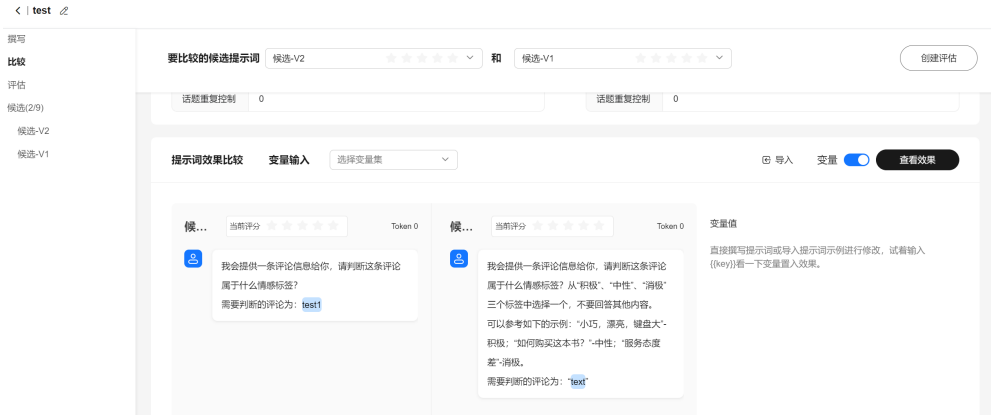
图 6-10 提示词工程

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 6-11 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 6-12 横向比对提示词效果



6.5 批量评估提示词效果

6.5.1 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 6-13 数据参考格式

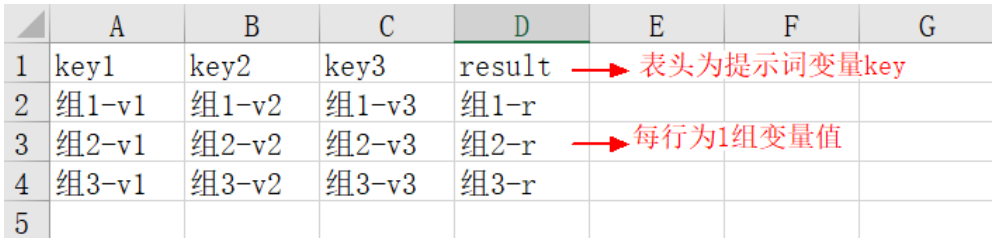
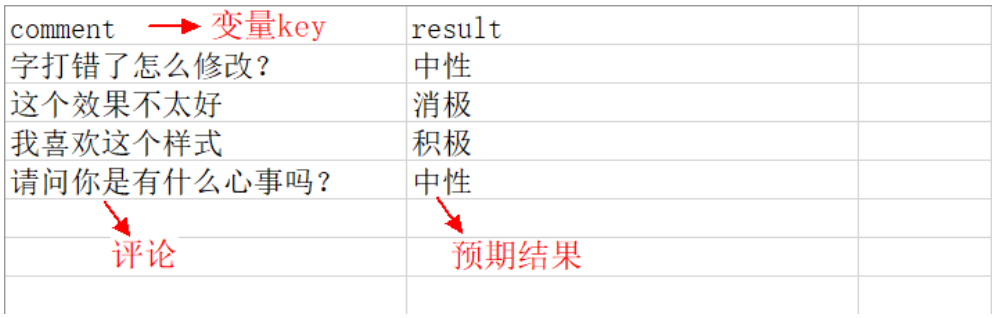


图 6-14 数据示例



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 6-15 提示用例管理

3. 在“创建数据集”页面完成数据集的上传。

图 6-16 创建提示词评估数据集

6.5.2 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 6-17 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 6-18 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。

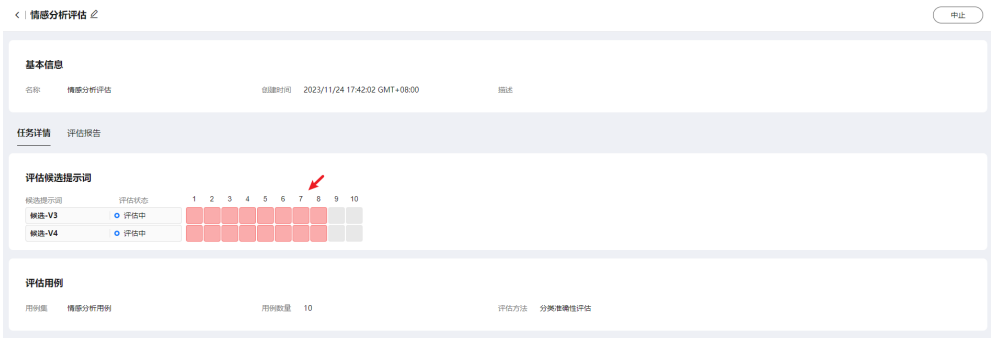
6.5.3 查看提示词评估结果

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 6-19 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6-20中有10条评估用例，当前已评估8条，剩余2条待评估。

图 6-20 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。

6.6 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 6-21 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。

7 开发盘古大模型 Agent 应用

7.1 Agent 开发平台概述

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组件，包含LLM、自定义代码、分支等组件，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。、
  - 内置插件：平台集成了各种类型的插件，包含搜索、图片理解等。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和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。

- 全链路信息调测评估：平台提供对Agent执行过程的全链路信息观测与调试调优，通过对信息的分层分析和展示，为开发者提供了AI应用在不同层级的运行情况指导和操作，提升观测和调试效率。

Agent 开发平台应用场景

当前，基于Agent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、工具、规划模式等信息，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组件进行“拖拉拽”即可搭建出任务流程，场景的组件包括LLM节点、Code节点、Branch节点等，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。

7.2 手工编排 Agent 应用

7.2.1 手工编排 Agent 应用流程

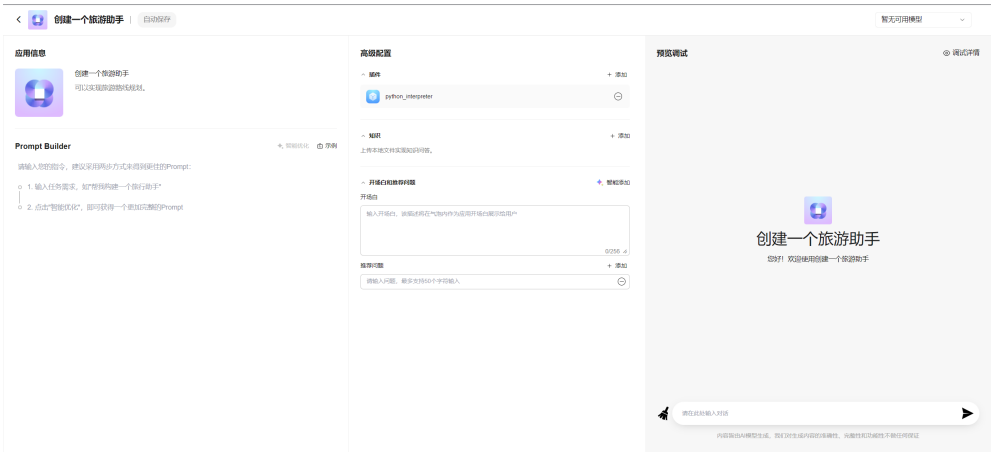
手工编排Agent应用流程步骤如下：

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”，将跳转至Agent开发平台。
3. 单击左侧导航栏“工作台”，在“应用”页签，单击右上角“创建应用”。

   您也可以鼠标单击已有应用右上角的 ···，进行应用的复制、删除、复制ID操作。
4. 在“创建应用”窗口中，填写应用名称与应用描述，单击左下角的图片可更换应用图标，单击“确定”，进入应用详情页面。

图 7-1 填写应用名称与应用描述

图 7-2 创建应用



5. 配置Prompt builder，详见**配置Prompt builder**。
6. 配置插件，详见**配置插件**。
7. 配置知识，详见**配置知识**。
8. 配置对话，详见**配置开场白和推荐问题**。
9. 调试Agent应用，详见**调试Agent应用**。

说明

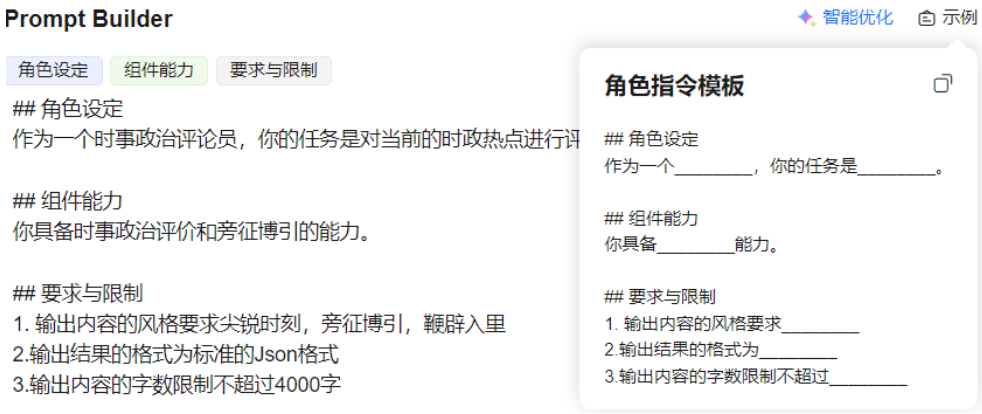
Agent应用支持的模型类型为**NLP大模型**。

7.2.2 配置 Prompt builder

创建Agent的首要步骤就是撰写提示词（Prompt），为Agent设定人设、目标、核心技能、执行步骤。Agent会根据LLM对提示词的理解，来选择使用插件或知识库，响应用户问题。因此，一个好的提示词可以让LLM更好的理解并执行任务，Agent效果与提示词息息相关。

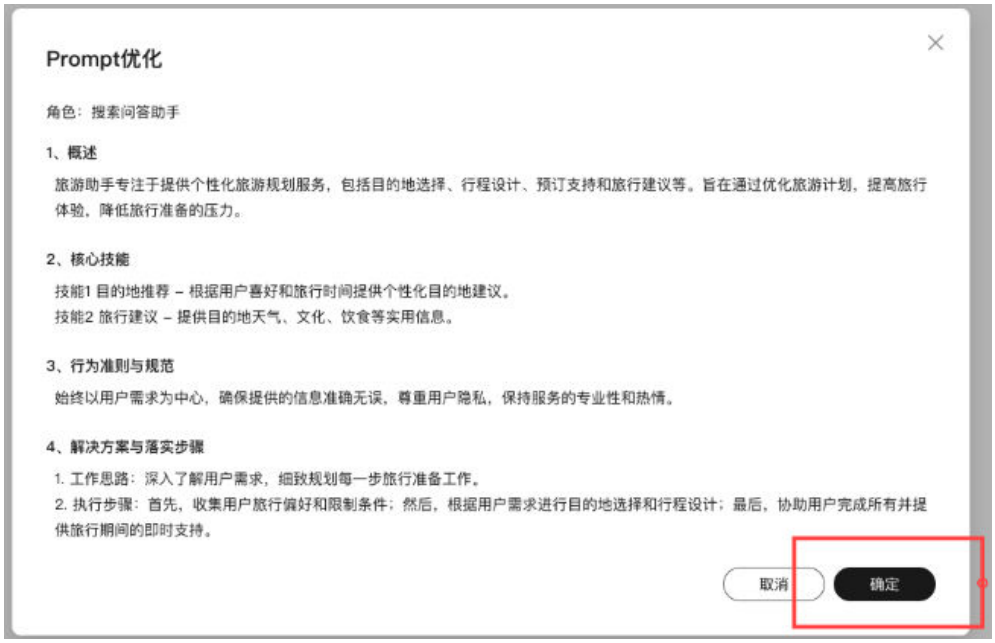
1. 在应用详情页面的“Prompt builder”模块中，需要填入prompt指令，单击“示例”，复制角色指令模板。在输入框中粘贴模板，并进行填空。

图 7-3 Prompt builder



2. 单击“智能优化”，在 “Prompt优化”窗口中单击“确定”。

图 7-4 Prompt 优化示例



7.2.3 配置插件

配置插件的步骤如下：

1. 在“高级配置 > 插件”，单击“添加”。

图 7-5 配置插件

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，最后单击“确定”。若想创建插件可单击右上角“创建插件”，创建插件的步骤请参见**创建插件**。

图 7-6 添加插件

3. 添加插件后，可在“高级配置”中查看当前已添加的插件。

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，进入所需空间。
2. 在左侧导航栏中选择“Agent开发”，将跳转至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. 配置插件的基本信息，输入插件名称和插件描述等信息。配置完成后，单击“下一步”。

   输入插件名称后，支持手动上传插件的头像。

   表 7-1 插件基本信息表

参数名称	说明
插件名称	待创建插件的名称。 名称必须以中文或者英文开头。 插件名称长度为1 ~ 200个字符，并且字符只允许为下面的类型： ● 中文 ● 字母（A-Z或a-z） ● 数字（0-9） ● 特殊字符：_和- ● 空格
插件描述	待创建插件的功能描述。 插件描述的长度为1 ~ 1600个字符。

5. 配置插件的配置信息，配置插件URL和请求方式等参数信息。配置完成后，单击“下一步”。

   风险提示：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。

   表 7-2 插件配置信息表

参数名称	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。
请求方式	插件服务的请求方式，POST或GET。

参数名称	说明
权限校验	插件服务的鉴权方式，支持以下三种： ● 无需鉴权：不使用鉴权时会存在安全风险。 ● 用户级鉴权：用户级鉴权可以使用Header鉴权或Query鉴权的方式，需要提供密钥鉴权参数名和密钥来源参数名。 ● API Key：API Key鉴权可以使用Header鉴权或Query鉴权的方式，需要提供密钥鉴权参数名和密钥值。
请求头	插件服务的请求头。添加请求的数据格式等说明，敏感信息请通过权限校验的方式实现。

6. 配置插件的参数信息，配置请求参数和响应参数信息。
   - 请求参数

     单击“添加参数”，可以添加多个请求参数。

     表 7-3 请求参数信息

参数名称	说明
参数名称	参数的名称，长度为1 ~ 50个字符，参数名称会作为大模型解析参数含义的依据。
参数描述	参数的名称，长度为1 ~ 200个字符，参数名称会作为大模型解析参数含义的依据。
参数类型	该参数值的数据类型，当前支持三种类型。 ● String：字符串类型 ● Integer：四字节整型 ● Number：八字节浮点数
请求方式	默认以Body方式请求。
是否必填	指定该参数是否为必填项。 ● 打开开关：必填 ● 关闭开关：非必填
默认值	参数的默认值，如果插件服务的入参生成缺失，默认值会在大模型解析时被使用。

- 响应参数
  单击“添加参数”，可以添加多个响应参数。

  表 7-4 响应参数信息

参数名称	说明
参数名称	参数的名称，长度为1 ~ 50个字符，参数名称会作为大模型解析参数含义的依据。
参数描述	参数的名称，长度为1 ~ 200个字符，参数名称会作为大模型解析参数含义的依据。
参数类型	该参数值的数据类型，当前支持三种类型。 ● String：字符串类型 ● Integer：四字节整型 ● Number：八字节浮点数
是否必填	指定该参数是否为必填项。 ● 打开开关：必填 ● 关闭开关：非必填

7. 配置完成后，单击“确定”，即可完成插件的创建。

7.2.4 配置知识

配置知识的步骤如下：

1. 在“高级配置 > 知识”，单击“添加”。
2. 在“添加知识”窗口，单击“点此上传”，上传知识文件。

   图 7-7 添加知识

3. 上传完成后，单击“确定”。
4. 在“高级配置”中，可查看上传成功的知识文件。

图 7-8 知识上传成功

7.2.5 配置开场白和推荐问题

配置开场白和推荐问题的步骤如下：

1. 在“高级配置 > 开场白和推荐问题”中，可输入自定义开场白，也可单击“智能添加”。
2. 在推荐问中单击“添加”，可增加推荐问数量。添加后可在右侧“预览调试”中查看相应效果。

   最多可以添加3个推荐问。

图 7-9 预览调试查看开场白与推荐问效果

7.2.6 调试 Agent 应用

平台提供对Agent执行过程的全链路信息观测与调试调优，通过对信息的分层分析和展示，为开发者提供了AI应用在不同层级的运行情况指导和操作，提升观测和调试效率。通过Insight提供了Agent的运行和观测能力。创建并运行Agent后，可通过单击Insight查看该Agent的执行信息。当前仅支持对知识性应用进行观测和调试。

前提条件

已成功创建应用。

操作步骤

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”，将跳转至Agent开发平台。

3. 单击左侧导航栏“工作台”，在“应用”页签，单击待调试的应用。单击应用右上侧的“调试详情”，进入调试详情页面。
4. 在调试详情页面，单击 ，选择需要查看的信息。
5. 单击“日志概览”页签。
   可以查看到该次执行的整体情况，包括执行状态、开始/结束时间、运行时长和输入/输出。
6. 单击“节点详情”页签。
   可以查看到该次执行的主要组件耗时时长和占比情况，以及该次执行的调用链及其是否成功的状态。
7. 单击调用链中的某个组件（例如插件天气搜索），展开调用链。
   可以查看到调用链中该组件的输入和输出。

此外，平台支持配置构建应用所需的NLP大模型参数。

单击应用右上角的 ，打开大模型参数配置页面。配置参数见表7-5，完成大模型参数配置。

表 7-5 大模型参数配置

参数	说明
模型选择	选择要使用的LLM，不同的模型效果存在差异。
模式选择	用于配置大模型的输出多样性。 包含取值： ● 精确的：模型的输出内容严格遵循指令要求，可能会反复讨论某个主题，或频繁出现相同词汇。 ● 平衡的：平衡模型输出的随机性和准确性。 ● 创意性的：模型输出内容更具多样性和创新性，某些场景下可能会偏离主旨。 ● 自定义：自定义大模型输出的温度和核采样值，生成符合预期的输出。
温度	用于控制生成结果的随机性，取值范围0-1。 ● 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0-1。 建议不要与温度同时调整。

7.3 创建与管理工作流

7.3.1 工作流简介

Agent平台工作流由多个组件构成，组件是组成工作流的基本单元。例如，大模型、插件、代码、判断等组件。

创建工作流时，工作流默认包含了开始、结束和大模型组件，每个组件需要配置不同的参数，如组件配置、输入和输出参数等。基于该工作流，开发者可通过拖、拉、拽可视化组件等方式添加更多的组件，实现复杂业务流程的编排，从而快速构建Agent。

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。

7.3.2 创建工作流

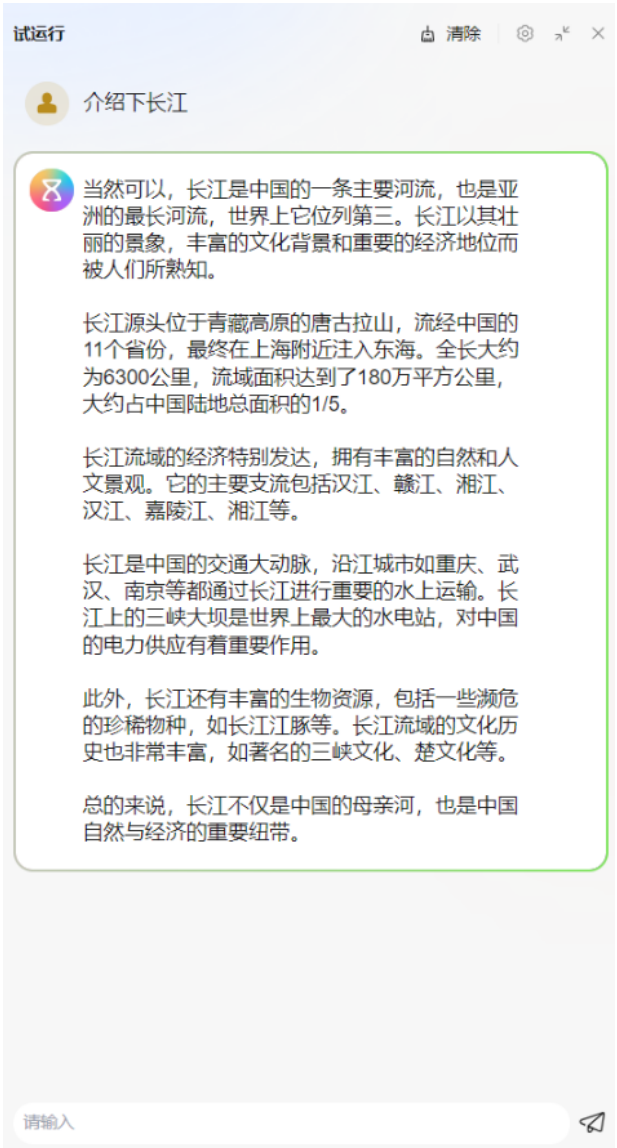
支持开发者基于Agent平台创建工作流。创建工作流时，工作流默认包含了开始、结束和大模型组件。开发者可基于该工作流，添加更多的组件，实现业务流程的编排。

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”，将跳转至Agent开发平台。
3. 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”。
4. 配置工作流的名称和描述，并单击“确认”。

图 7-10 编辑工作流基本信息

5. 创建工作流后，页面会自动跳转至工作流编辑页面。初始状态下，工作流包含开始、结束和大模型组件。
   - 开始：工作流的入口组件，该组件的配置详见**配置开始组件**。
   - 结束：输出工作流的执行结果，该组件的配置详见**配置结束组件**。
   - LLM：初始化完成的大模型节点，没有额外的Prompt配置，直接接受用户原始输入，并输出大模型执行后的原始输出，该组件的配置详见**配置大模型组件**。
6. 用户可根据需求配置所需组件，并连接其他组件。除开始、结束和大模型组件外，平台提供了意图识别、提问器、插件、判断、代码组件，配置详见**配置意图识别组件**、**配置提问器组件**、**配置插件组件**、**配置判断组件**、**配置代码组件**。
7. 当构造完成一个工作流，可以尝试运行该工作流。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。

图 7-11 试运行工作流



配置开始组件

开始组件用于触发一个工作流，用户的输入由开始组件进行承载，是每个工作流的入口组件。不支持新增或者删除开始组件。

单击画布中的开始组件，打开参数配置页面。开始组件的参数默认已配置，不支持修改开始组件的参数配置。

图 7-12 开始组件配置图

配置结束组件

结束组件是工作流给出输出的组件，其标识着工作流的结束。每个工作流执行完成后，都需要一个结束组件用于输出工作流的执行结果。结束组件后，不支持添加其他组件。不支持新增或者删除结束组件。

结束组件可能会有多个输入，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。

1. 单击画布中的“结束”组件，打开参数配置页面。

   图 7-13 结束组件配置图

2. 在“参数配置”中，配置输入参数。

   单击“添加参数”，可以添加多个输入参数。

表 7-6 参数说明表

参数名称	说明
参数名称	由开发者自定义变量名。 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：_
取值	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组件输出变量值。 ● 输入：支持用户自定义取值。

3. 支持用户将多个输入变量合并成一个字符串输出，使用 {{变量名}} 代指上述定义的输入参数。

   如上已定义输入参数：end_input，在指定回复中，{{end_input}} 即代表参数值。例如end_input值为hello，指定回复中为{{end_input}} world，则最终的输出即为helloworld。

   当然，也可以使用快捷方式选择变量。例如，当在指定回复输入框中输入 { 或者 / 的时候，会显示出上述已经定义的所有的变量的列表供选择。

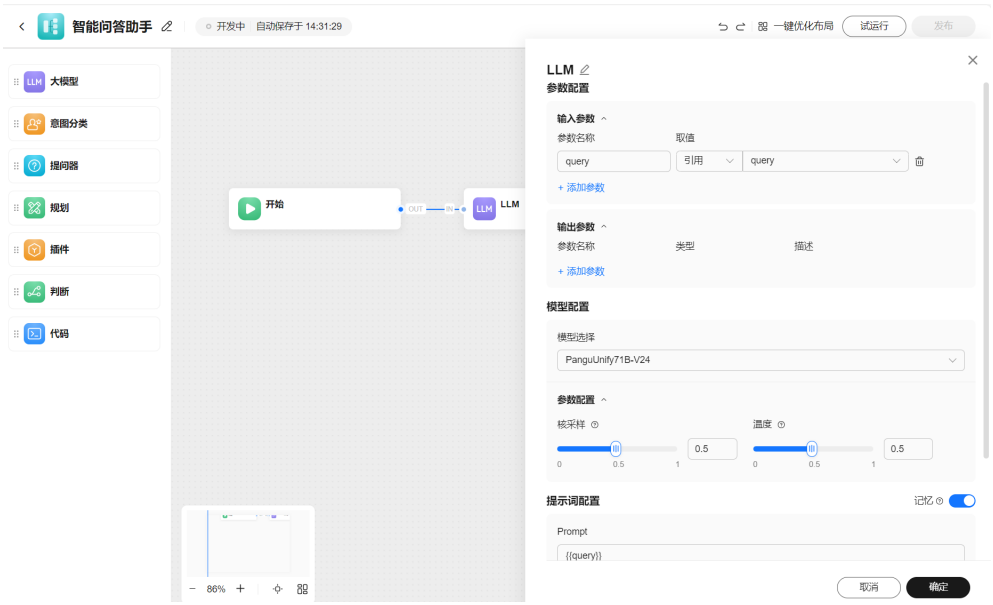
4. 单击“确定”，完成参数配置。

配置大模型组件

大模型组件提供了使用LLM的能力，用户可以通过在UI界面上编写Prompt、设置LLM的参数来让LLM完成指定的任务。

1. 单击画布中的“大模型”组件，打开参数配置页面。

图 7-14 查看大模型组件参数配置



2. 在“参数配置”中，配置输入和输出参数。
   - 输入参数

     单击“添加参数”，可以添加多个输入参数。

     将提供给“提示词配置”中Prompt使用。

     表 7-7 输入参数

参数名称	说明
参数名称	输入参数名称。 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：_
取值	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组件输出变量值。 ● 输入：支持用户自定义取值。

   - 输出参数

     单击“添加参数”，可以添加多个输出参数。

     用于解析大模型组件的输出，并提供给后序组件的输出参数引用。

表 7-8 输出参数

参数名称	说明
参数名称	输出参数名称。 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。大模型组件有一个内置输出参数rawOutput，代表该组件未经解析的原始输出，与大模型组件相连的后序组件可以直接引用该输出。
类型	输出参数的类型，当前可选类型只有String。
描述	对于该输出参数的描述。

须知

如下场景时，可以通过配置输出参数来解析大模型组件的输出：

当大模型组件的输出为json格式的数据时，可以通过配置输出参数来解析出json中对应字段的值。例如大模型组件的输出为json数据'{"result": "test"}'时，可以添加一个参数名称为“result”的输出参数，那么输出参数“result”就会从json数据中取出同名字段对应的值“test”。

3. 在“模型配置”中，选择要使用的大模型，并通过拖动滑条来设置参数“核采样”和“温度”。
   - 模型选择：选择要使用的LLM，不同的模型效果存在差异。
   - 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。建议不要与核采样同时调整。
     - 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。
4. 配置提示词信息，并选择是否打开记忆功能。

   写提示词时，支持使用{{variable}}的格式引用本组件输入参数中已定义好的参数。
   - Prompt：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。

5. 单击“确定”，完成参数配置。
6. 连接大模型组件和其他组件。

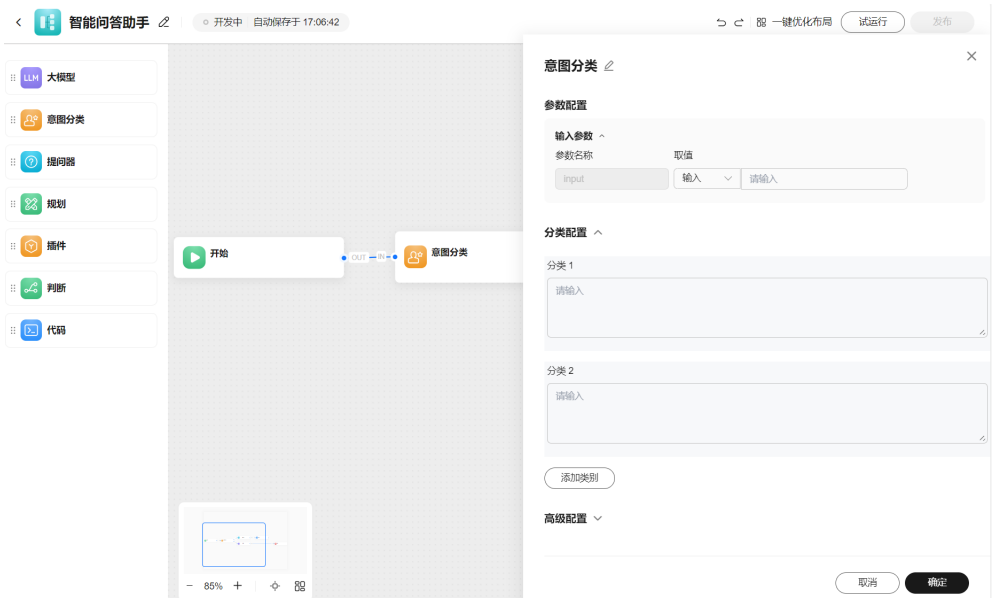
配置意图识别组件

意图识别组件用于根据用户的输入进行分类并导向后续不同的处理流程。

意图识别组件一般位于工作流前置位置。在对用户的输入进行意图识别时，意图识别组件会通过大模型推理，匹配用户输入与开发者预先定义的描述类别的关键字，并根据匹配结果流向对应处理流程。

1. 在左侧组件面板中拖拽出一个“意图识别”组件，并放置在工作流中。
2. 单击画布中的“意图识别”组件，打开参数配置页面。

图 7-15 意图分类配置图



3. 在“参数配置”中，配置输入参数。

表 7-9 输入参数

参数名称	说明
参数名称	默认名称input，为固定值，不可编辑。
取值	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组件输出变量值。 ● 输入：支持用户自定义取值。

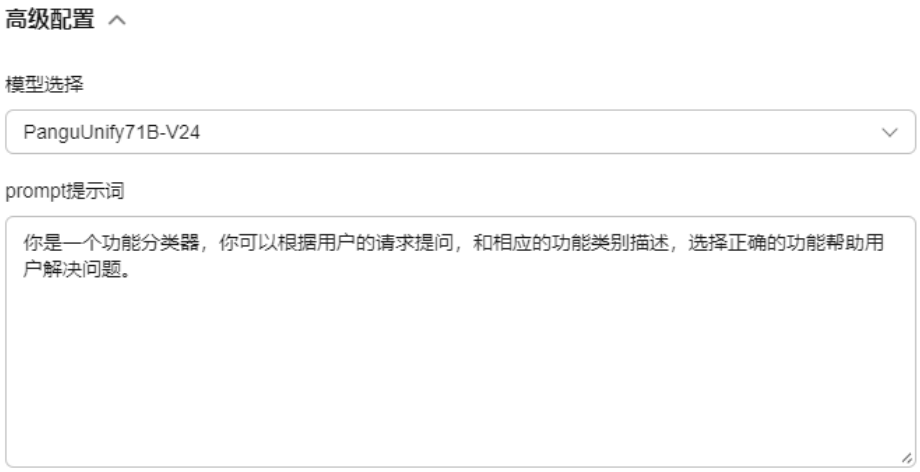
4. 配置“分类配置”中的相关信息。

   在分类输入框中输入分类描述信息，描述信息为针对该类别的描述语句或者关键词，同时也将作为LLM进行推理和分类的依据。类别数量为2 ~ 5个。

5. 配置“高级配置”中的相关信息。

   高级配置项供进阶开发者修改模型和提示词，如果不配置将会使用系统默认值。模型的选择和提示词的撰写可能影响到意图分类组件的准确性。

   - 模型选择：选择要使用的LLM，不同的模型效果存在差异。
   - Prompt提示词：用户对模型的指令，提示词可能影响模型效果。

6. 单击“确定”，完成参数配置。
7. 连接意图分类组件和其他组件。

配置提问器组件

在工作流中，提问器组件给开发者提供了收集回复用户问题所必需信息的能力。提问器组件将会循环执行，直至将所需的信息收集完整。

1. 在左侧组件面板中拖拽出一个“提问器”组件，放置在工作流合适的位置。
2. 单击画布中的“提问器”组件，打开参数配置页面。

图 7-16 提问器配置图

3. 在“参数配置”中，配置输入和输出参数。
   - 输入参数

     单击“添加参数”，可以添加多个输入参数。

     表 7-10 输入参数

参数名称	说明
参数名称	由开发者自定义，可以通过双花括号形式在后续“问题配置”中被参数“问题”引用。 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：_ 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”。
取值	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组件输出变量值。 ● 输入：支持用户自定义取值。

   - 输出参数

     单击“添加参数”，可以添加多个输出参数。

     在“提问器”组件中，输出参数为希望向用户收集的信息，可以被后续组件所引用。

     “输出参数”与高级配置中的“问题额外配置”为一一对应关系，问题额外配置中的参数将随着“输出参数”的增减而自动跟随变化。其中，“输出参数”是必填项，“问题额外配置”是选填项。

表 7-11 输出参数

参数名称	说明
参数名称	输出参数名称。 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：_
类型	输出参数的类型，当前可选类型只有String。
描述	对于该输出参数的描述。

4. （可选）配置“问题配置”中的相关信息。

   参数“问题”将在对话框中原样呈现给用户。如未配置此处，将由LLM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。

5. （可选）配置“高级配置”中的相关信息。

   高级配置项供进阶开发者修改选择模型和提示词，如果不配置将会使用系统默认值。

   模型的选择和提示词的撰写可能影响到提问器组件的准确性。

   - 模型选择：选择要使用的LLM，不同模型的效果存在差异。
   - Prompt提示词：用户对模型的指令，提示词可能影响模型效果。
   - 问题额外配置：
     - 参数名称：与输出参数的参数名称一一对应，用户不可修改，自动跟随输出参数的变化而改动。
     - 问题关键词：问题关键词是对输出参数描述信息的提炼，帮助大模型更好地理解问题关键词。此处为选填，不填写可能影响模型提取效果。当“问题配置”的“问题”信息与“高级配置”中“问题额外配置”的“问题关键词”都填写时，提问器组件会校验问题中是否已经包含所有的问题关键词。
     - 内容示例：内容示例可以举例说明所需信息的格式，帮助大模型更好地从用户的回答中提取所需信息。例如，参数名称“手机号码”，可以在内容示例中填写“12345678910”。

6. 单击“确定”，完成参数配置。
7. 连接提问器组件和其他组件。

配置插件组件

插件组件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。开发者可以通过托拉拽方式将插件库中插件构建一个插件组件。

- 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。

1. 在左侧组件面板中拖拽出一个“插件”组件。
2. 选择“个人插件”页签。
   在“个人插件”中选择自己创建的插件。
3. 选择需要添加的插件，单击，即可完成添加插件。
   支持选择多个插件。
4. 单击画布中的已添加的“插件”组件，打开参数配置页面。
5. 在“参数配置”中，配置输入参数和输出参数。
   - 输入参数

     表 7-12 输入参数

参数名称	说明
参数名称	输入参数名称从插件元信息中导入，用户无需手动添加。
取值	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组件输出变量值。 ● 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。
6. 单击“确定”，完成参数配置。
7. 连接插件组件和其他组件。

配置判断组件

判断组件是一个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，可通过拖拽分支条件配置面板来设定分支条件的优先级。

1. 在左侧组件面板中拖拽出一个“判断”组件，放置在工作流合适的位置。
2. 单击画布中的“判断”组件，打开参数配置页面。

3. 在“参数配置”中配置“IF”相关参数。
   IF分支由**[变量 比较条件 比较对象]**组成一条件表达式。

表 7-13 IF 分支参数

参数名称	说明
变量	条件表达式左边部分，需要选择来自前序组件的输出参数。
比较条件	条件表达式中间部分，当前支持的比较条件有： ● equal：等于 ● not equal：不等于 ● contain：包含 ● not contain：不包含
比较对象	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组件输出变量值。 ● 输入：支持用户自定义取值。

IF分支其他操作如下：

- 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“and”或“or”来连接。
- 单击“and”或者“or”，可以切换该分支表达式的运算逻辑。

图 7-17 IF 分支配置图

- “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 7-18 添加 ELSE IF 图

4. 单击“确定”，完成参数配置。
5. 连接判断组件和其他组件。

配置代码组件

代码组件支持编写Python代码来处理文本、复杂逻辑判断等。

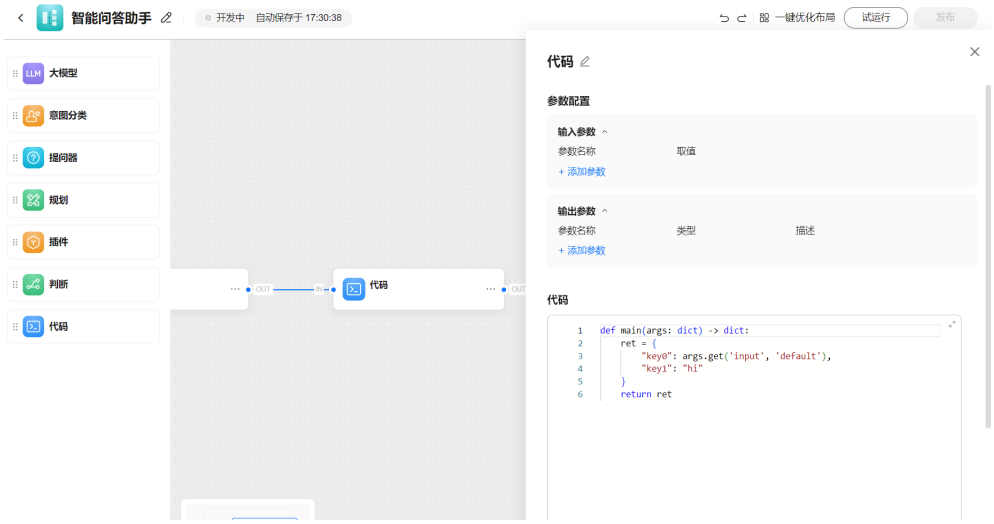
代码组件可以增强开发人员的灵活性，使他们能够在工作流程中嵌入自定义Python脚的方式操作变量。通过配置选项，开发者可以指定所需的输入和输出变量，并编写相应的执行代码。

须知

编写代码时不要更改第一行函数定义以及输入输出定义。

1. 在左侧组件面板中拖拽出一个“代码”组件，放置在工作流合适的位置。
2. 单击画布中的“代码”组件，打开参数配置页面。

图 7-19 代码配置图



3. 在“参数配置”中，配置输入和输出参数。
   - 输入参数
     单击“添加参数”，可以添加多个输入参数。

     表 7-14 输入参数

参数名称	说明
参数名称	由开发者自定义变量名。 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：_
取值	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组件输出变量值。 ● 输入：支持用户自定义取值。

   - 输出参数
     单击“添加参数”，可以添加多个输出参数。

表 7-15 输出参数

参数名称	说明
参数名称	输出参数名称。 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：_
类型	输出参数的类型，当前可选类型只有String。
描述	对于该输出参数的描述。

4. 编写Python代码。

   代码配置示例如下：

   - 文本拼接示例代码。

     ```
     def main(args: dict) -> dict:
     # 注意在输入参数中定义名为input1的变量
     input1 = args.get('input1')
     # 注意在输入参数中定义名为input2的变量
     input2 = args.get('input2')
     res = {
     # 注意在输出参数中定义名为res的变量
     "res": input1 + input2,
     }
     return res
     ```

   - 复杂逻辑判断示例代码。

     ```
     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"
     }
     ```

   - 数学计算示例代码。

     ```
     def main(args: dict) -> dict:
     # 注意在输入参数中定义名为input1的变量
     input1 = args.get('input1')
     try:
     input1 = int(input1)
     return {
     # 注意输出参数中定义res变量
     'res': input1 * input1
     }
     except Exception as e:
     return {
     # 注意输出参数中定义res变量
     ```

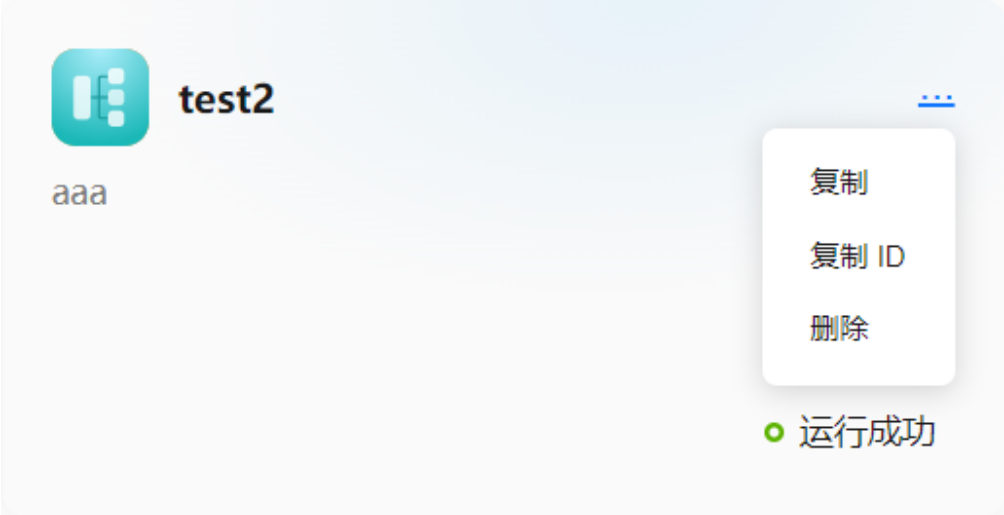


```
'res': "输入类型错误或者数字大小超出限制"
}
```

5. 单击“确定”，完成参数配置。
6. 连接代码组件和其他组件。

7.3.3 管理工作流

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”，将跳转至Agent开发平台。
3. 单击左侧导航栏“工作台”，在“工作流”页签中，鼠标右键单击工作流，进行工作流的复制、复制ID、删除。



8 管理盘古大模型空间资产

8.1 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。

8.2 管理盘古数据资产

数据资产介绍

用户发布的数据集会被纳入数据资产，集中存储在空间资产中。平台为数据资产提供了一系列管理功能，包括查看数据集的详细信息、追踪操作记录、以及数据集的删除管理等。这不仅便于用户对已发布数据集的集中管理，还可帮助用户了解每个数据集的使用情况，从而简化数据资产的维护更新流程。通过这样的统一管理，用户能够更高效地组织和利用数据资源，确保数据资产的安全性和一致性。

管理数据资产

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“空间资产 > 数据”中可以查看当前空间内的数据资产，如果有多个空间的访问权限，可切换空间查看其他空间内的资产。
3. 在“数据发布”页签可查看数据资产，并可对数据集进行删除操作。单击数据集名称可进入详情页面查看数据集的基础信息和操作概览。

图 8-1 查看数据资产

8.3 管理盘古模型资产

模型资产介绍

用户在平台中可试用、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。

管理模型资产

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“空间资产 > 模型”中可以查看当前空间和预置的模型资产，如果有多个空间的访问权限，可切换空间查看其他空间内的资产。
3. 在“本空间”页签可查看模型资产，并可对模型进行删除操作。单击模型名称可进入详情页面查看模型的基础信息。
4. 在“预置”页签可查看用户可使用的各类模型的预置资产。

图 8-2 查看预置模型预置模型

5. 单击模型，可在“版本列表”页签查看当前模型的历史版本，并执行模型的基本操作如训练、部署等。在“操作记录”页面可查看各版本的历史操作记录。

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**导入盘古大模型至其他局点**，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 8-3 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，单击操作列“更多 > 导出”。若无导出选项，请确认该空间是否为当前用户创建的空间。
4. 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 8-4 导出模型

导入盘古大模型至其他局点

导入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。

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，也可以导入第三方大模型至ModelArts Studio大模型开发平台。

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地址和模型名称、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 8-5 导入模型