

盘古大模型

API 参考

文档版本

01

发布日期

2025-12-01



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

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

商标声明



HUAWEI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 使用前必读.....	1
1.1 概述.....	1
1.2 调用说明.....	4
1.3 请求 URI.....	4
1.4 基本概念.....	6
2 如何调用 REST API.....	8
2.1 构造请求.....	8
2.2 认证鉴权.....	10
2.3 返回结果.....	15
3 API.....	17
3.1 模型推理接口.....	17
3.1.1 NLP 大模型.....	17
3.1.1.1 文本对话.....	17
3.1.1.2 CV 大模型.....	51
3.1.2.1 Pangu-CV-图像分类-2.1.0.....	51
3.1.2.2 Pangu-CV-物体检测-S-2.1.0.....	58
3.1.2.3 Pangu-CV-物体检测-N-2.1.0.....	65
3.1.2.4 Pangu-CV-物体检测-S-3.1.0.....	71
3.1.2.5 Pangu-CV-语义分割-2.1.0.....	78
3.1.2.6 Pangu-CV-实例分割-1.1.0.....	83
3.1.2.7 Pangu-CV-万物检测-V2.....	88
3.1.2.8 Pangu-CV-视觉交互检测-V3.....	93
3.1.3 预测大模型.....	103
3.1.3.1 盘古统一编码时序预测分类大模型.....	103
3.1.3.2 盘古统一编码时序预测回归大模型.....	108
3.1.3.3 盘古统一编码表格预测分类大模型.....	114
3.1.3.4 盘古统一编码表格预测回归大模型.....	118
3.1.3.5 盘古时序异常检测大模型.....	122
3.1.3.5.1 时序异常检测预训练工作流.....	122
3.1.3.5.2 时序异常检测微调工作流.....	128
3.1.3.6 盘古融合推荐异常检测大模型.....	133
3.1.3.7 盘古融合推荐分类大模型.....	138

3.1.3.7.1 数据分类预测.....	138
3.1.3.7.2 排列特征重要性.....	142
3.1.3.8 盘古融合推荐回归大模型.....	147
3.1.3.8.1 数据回归预测.....	147
3.1.3.8.2 排列特征重要性.....	151
3.1.3.9 盘古表格直推预测大模型.....	156
3.1.4 科学计算大模型.....	161
3.1.4.1 气象类模型.....	161
3.1.4.1.1 创建推理作业.....	161
3.1.4.1.2 查询推理作业详情.....	173
3.1.5 行业大模型.....	177
3.1.5.1 行业 NLP 大模型.....	177
3.1.6 三方大模型.....	183
3.1.6.1 三方 NLP 大模型.....	183
3.1.6.2 Qwen 三方 VL 大模型.....	206
3.1.7 多模态模型.....	227
3.1.7.1 图像问答.....	227
3.1.7.2 创建文生视频任务.....	253
3.1.7.3 查询文生视频任务详情.....	257
3.1.7.4 创建图生视频任务.....	260
3.1.7.5 查询图生视频任务详情.....	265
3.1.7.6 创建视频续写任务.....	267
3.1.7.7 查询视频续写任务详情.....	271
3.1.7.8 上传图片或视频.....	274
3.1.8 图像搜索模型.....	276
3.1.8.1 服装图像标签搜索模型.....	276
3.1.9 专业大模型.....	285
3.1.9.1 Embedding 模型.....	286
3.1.9.2 Rerank 模型.....	291
3.1.9.3 搜索规划模型.....	298
3.2 数据工程接口.....	306
3.2.1 查询数据血缘.....	306
3.2.2 数据集彻底删除.....	309
3.3 Agent 应用接口.....	313
3.3.1 调用应用.....	313
3.3.2 调用工作流.....	320
3.4 Token 计算器.....	335
4 附录.....	339
4.1 状态码.....	339
4.2 错误码.....	342
4.3 获取项目 ID.....	345
4.4 获取空间 ID.....	347

4.5 获取模型部署 ID.....	348
4.6 获取用户 ID.....	349
4.7 获取负载均衡 ID.....	350

1 使用前必读

[概述](#)
[调用说明](#)
[请求URI](#)
[基本概念](#)

1.1 概述

ModelArts Studio 大模型开发平台支持纳管盘古大模型及三方大模型，模型在平台部署后，可以通过 API 调用推理接口。

表 1-1 API 清单

类别	模型	API	功能
模型推理接口	NLP大模型	文本对话	基于对话问答功能，用户可以与模型进行自然而流畅的对话和交流。
	CV大模型	Pangu-CV-图像分类-2.1.0	根据在图像信息中所反映的不同特征，对图像进行定量分析，把图像划归为若干个类别中的某一种。适用于动植物分类、车辆类型分类、车牌分类、废钢定级、零部件分类等任务。
		Pangu-CV-物体检测-S-2.1.0	找出图像中所有感兴趣的目标，确定它们的位置和类别。物体检测-S模型特点是小参数量，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。
		Pangu-CV-物体检测-N-2.1.0	找出图像中所有感兴趣的目标，确定它们的位置和类别。物体检测-N模型特点是参数量适中，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。

类别	模型	API	功能
预测大模型	Pangu-CV-物体检测-S-3.1.0	Pangu-CV-物体检测-S-3.1.0	盘古计算机视觉物体检测大模型，任务是找出图像中所有感兴趣的目标，确定它们的位置和类别。
		Pangu-CV-语义分割-2.1.0	将数字图像细分为多个图像子区域，适用于车道分割、建筑分割、选煤厂筛面状态识别等任务。
		Pangu-CV-实例分割-1.1.0	能够对输入图片中的不同类别物体及物体个体进行分割识别，输出每个实例的类别标签、置信度及坐标信息。
	盘古统一编码时序预测分类大模型	盘古统一编码时序预测分类大模型	基于时序预测基模型实现分类预测能力。时序分类预测有很多应用场景，例如：基于工业设备传感器一段时间采集的连续数据，实现设备正常或异常状态的预测。
	盘古统一编码时序预测回归大模型	盘古统一编码时序预测回归大模型	2025年3月首次发布的模型，基于时序预测基模型实现对未来一组观测值的预测，最多支持128个未来时间点的预测。
	盘古统一编码表格预测分类大模型	盘古统一编码表格预测分类大模型	基于统一编码大模型实现表格分类预测能力，针对特定场景的分类任务，用户传入分类数据，使用模型对指定的预测目标进行分类预测。
	盘古统一编码表格预测回归大模型	盘古统一编码表格预测回归大模型	基于统一编码大模型实现表格回归预测能力，针对特定场景的回归任务，用户传入回归数据，使用模型对指定的预测目标进行回归预测。
	时序异常检测预训练工作流	时序异常检测预训练工作流	可实现时序数据的异常检测任务。支持仅使用设备历史数据对时序异常检测大模型进行全量训练，使其更好地学习并重构正常数据的潜在特征分布。
	时序异常检测微调工作流	时序异常检测微调工作流	可实现时序数据的异常检测任务。支持使用历史正异常数据对模型进行微调，实现对异常状态和正常状态的识别。
	盘古融合推荐异常检测大模型	盘古融合推荐异常检测大模型	针对特定场景的异常检测任务，用户传入异常检测数据，使用模型对指定的预测目标进行异常检测预测。
	数据分类预测-盘古融合推荐分类大模型	数据分类预测-盘古融合推荐分类大模型	针对特定场景的分类任务，用户传入分类数据，使用模型对指定的预测目标进行分类预测。
	排列特征重要性-盘古融合推荐分类大模型	排列特征重要性-盘古融合推荐分类大模型	支持用户传入一批数据样本，并据此计算各个特征的排列重要性（permutation importance）。

类别	模型	API	功能
盘古大模型	盘古融合推荐回归大模型	数据回归预测 -盘古融合推荐回归大模型	针对特定场景的回归任务，用户传入回归数据，使用模型对指定的预测目标进行回归预测。
		排列特征重要性 -盘古融合推荐回归大模型	支持用户传入一批数据样本，并据此计算各个特征的排列重要性（permutation importance）。
	科学计算大模型	创建推理作业	支持调用科学计算大模型创建气象/降水模型的推理作业。
		查询推理作业详情	根据 创建推理作业 的作业ID获取科学计算大模型的结果数据。
	行业大模型	行业NLP大模型	基于文本对话功能，用户可以与医学行业大模型进行自然而流畅的对话和交流，实现医学领域的文本生成、行业推理等场景
	三方大模型	三方NLP大模型	DeepSeek API是基于DeepSeek大模型推出的接口服务，它支持多场景文本交互，能够快速生成高质量对话、文案、故事等内容，可用于文本摘要、智能问答、内容创作等场景。
	多模态模型	图像问答	多模态图像问答接口提供图片描述、视觉问答、OCR等能力，常用于图像理解、视觉问答等理解和生成任务。
	图像搜索模型	服装图像标签搜索模型	基于服装图片返回图像中的服装类型、款式信息、工艺类别及置信度信息。
	专业大模型	Embedding模型	Embedding模型服务在创建知识库中，文本处理阶段，用于对文本文档进行前片，转换成向量化表示。在知识检索阶段，根据用户输入的query对切片进行召回。
		Rerank模型	Rerank模型服务用于对召回的切片，按照query与切片的相关度进行精细化排序，以确保召回相关度top切片。
		搜索规划模型	Pangu-SearchPlan模型，用于RAG场景，提供通用意图分类/多轮查询改写/复杂查询分解/时间抽取等功能，在RAG任务中生产用于检索的query，以及根据query分类路由到后续不同的流程。
数据工程接口	-	查询数据血缘	客户通过obs导入原始数据集，可基于该obs路径查询所有基于该路径创建的原始数据集及后续的血缘信息。

类别	模型	API	功能
	-	数据集彻底删除	只针对从obs上传的数据，在删除数据集的时候要关联删除OBS下对应的原始数据，客户认为原始数据应该在客户侧大数据中心长期归档，不应该在OBS长期保留。
Agent应用接口	-	调用应用	通过调用创建好的应用API，输入问题，将得到应用执行的结果。
	-	调用工作流	通过调用创建好的工作流API，输入问题，将得到工作流执行的结果。
Token计算器	-	Token计算器	为了帮助用户更好地管理和优化Token消耗，平台提供了Token计算器工具。Token计算器可以帮助用户在模型推理前评估文本的Token数量，提供费用预估，并优化数据预处理策略。

说明

用户在部署服务的过程中，建议开启“安全护栏”功能，以保证内容的安全性。

1.2 调用说明

盘古大模型提供了REST (Representational State Transfer) 风格的API，支持您通过HTTPS请求调用，调用方法请参见[如何调用REST API](#)。

调用API时，需要用户网络可以访问公网。

1.3 请求 URI

服务的请求URI即API服务的终端地址，通过该地址与API进行通信和交互。

URI获取步骤如下：

步骤1 登录ModelArts Studio大模型开发平台。

步骤2 进入所需工作空间。

步骤3 获取请求URI。

- 获取模型请求URI。
 - 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的请求URI。

图 1-1 部署后的模型调用路径



若调用的是用户自己部署的NLP推理服务，可在“API调用”页签中，可获取V1接口URL或者V2接口URI。

图 1-2 NLP 服务的调用路径



- 若调用预置模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“预置服务”页签，模型列表单击“调用路径”，获取该模型的请求URI。

图 1-3 预置模型的调用路径



- 获取Agent应用请求URI。
 - 单击左侧导航栏“Agent开发”，进入“工作台 > 应用”页面，选择需要部署的应用，单击“... > 调用路径”。
 - 在“调用路径”页面可获取Agent应用请求URI。

图 1-4 调用路径



----结束

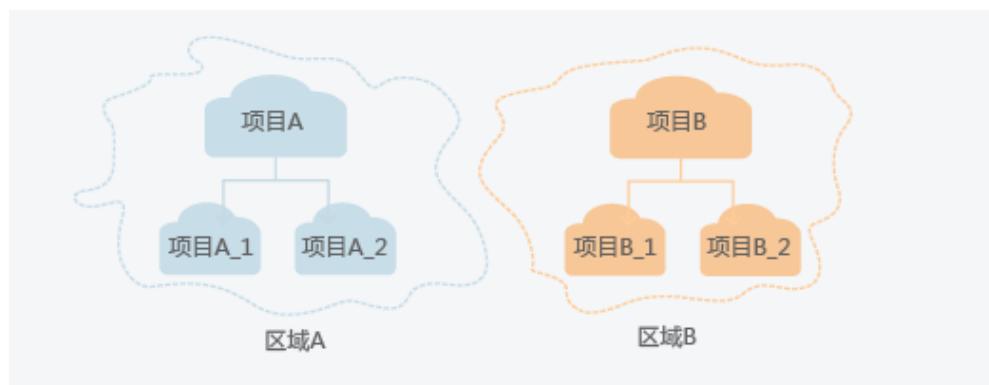
1.4 基本概念

- 账号

用户注册华为云时的账号，账号对其所拥有的资源及云服务具有完全的访问权限，可以重置用户密码、分配用户权限等。由于账号是付费主体，为了确保账号安全，建议您不要直接使用账号进行日常管理工作，而是创建用户并使用他们进行日常管理工作。

- **用户**
由账号在IAM中创建的用户，是云服务的使用人员，具有身份凭证（密码和访问密钥）。
在我的凭证下，您可以查看账号ID和用户ID。通常在调用API的鉴权过程中，您需要用到账号、用户和密码等信息。
- **区域 (Region)**
从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
详情请参见[区域和可用区](#)。
- **可用区 (AZ, Availability Zone)**
一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。
- **项目**
华为云的区域默认对应一个项目，这个项目由系统预置，用来隔离物理区域间的资源（计算资源、存储资源和网络资源），以默认项目为单位进行授权，用户可以访问您账号中该区域的所有资源。如果您希望进行更加精细的权限控制，可以在区域默认的项目中创建子项目，并在子项目中购买资源，然后以子项目为单位进行授权，使得用户仅能访问特定子项目中资源，使得资源的权限控制更加精确。

图 1-5 项目隔离模型



2 如何调用 REST API

构造请求
认证鉴权
返回结果

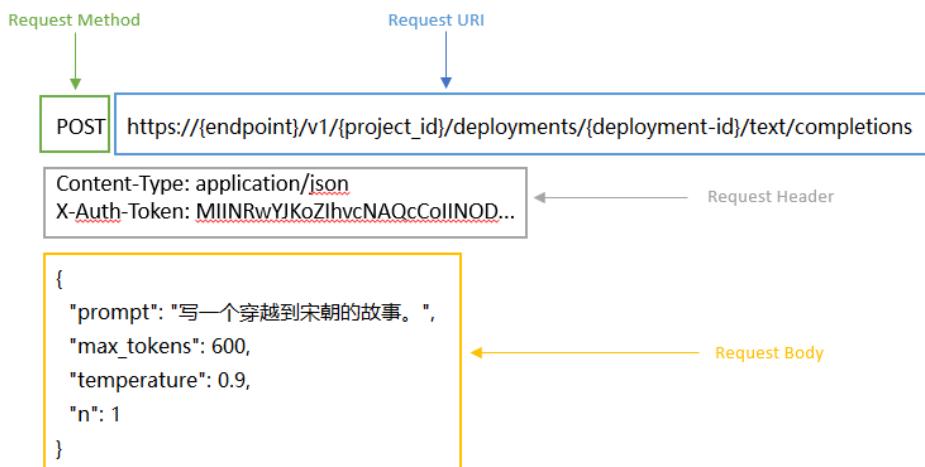
2.1 构造请求

本节介绍REST API请求的组成，并以调用服务的[获取用户Token](#)接口说明如何调用API。

您还可以通过这个视频教程了解如何构造请求调用API：<https://bbs.huaweicloud.com/videos/102987>。

请求示例如图2-1，一个请求主要由[请求URI](#)、[请求方法](#)、[请求消息头](#)和[请求消息体](#)组成。

图 2-1 请求示例图



请求 URI

请求URI由如下部分组成：

{URI-scheme}://{endpoint}/{resource-path}?{query-string}

表 2-1 请求 URI

参数	说明
URI-scheme	传输请求的协议，当前所有API均采用HTTPS协议。
endpoint	承载REST服务端点的服务器域名或IP。
resource-path	资源路径，即API访问路径。从具体API的URI模块获取。
query-string	查询参数，可选，查询参数前面需要带一个“？” ，形式为“参数名=参数取值”。

获取请求URI的步骤详见[请求URI](#)，示例如下：

`https://{{endpoint}}/v1/{{project_id}}/deployments/{{deployment_id}}/chat/completions`

请求方法

HTTP请求方法，表示服务正在请求操作类型，包括：

- **GET**: 请求服务器返回指定资源。
- **PUT**: 请求服务器更新指定资源。
- **POST**: 请求服务器新增资源或执行特殊操作。
- **DELETE**: 请求服务器删除指定资源，如删除对象等。
- **HEAD**: 请求服务器资源头部。
- **PATCH**: 请求服务器更新资源的部分内容。当资源不存在的时候，PATCH可能会去创建一个新的资源。

在接口的URI部分，请求方法为“POST”，例如：

`POST https://{{endpoint}}/v1/{{project_id}}/deployments/{{deployment_id}}/chat/completions`

请求消息头

附加请求头字段，如指定的URI和HTTP方法所要求的字段。例如，定义消息体类型的请求头“Content-Type”，请求鉴权信息等。

以下公共消息头需要添加到请求中。

- **Content-Type**: 消息体的类型（格式），必选，默认取值为“application/json”。
- **X-Auth-Token**: 用户Token，可选，当使用Token方式认证时，必须填充该字段。用户Token请参考[认证鉴权](#)中的“Token认证”。

□ 说明

公有云API同时支持使用AK/SK认证，AK/SK认证是使用SDK对请求进行签名，签名过程会自动往请求中添加Authorization（签名认证信息）和X-Sdk-Date（请求发送的时间）请求头。AK/SK认证的详细说明请参见：[AK/SK](#)。

添加消息头后的请求如下所示：

```
POST https://{{endpoint}}/v1/{{project_id}}/deployments/{{deployment_id}}/chat/completions
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZlhcNAQcCollNOD...
```

请求消息体

请求消息体通常以结构化格式发出，与请求消息头中Content-Type对应，传递除请求消息头之外的内容。若请求消息体中参数支持中文，则中文字符必须为UTF-8编码。

每个接口的请求消息体内容不同，也并不是每个接口都需要有请求消息体（或者说消息体为空），GET、DELETE操作类型的接口就不需要消息体，消息体具体内容需要根据具体接口而定。

将消息体加入后的请求如下所示，详细参数解释可参考文档API章节。

```
POST https://{{endpoint}}/v1/{{project_id}}/deployments/{{deployment_id}}/chat/completions
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZlhcNAQcCollNOD...
{
  "messages": [
    {
      "content": "介绍下长江，以及长江中典型的鱼类"
    }
  ],
  "temperature": 0.9,
  "max_tokens": 600
}
```

综上，您可以使用[curl](#)、[Postman](#)或直接编写代码等方式发送请求调用API。对于接口，您可以从响应消息部分看到返回参数及参数说明。

2.2 认证鉴权

调用接口有如下认证方式，您可以选择其中一种进行认证鉴权。

- Token认证：通过Token认证调用请求。
- API Key认证：当用户部署的模型服务期望开放给其他用户调用时，使用原有的Token认证需要进行动态认证鉴权和凭证管理，操作繁杂。此时可使用API Key认证。该方式不仅相比Token认证更简便，还与业界主流模型调用规范保持一致。

Token 认证

Token在计算机系统中代表令牌（临时）的意思，拥有Token就代表拥有某种权限。Token认证就是在调用API的时候将Token加到请求消息头，从而通过身份认证，获得操作API的权限。

□ 说明

- Token的有效期为24小时，需要使用一个Token鉴权时，可以先缓存，避免频繁调用。
- 如果您的华为云账号已升级为华为账号，将不支持获取账号Token。建议为您自己创建一个IAM用户，获取IAM用户的Token。

获取Token方法：

Token可通过调用“获取Token”接口获取，接口调用示例如下：

- 伪码

```
POST https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens
Content-Type: application/json
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username", //IAM用户名
                    "password": "*****", //IAM用户密码
                    "domain": {
                        "name": "domainname" //账号名
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "cn-southwest-2" //盘古大模型当前部署在“西南-贵阳一”区域，取值为cn-southwest-2
            }
        }
    }
}
```

- Python

```
import requests
import json

url = "https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens"
payload = json.dumps({
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "*****",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "cn-north-4"
            }
        }
    }
})
headers = {
    'Content-Type': 'application/json'
}

response = requests.request("POST", url, headers=headers, data=payload)

print(response.headers["X-Subject-Token"])
```

获取Token步骤：

本示例中，通过使用Postman软件获取Token。

1. 登录“[我的凭证 > API凭证](#)”页面，获取user name、domain name、project id。

由于盘古大模型当前部署在“西南-贵阳一”区域，需要获取与“西南-贵阳一”区域对应的project id。

图 2-2 获取 user name、domain name、project id

The screenshot shows the 'API凭证' (API Credentials) page in the IAM console. At the top, it displays user information: IAM用户名 (IAM Username) pengi..., user name (User Name) HS..., domain name (Domain Name) HS...; IAM用户ID (IAM User ID) 1644..., user id (User ID); and 帐号名 (Account Name) HS..., 帐号ID (Account ID) 60289..., domain id (Domain ID). Below this is a section titled '项目列表' (Project List) showing a grid of projects. A red arrow points to the project '西南-贵阳一' (Southwest-Guiyang One) in the list.

2. 打开Postman，新建一个POST请求，并输入“西南-贵阳一”区域的“获取Token”接口。并填写请求Header参数。
 - 接口地址为：<https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens>
 - 请求Header参数名为Content-Type，参数值为application/json

图 2-3 填写获取 Token 接口

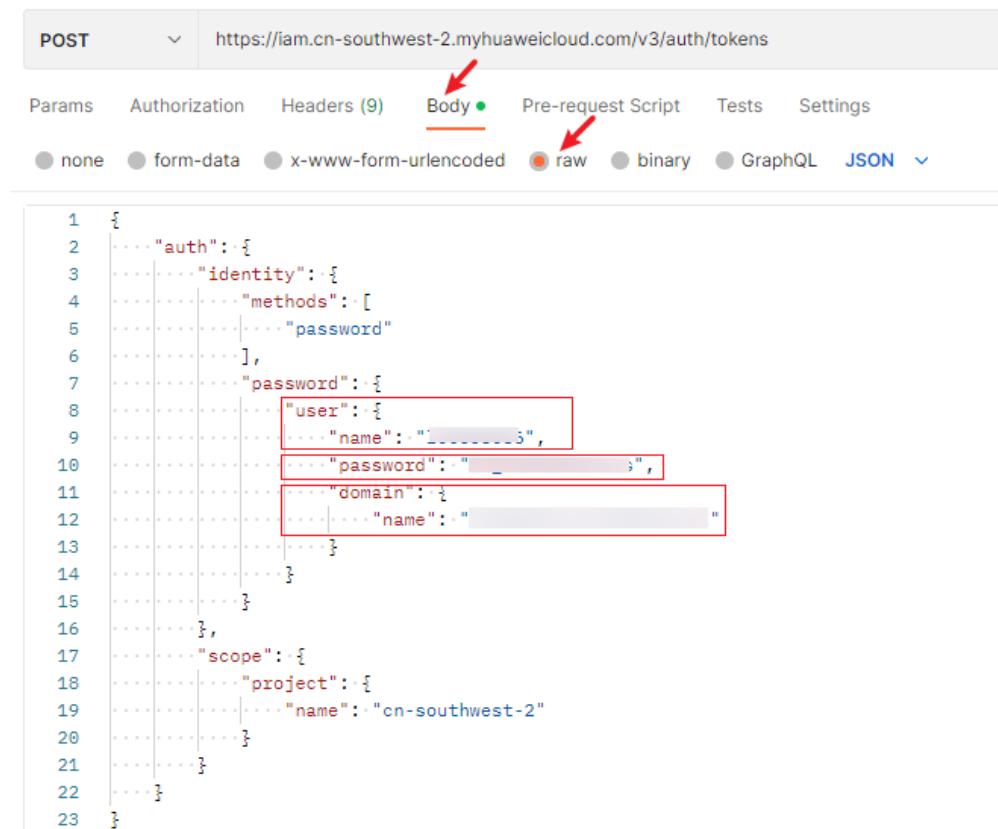
The screenshot shows the Postman interface with a POST request to the URL <https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens>. The 'Headers' tab is selected, showing a single header entry: Content-Type: application/json. A red arrow points to the URL field.

3. 填写“获取token”接口的请求体。在Postman中选择“Body > raw”选项，参考[图2-4](#)复制并填入以下代码，并填写user name、domain name、password。

```
{  
    "auth": {  
        "identity": {  
            "methods": [  
                "password"  
            ],  
            "password": {  
                "user": {  
                    "name": "username", //IAM用户名  
                    "password": "*****", //华为云账号密码  
                    "domain": {  
                        "name": "domainname" //账号名  
                    }  
                }  
            }  
        }  
    }  
}
```

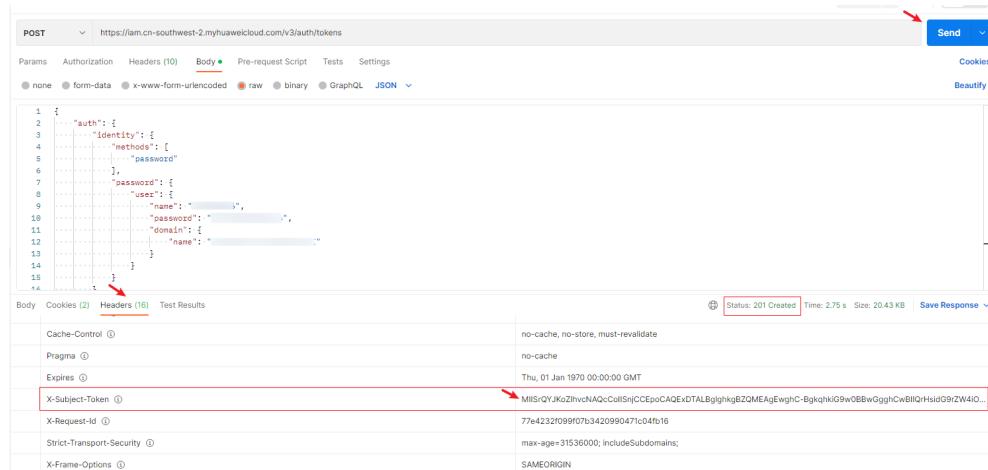
```
        }
    }
},
"scope": {
    "project": {
        "name": "cn-southwest-2" // 盘古大模型当前部署在“西南-贵阳一”区域，取值为cn-southwest-2
    }
}
}
```

图 2-4 填写请求 Body



4. 单击Postman界面“Send”按钮，发送请求。当接口返回状态为201时，表示Token接口调用成功，此时单击“Headers”选项，找到并复制“X-Subject-Token”参数对应的值，该值即为需要获取的Token。

图 2-5 获取 Token



您还可以通过这个视频教程了解如何使用Token认证：<https://bbs.huaweicloud.com/videos/101333>。

API Key 认证

当用户部署的API服务期望开放给其他用户调用时，原有[Token认证](#)无法支持，可使用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为“API Key”值），而不需要对请求内容签名。

使用该鉴权方式前，请确保有已部署的大模型。

获取API Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“应用接入”，单击界面右上角“创建应用接入”。
3. 在“关联服务”中，选择“全部服务”，或者在“指定服务”中选择已部署好的大模型服务，单击“确定”。
4. 在弹窗中复制API Key的值。

说明

注意：API Key仅可在此弹窗中复制一次，请妥善保存。关闭弹窗后，如果忘记API Key值，请重新创建。

图 2-6 复制 API Key



2.3 返回结果

状态码

请求发送以后，您会收到响应，包含状态码、响应消息头和消息体。

状态码是一组从1xx到5xx的数字代码，状态码表示了请求响应的状态，完整的状态码列表请参见[状态码](#)。

对于Pangu服务接口，如果调用后返回状态码为“200”，则表示请求成功。

响应消息头

对应请求消息头，响应同样也有消息头，如“Content-Type”。

响应消息体

响应消息体通常以结构化格式返回，与响应消息头中Content-Type对应，传递除响应消息头之外的内容。

接口调用成功后将返回如下响应体。

```
{  
    "id": "180f5745-4ee4-42a9-9869-23f829654bb7",  
    "created": 1724915285,  
    "choices": [  
        {  
            "index": 0,  
            "text": "故事标题：《穿越宋朝的奇妙之旅》在一个阴雨绵绵的夜晚，一个名叫李晓的年轻人正在阅读一本关于宋朝的历史书籍。突然，他感到一阵眩晕，当他再次睁开眼睛时，他发现自己身处一个完全陌生的地方。李晓发现自己穿越到了宋朝。他身处一座繁华的城市，人们穿着古代的服饰，用着他听不懂的语言交谈。他意识到自己真的穿越了。李晓在宋朝的生活充满了挑战。他必须学习如何使用新的语言，适应新的生活方式。他开始学习宋朝的礼仪，尝试理解这个时代的文化。在宋朝，李晓遇到了许多有趣的人。他遇到了一位名叫赵敏拿来的女孩，她聪明伶俐，让李晓对她产生了深深的喜爱。他还遇到了一位名叫王安石的大儒，他的智慧和博学让李晓深感敬佩。在宋朝的生活中，李晓也遇到了许多困难。他必须适应新的食物，新的气候，甚至新的疾病。但是，他从未放弃，他始终坚信，只要他坚持下去，他就能适应这个世界。在宋朝的生活中，李晓也找到了新的目标。他开始学习宋朝的书法，尝试理解这个时代的艺术。他还开始学习宋朝的医学，尝试理解这个时代的科学。在宋朝的生活中，李晓也找到了新的爱情。他与赵敏拿来建立了深厚的感情，他们的conseillé情深深地打动了李晓。最后，李晓决定留在宋朝。他发现，他已经深深地爱上了这个时代，爱上了这里的人，爱上了这里的食物，爱上了这里的一切。这就是李晓在宋朝的hesion之旅，一个充满挑战和机遇的旅程。虽然他来自未来，但他已经成为了一部分，成为了这个时代的一部分。",  
            "ppl": 1.77809815678146e-36  
        }  
    ],  
    "usage": {  
        "completion_tokens": 365,  
        "prompt_tokens": 9,  
        "total_tokens": 374  
    }  
}
```

当接口调用出错时，会返回错误码及错误信息说明。

token有效期为24小时，下面的报错表示token过期。

```
{  
    "error_msg": "Incorrect IAM authentication information: token expires,  
    expires_at:2023-06-29T02:16:41.581000Z",  
    "error_code": "APIG.0301",  
    "request_id": "469967f55e6b225xxx"  
}
```

其中，error_code表示错误码，error_msg表示错误描述信息。

3 API

模型推理接口
数据工程接口
Agent应用接口
Token计算器

3.1 模型推理接口

3.1.1 NLP 大模型

3.1.1.1 文本对话

功能介绍

基于文本对话功能，用户可以与模型进行自然而流畅的对话和交流。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

获取URI方式请参见[请求URI](#)。

NLP推理服务提供两种推理接口调用：

- 盘古推理接口（V1推理接口）。
- 业界通用的OpenAI格式接口（V2推理接口）。

两种接口定义如[表3-1](#)所示。

表 3-1 NLP 服务推理接口

API分类	API访问路径 (URI)
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

V1推理接口URI需要输入额外参数，参数说明如[表2 V1推理接口路径参数](#)所示：

表 3-2 V1 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
 - 使用[Token认证](#)方式的请求Header参数见[表3-3](#)。

表 3-3 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。 如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

- 使用[API Key认证](#)方式的请求Header参数见[表3-4](#)。

表 3-4 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

2. V2接口仅支持API Key鉴权方式。请求Header参数见[表3-5](#)

表 3-5 请求 Header 参数 (OpenAI 格式的 API Key 认证)

参数	是否必选	参数类型	描述
Authorization	是	String	参数解释: 用户创建应用接入获取的API Key， 拼接“Bearer”后的字符串。示例： <code>Bearer d59*****9C3</code> 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

请求Body参数

V1、V2推理接口请求Body参数一致，如[表6 请求Body参数](#)描述。

表 3-6 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	<p>参数解释: 多轮对话问答对，包含两个属性：role和content。</p> <ul style="list-style-type: none">• role表示对话的角色，取值是system或user。如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。• content表示对话的内容，可以是任意文本。 <p>messages参数可以帮助模型根据对话的上下文生成合适的回复。</p> <p>约束限制: 数组长度：1 - 20</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
model	V1推理接口： 否 V2推理接口： 是	String	<p>参数解释: 使用的推理服务模型名称，为推理服务部署时指定的Deployed_Model，可在推理服务详情页面查询到。V2推理接口必须指定此参数，V1推理接口不需要此参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
stream	否	boolean	<p>参数解释: 流式开关。</p> <p>开启流式开关后，API会在生成文本的过程中，实时地将生成的文本发送给客户端，而不是等到生成完成后一次性将所有文本发送给客户端。</p> <p>约束限制: 如果开启流式，请赋值true，同时n参数只能设置为1。</p> <p>取值范围: 不涉及</p> <p>默认取值: false</p>
temperature	否	Float	<p>参数解释: 用于控制生成文本的多样性和创造力。</p> <p>参数的取值范围是(0, 1]，取值接近0表示最低的随机性，1表示最高的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，例如0.9，适合完成创造性的任务。</p> <p>temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好。</p> <p>约束限制: 不涉及</p> <p>取值范围: (0, 1]</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
top_p	否	Float	<p>参数解释: 一种替代温度采样的方法，称为nucleus sampling，其中模型考虑具有top_p概率质量的标记的结果。</p> <p>约束限制: 不涉及</p> <p>取值范围: (0, 1]</p> <p>默认取值: 不涉及</p>
max_tokens	否	Integer	<p>参数解释: 生成文本的最大token数量，单位token。</p> <p>说明 token是指模型处理和生成文本的基本单位。token可以是词或者字符的片段。模型的输入和输出的文本都会被转换成token，然后根据模型的概率分布进行采样或者计算。</p> <p>约束限制: 输入的文本加上生成的文本总量不能超过模型所能处理的最大长度。</p> <p>取值范围: 最小值: 1 最大值: 模型支持的最大token数</p> <p>默认取值: 模型支持的最大token数</p>

参数	是否必选	参数类型	描述
presence_penalty	否	Float	<p>参数解释: 用于调整模型对新Token的处理方式。即如果一个Token已经在之前的文本中出现过，那么模型在生成这个Token时会受到一定的惩罚。当presence_penalty的值为正数时，模型会更倾向于生成新的、未出现过的Token，即模型会更倾向于谈论新的话题。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: -2 最大值: 2</p> <p>默认取值: 0 (表示该参数未生效)</p>
frequency_penalty	否	Float	<p>参数解释: 用于调整模型对频繁出现的Token的处理方式。即如果一个Token在训练集中出现的频率较高，那么模型在生成这个Token时会受到一定的惩罚。当frequency_penalty的值为正数时，模型会更倾向于生成出现频率较低的Token，即模型会更倾向于使用不常见的词汇。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: -2 最大值: 2</p> <p>默认取值: 0 (表示该参数未生效)</p>

参数	是否必选	参数类型	描述
stop	否	Union[String, List[String]]	<p>参数解释: 停止符，控制模型输出何时结束的参数。可以指定一个或多个字符串作为"停止符"。当模型在生成过程中产生了任何一个停止符时，它会立即停止生成更多内容。停止符本身不会出现在最终输出中。例如：["你", "好"]，生成文本时遇到"你"或者"好"将停止文本生成。</p> <p>约束限制: 非空字符串或字符串列表。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
moderation_config	否	moderation_config objects	<p>参数解释: 盘古护栏审核配置。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
enable_search	否	boolean	<p>参数解释: 是否启动搜索增强。</p> <p>约束限制: 不涉及</p> <p>取值范围: <ul style="list-style-type: none">• false: 关闭• true: 启动</p> <p>默认取值: 不涉及</p>

表 3-7 moderation_config

参数	参数类型	描述
black_glossary_names	List<String>	参数解释: 黑名单词库列表。 约束限制: 仅对开启高级版内容审核的推理服务生效。 取值范围: 不涉及 默认取值: 不涉及
white_glossary_names	List<String>	参数解释: 白名单词库列表。 约束限制: 仅对开启高级版内容审核的推理服务生效。 取值范围: 不涉及 默认取值: 不涉及
question_moderation	boolean	参数解释: 是否开启对提示词进行内容审核，仅对订购用户生效。 约束限制: 不涉及 取值范围: <ul style="list-style-type: none">• true: 审核• false: 不审核 默认取值: true
answer_moderation	boolean	参数解释: 是否开启对推理结果进行内容审核，仅对订购用户生效。 约束限制: 不涉及 取值范围: <ul style="list-style-type: none">• true: 审核• false: 不审核 默认取值: true

参数	参数类型	描述
show_result	boolean	参数解释: 是否返回内容审核不通过原因 缺省值: false 约束限制: 不涉及 取值范围: true: 返回 false: 不返回 默认取值: 不涉及

表 3-8 message

参数	是否必选	参数类型	描述
role	V1推理接口: 否 V2推理接口: 是	String	参数解释: 对话的角色。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。多轮对话中，用户输入提示词的role设置为user，推理结果的role设置为assistant。 约束限制: 不涉及 取值范围: system、user、assistant 默认取值: 不涉及

参数	是否必选	参数类型	描述
content	是	String	参数解释: 对话的内容，可以是任意文本，单位token。 约束限制: 设置多轮对话时，message中content个数不能超过20。 取值范围: 最小长度: 1 最大长度: N1模型支持32K/128K, N2模型支持32K/128K, N4模型支持4K/32K。 默认取值: 不涉及

响应参数

非流式

状态码: 200

表 3-9 响应 Body 参数

参数	参数类型	描述
id	String	参数解释: 用来标识每个响应的唯一字符串。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 固定为"chat.completion"。 约束限制: 不涉及 取值范围: 不涉及 默认取值: chat.completion

参数	参数类型	描述
created	Integer	参数解释: 响应生成的时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
model	String	参数解释: 请求模型名称标识。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
choices	Array of choices objects	参数解释: 生成的文本列表。包含以下属性： <ul style="list-style-type: none">message: 生成的文本内容。index: 生成的文本在列表中的索引，从0开始。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
usage	usage object	参数解释: 该参数可以帮助用户了解和控制模型的使用情况，避免超出Tokens限制。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-10 choices

参数	参数类型	描述
message	message object	参数解释: 生成的文本内容。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
index	Integer	参数解释: 生成的文本在列表中的索引，从0开始。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
finish_reason	String	参数解释: 模型停止生成token的原因。 约束限制: 不涉及 取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource] <ul style="list-style-type: none">• stop：模型自然停止生成，或遇到stop序列中列出的字符串。• length：输出长度达到了模型上下文长度限制，或达到了max_tokens的限制。• content_filter：输出内容因触发过滤策略而被过滤。• tool_calls：模型决定调用外部工具（函数/API）来完成任务。• insufficient_system_resource：系统推理资源不足，生成被打断。 默认取值: stop

参数	参数类型	描述
stop_reason	Union[Integer, String]	<p>参数解释: 导致生成停止的token id或者字符串。如果是遇到EOS token则返回默认值。如果是因为用户请求参数中指定的stop参数中的字符串或者token id，则返回对应的字符串或者token id。不是openai接口标准字段，但vllm接口支持。</p> <p>缺省值: None</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-11 message

参数	参数类型	描述
role	String	<p>参数解释: 对话的角色。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
content	String	<p>参数解释: 对话的内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小长度: 1 最大长度: 不同模型支持的token长度请参见《产品介绍》“模型能力与规格”章节。</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
reasoning_content	String	参数解释: 大模型推理过程中在给出最终答案之前的推理内容。 约束限制: 该字段仅适用于Pangu-NLP-Reasoner系列模型。 取值范围: 不涉及 默认取值: 不涉及

表 3-12 usage

参数	参数类型	描述
completion_tokens	Number	参数解释: 推理服务返回结果的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
prompt_tokens	Number	参数解释: 用户输入的提示词及默认人设的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Number	参数解释: 总消耗Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

流式 (stream参数为true)

状态码： 200

表 3-13 流式输出的数据单元

参数	参数类型	描述
data	CompletionStreamResponse	<p>参数解释： stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>

表 3-14 CompletionStreamResponse

参数	参数类型	描述
id	String	<p>参数解释： 该对话的唯一标识符。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>
created	Integer	<p>参数解释： 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>

参数	参数类型	描述
model	String	参数解释: 生成该completion的模型名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 对象的类型, 其值为chat.completion.chunk。 约束限制: 不涉及 取值范围: 不涉及 默认取值: chat.completion.chunk
choices	ChatCompletionResponseStreamChoice	参数解释: 模型生成的completion的选择列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
usage	UsageInfo	参数解释: 该对话请求的Token用量信息。该参数可以帮助用户了解和控制模型的使用情况, 避免超出Tokens限制。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-15 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	参数解释: 该completion在模型生成的completion的选择列表中的索引。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
finish_reason	String	参数解释: 模型停止生成token的原因。 约束限制: 不涉及 取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource] <ul style="list-style-type: none">• stop: 模型自然停止生成, 或遇到stop序列中列出的字符串。• length : 输出长度达到了模型上下文长度限制, 或达到了max_tokens的限制。• content_filter: 输出内容因触发过滤策略而被过滤。• tool_calls: 模型决定调用外部工具(函数/API)来完成任务。• insufficient_system_resource: 系统推理资源不足, 生成被打断。 默认取值: 不涉及
delta	DeltaMessage	参数解释: V2推理接口流式返回的一个completion增量。 约束限制: V1推理接口返回体不包含此参数。 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
message	DeltaMessage	<p>参数解释: V1推理接口流式返回的一个completion增量。</p> <p>约束限制: V2推理接口返回体不包含此参数。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-16 DeltaMessage

参数	参数类型	描述
role	String	<p>参数解释: 产生这条消息的角色。</p> <p>约束限制: 不涉及</p> <p>取值范围: assistant</p> <p>默认取值: assistant</p>
content	String	<p>参数解释: completion增量的内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
reasoning_content	String	<p>参数解释: 大模型推理过程中在给出最终答案之前的推理内容（模型的思考过程）。</p> <p>约束限制: 仅适用于Pangu-NLP-Reasoner系列模型。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-17 UsageInfo

参数	参数类型	描述
prompt_tokens	Integer	参数解释: 用户输入的提示词及默认人设的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
completion_tokens	Integer	参数解释: 推理服务返回结果的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Integer	参数解释: 总消耗Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-18 内容审核拦截场景，流式输出的数据单元

参数	参数类型	描述
suggestion	String	参数解释: 审核结果: block表示未通过。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
reply	String	参数解释: 兜底回复：审核未通过时兜底回复为有效回复，兜底策略。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

接口报错的场景下，V1推理接口返回的报错信息符合华为云规范；V2推理接口则会对外透传推理服务返回的错误信息，通常符合OpenAI接口格式。

表 3-19 V1 推理接口响应错误信息 Body 参数

参数	参数类型	描述
error_msg	String	参数解释: 错误信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
error_code	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
details	List<Object>	参数解释: 推理服务返回的报错信息，具体的格式、内容取决于推理服务。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-20 V2 推理接口响应错误信息 Body 参数

参数	参数类型	描述
error	ErrorResp	参数解释: 错误信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
id	String	参数解释: 请求ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-21 ErrorResp

参数	参数类型	描述
code	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
type	String	参数解释: 错误类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
message	String	参数解释: 错误详情。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

- 单轮问答

V1推理接口:

POST https://[endpoint]/v1/{project_id}/deployments/{deployment_id}/chat/completions

Request Header:

Content-Type: application/json

X-Auth-Token:

MIINRwYJKoZIhvcNAQcCoIIINODCCDTQCAQExDTALBglghkgBZQMEA...gguVBgkqhkiG...

Request Body:

```
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user",  
            "content": "介绍下长江，以及长江中典型的鱼类"  
        }  
    ]  
}
```

```
        ],
        "temperature": 0.9,
        "max_tokens": 600
    }
}

V2推理接口:  
POST https://{{endpoint}}/api/v2/chat/completions

Request Header:  
Content-Type: application/json  
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

Request Body:  
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user",  
            "content": "介绍下长江，以及长江中典型的鱼类"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

- 单轮流式问答 (stream参数值为true)

```
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user",  
            "content": "五岳分别是哪些山"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600,  
    "stream": "true"  
}
```

- 带有人设的单轮问答 (role参数值为system)

```
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "system",  
            "content": "你的名字叫智子，是一名幼儿园老师，请用幼儿园老师的口吻回答问题，注意语气温和亲切，通过提问、引导、赞美等方式，激发学生的思维和想象力."  
        },  
        {  
            "role": "user",  
            "content": "介绍下长江，以及长江中典型的鱼类"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

- 带有人设的单轮流式问答 (role参数值为system, stream参数值为true)

```
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "system",  
            "content": "你的名字叫智子，是一名幼儿园老师，请用幼儿园老师的口吻回答问题，注意语气温和亲切，通过提问、引导、赞美等方式，激发学生的思维和想象力."  
        },  
        {  
            "role": "user",  
            "content": "写一首诗"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

```
        "temperature": 0.9,  
        "max_tokens": 600,  
        "stream": "true"  
    }  
  
● 多轮问答  
# 多轮问答传参方法：在完成第一轮问答，进行第二轮问答时，需要将第一轮的问题和答案、第二轮问题作为参数传入……依次类推，完成多轮对话。  
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user",  
            "content": "介绍下长江，以及长江中典型的鱼类" //第一轮问题  
        },  
        {  
            "role": "assistant",  
            "content": "长江是中国第一大河,全长6300多公里。它发源于唐古拉山脉南麓,流经中国11个省市自治区和两个特别行政区,最终注入东海。\\n长江中典型的鱼类有:鲢鱼、鳙鱼、草鱼、鲤鱼等。" //第一轮答案  
        },  
        {  
            "role": "user",  
            "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题  
        },  
        {  
            "role": "assistant",  
            "content": "1. 湖南 - 长沙岳麓山、张家界天门山等\\n2. 四川 - 九寨沟、峨眉山等" //第二轮答案  
        },  
        {  
            "role": "user",  
            "content": "提到的湖南景点，详细说明下" //第三轮问题  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

```
● 多轮流式问答（stream参数值为true）  
# 多轮问答传参方法：在完成第一轮问答，进行第二轮问答时，需要将第一轮的问题和答案、第二轮问题作为参数传入……依次类推，完成多轮对话。  
{  
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user",  
            "content": "介绍下长江，以及长江中典型的鱼类" //第一轮问题  
        },  
        {  
            "role": "assistant",  
            "content": "长江是中国第一大河,全长6300多公里。它发源于唐古拉山脉南麓,流经中国11个省市自治区和两个特别行政区,最终注入东海。\\n长江中典型的鱼类有:鲢鱼、鳙鱼、草鱼、鲤鱼等。" //第一轮答案  
        },  
        {  
            "role": "user",  
            "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题  
        },  
        {  
            "role": "assistant",  
            "content": "1. 湖南 - 长沙岳麓山、张家界天门山等\\n2. 四川 - 九寨沟、峨眉山等" //第二轮答案  
        },  
        {  
            "role": "user",  
            "content": "张家界具体在哪里" //第三轮问题  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600,
```

- 带有人设的多轮流式问答 (role参数值为system, stream参数值为true)

```
        "stream": "true"
    }
}

{
    "model": "pangu-nlp-n1-32k", // 仅V2接口需要此参数
    "messages": [
        {
            "content": "你是一个热心的导游", //人设
            "role": "system"
        },
        {
            "role": "user",
            "content": "介绍下长江，以及长江中典型的鱼类" //第一轮问题
        },
        {
            "role": "assistant",
            "content": "长江，亚洲最长河流，世界第三长河流，它发源于唐古拉山脉南麓，流经我国多个省份，最终注入东海。长江流域资源丰富，特别是鱼类资源，例如：鮰鱼、白鲟、白鳍豚等，都是长江特有的珍稀物种。" //第一轮答案
        },
        {
            "role": "user",
            "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题
        },
        {
            "role": "assistant",
            "content": "长江流经的省份众多，这里为您介绍两个：湖北和四川。湖北，长江流经此省9年，有武汉、宜昌等知名城市，人文风景丰富，如黄鹤楼、古琴台等。四川，长江流经此省14年，有成都、重庆等大城市，人文风景包括都江堰、乐山大佛等世界文化遗产。" //第二轮答案
        },
        {
            "role": "user",
            "content": "张家界具体在哪里" //第三轮问题
        }
    ],
    "temperature": 0.5,
    "max_tokens": 800,
    "stream": true
}
```

响应示例

状态码： 200

OK

- 非流式问答响应

```
{
    "id": "chat-3f7d8e038f244ae5b7cb0cecf352e154",
    "object": "chat.completion",
    "created": 1745725506,
    "model": "pangu-nlp-n1-32k_kpyq10",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "你好！有什么我可以帮助你的吗？",
                "tool_calls": []
            },
            "logprobs": null,
            "finish_reason": "stop",
            "stop_reason": null
        }
    ],
    "usage": {
        "prompt_tokens": 64,
```

```
        "total_tokens": 73,  
        "completion_tokens": 9  
    },  
    "prompt_logprobs": null  
}
```

- 带有思维链的非流式问答响应

```
{  
    "id": "4c0dcef2-7f8f-4c57-93c7-484c03b0a216",  
    "object": "chat.completion",  
    "created": 1741937420,  
    "model": "pangu-nlp-n1-32k_kpyq10",  
    "choices": [  
        {  
            "index": 0,  
            "message": {  
                "role": "assistant",  
                "content": "\n\n你好！很高兴见到你，有什么我可以帮忙的吗？",  
                "reasoning_content": "嗯，用户发来“你好”，这是一个常见的中文问候。我要用中文回应用户，保持友好和亲切。首先要分析用户的需求，可能只是想打个招呼，或者有具体的问题想问。\\n\\n接下来要考虑用户可能的背景，可能来自不同的地方，但使用简体中文，所以可能是中国大陆的用户。需要保持回答的通用性，适合所有使用中文的人。\\n\\n还要判断是否有更深层次的需求，用户可能在测试聊天机器人，或者真的需要帮助。这时候需要给出一个开放式的回应，鼓励用户进一步说明需求。例如询问有什么可以帮忙的，这样既回应了问候，又引导用户提出具体问题。\\n\\n另外要注意语言的流畅和自然，避免过于机械化的回答。保持礼貌和专业，同时适当使用表情符号增加亲和力，比如用[微笑]或者[]之类的，不过可能根据平台的不同需要调整。\\n\\n最后要确保回答符合规范，不涉及敏感话题，内容积极健康。检查是否有拼写或语法错误，确保回复准确无误。\\n",  
                "tool_calls": []  
            },  
            "logprobs": null,  
            "finish_reason": "stop",  
            "stop_reason": null  
        }  
    ],  
    "usage": {  
        "prompt_tokens": 22,  
        "total_tokens": 377,  
        "completion_tokens": 355  
    },  
    "prompt_logprobs": null  
}
```

- 流式问答响应

V1推理接口返回体:

```
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":174572  
5837,"model":"pangu-nlp-n1-32k_kpyq10","choices":  
[{"index":0,"logprobs":null,"finish_reason":null,"message":{"role":"assistant"}],"usage":  
{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":174572  
5837,"model":"pangu-nlp-n1-32k_kpyq10","choices":  
[{"index":0,"logprobs":null,"finish_reason":null,"message":{"content":"你好"}],"usage":  
{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":174572  
5837,"model":"pangu-nlp-n1-32k_kpyq10","choices":  
[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"！有什么我  
可以帮你的吗？"}],"usage": {"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":174572  
5837,"model":"pangu-nlp-n1-32k_kpyq10","choices":[],"usage":  
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}  
  
event:{ "usage":  
{"completionTokens":9,"promptTokens":64,"totalTokens":73}, "tokens":64, "token_number":9}
```

```
data:[DONE]
V2推理接口返回体:
data:{ "id": "chat-
b9417f06b6524362ae09844cc9b0172d", "object": "chat.completion.chunk", "created": 1745725924, "model": "pangu-nlp-n1-32k_kpyq10", "choices": [ { "index": 0, "delta": { "role": "assistant", "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 64, "total_tokens": 64, "completion_tokens": 0 } } ] }

data:{ "id": "chat-
b9417f06b6524362ae09844cc9b0172d", "object": "chat.completion.chunk", "created": 1745725924, "model": "pangu-nlp-n1-32k_kpyq10", "choices": [ { "index": 0, "delta": { "content": "你好" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 64, "total_tokens": 65, "completion_tokens": 1 } ] }

data:{ "id": "chat-
b9417f06b6524362ae09844cc9b0172d", "object": "chat.completion.chunk", "created": 1745725924, "model": "pangu-nlp-n1-32k_kpyq10", "choices": [ { "index": 0, "delta": { "content": "！有什么我可以帮助你的吗？" }, "logprobs": null, "finish_reason": "stop", "stop_reason": null }, "usage": { "prompt_tokens": 64, "total_tokens": 73, "completion_tokens": 9 } ] }

data:{ "id": "chat-
b9417f06b6524362ae09844cc9b0172d", "object": "chat.completion.chunk", "created": 1745725924, "model": "pangu-nlp-n1-32k_kpyq10", "choices": [ ], "usage": { "prompt_tokens": 64, "total_tokens": 73, "completion_tokens": 9 } }

data:[DONE]
```

- 带有思维链的流式问答响应

```
V1推理接口返回体:
data:{ "id": "chat-
cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "role": "assistant", "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 23, "completion_tokens": 0 } } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 24, "completion_tokens": 1 } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 25, "completion_tokens": 2 } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 26, "completion_tokens": 3 } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "好的" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 27, "completion_tokens": 4 } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "用户" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 28, "completion_tokens": 5 } ] }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "发" }, "logprobs": null, "finish_reason": null }, "usage": { "prompt_tokens": 23, "total_tokens": 30, "completion_tokens": 7 } ] }
```

```
data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "来了"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 31, "completion_tokens": 8} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": " "}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 32, "completion_tokens": 9} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "你好"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 33, "completion_tokens": 10} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": " " }, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 34, "completion_tokens": 11} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "我需要"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 35, "completion_tokens": 12} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "回答"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 36, "completion_tokens": 13} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "他们"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 38, "completion_tokens": 15} }

...
data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "符合"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 178, "completion_tokens": 155} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "要求"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 179, "completion_tokens": 156} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "。"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 180, "completion_tokens": 157} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": "\n"}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 181, "completion_tokens": 158} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [ { "index": 0, "message": { "reasoning_content": ""}, "logprobs": null, "finish_reason": null}], "usage": { "prompt_tokens": 23, "total_tokens": 182, "completion_tokens": 159} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created":
```

```
1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 183, "completion_tokens": 160}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "\n"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 184, "completion_tokens": 161}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "你好"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 185, "completion_tokens": 162}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "我是"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 186, "completion_tokens": 163}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "华为"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 187, "completion_tokens": 164}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "云"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 189, "completion_tokens": 166}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "开发的"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 190, "completion_tokens": 167}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "盘"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 191, "completion_tokens": 168}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "古"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 192, "completion_tokens": 169}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "大"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 193, "completion_tokens": 170}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "语言"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 194, "completion_tokens": 171}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "模型"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 195, "completion_tokens": 172}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "。"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 196, "completion_tokens": 173}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "有什么"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 197, "completion_tokens": 174}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "可以"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 198, "completion_tokens": 175}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "帮"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 199, "completion_tokens": 176}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "您的"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 200, "completion_tokens": 177}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": "?"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 201, "completion_tokens": 178}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "message": {"content": ""}, "logprobs": null, "finish_reason": "stop", "stop_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 202, "completion_tokens": 179}}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [], "usage": {"prompt_tokens": 23, "total_tokens": 202, "completion_tokens": 179}}
```

```
data:[DONE]  
V2推理接口返回体:  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"role": "assistant"}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 23, "completion_tokens": 0}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 24, "completion_tokens": 1}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 25, "completion_tokens": 2}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 26, "completion_tokens": 3}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "好的"}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 27, "completion_tokens": 4}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": " ", "logprobs": null, "finish_reason": null}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 28, "completion_tokens": 5}}]}
```

```
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"logprobs": null, "finish_reason": null}, "logprobs": null, "finish_reason": null}, {"usage": {"prompt_tokens": 23, "total_tokens": 28, "completion_tokens": 5}}]}
```

```
1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "用户"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 29, "completion_tokens": 6}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "发"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 30, "completion_tokens": 7}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "来了"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 31, "completion_tokens": 8}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": " "}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 32, "completion_tokens": 9}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "你好"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 33, "completion_tokens": 10}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": " " " "}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 34, "completion_tokens": 11}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "我需要"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 35, "completion_tokens": 12}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "回复"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 37, "completion_tokens": 14}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "他们"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 38, "completion_tokens": 15}]

...
data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "符合"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 178, "completion_tokens": 155}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "要求"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 179, "completion_tokens": 156}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "。"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 180, "completion_tokens": 157}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": "\n"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23,
```

```
"total_tokens": 181, "completion_tokens": 158}]

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 182, "completion_tokens": 159} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 183, "completion_tokens": 160} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "\n\n"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 184, "completion_tokens": 161} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "你好"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 185, "completion_tokens": 162} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "！"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 186, "completion_tokens": 163} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "我是"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 187, "completion_tokens": 164} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "华为"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 188, "completion_tokens": 165} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "云"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 189, "completion_tokens": 166} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "开发的"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 190, "completion_tokens": 167} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "盘"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 191, "completion_tokens": 168} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "古"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 192, "completion_tokens": 169} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "大"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 193, "completion_tokens": 170} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "语言"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 194, "completion_tokens": 171} }

data:{ "id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "模型"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 195, "completion_tokens": 172} }
```

```
    "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 195, "completion_tokens": 172}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "。"}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 196, "completion_tokens": 173}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "有什"}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 197, "completion_tokens": 174}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "么"}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 198, "completion_tokens": 175}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "帮"}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 199, "completion_tokens": 176}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "您的"}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 200, "completion_tokens": 177}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": "? "}}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 23, "total_tokens": 201, "completion_tokens": 178}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [{"index": 0, "delta": {"content": ""}}, "logprobs": null, "finish_reason": "stop", "stop_reason": null], "usage": {"prompt_tokens": 23, "total_tokens": 202, "completion_tokens": 179}}  
  
data:{"id": "chat-cfc9f0e8404540af97aa97bfc5c5090f", "object": "chat.completion.chunk", "created": 1747364482, "model": "pangu-nlp-n2-reasoner-32k", "choices": [], "usage": {"prompt_tokens": 23, "total_tokens": 202, "completion_tokens": 179}}  
  
data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2 CV 大模型

3.1.2.1 Pangu-CV-图像分类-2.1.0

功能介绍

根据在图像信息中所反映的不同特征，对图像进行定量分析，把图像划归为若干个类别中的某一种。适用于动植物分类、车辆类型分类、车牌分类、废钢定级、零部件分类等任务。

说明书

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/

获取URI方式请参见[请求URI](#)。

表 3-22 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	<p>参数解释： 项目ID，获取方法请参见获取项目ID。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>
deployment_id	是	String	<p>参数解释： 模型的部署ID，获取方法请参见获取模型部署ID。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-23](#)。

表 3-23 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-24](#)。

表 3-24 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-25 请求 Body 参数

参数	是否必选	参数类型	描述
images	是	String/ List[String]	参数解释: 被检测图片的base64编码。 约束限制: <ul style="list-style-type: none">建议单次请求体大小不超过4M。建议使用JPG、PNG、JPEG、BMP格式的图片。默认只支持RGB三通道图片数据，其他通道的数据暂不支持。单张图片请求时，参数类型为String，为图片的base64编码；批量请求时，参数类型为List[String]，以列表形式存放各图片的base64编码，单次请求不超过24张。 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
mode	否	String	<p>参数解释: 取值为"single"或者"multiple"，分别表示单标签分类和多标签分类模式，二者取一，默认设置为训练得到的模型对应的模式。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• single: 单标签分类• multiple: 多标签分类 <p>默认取值: 不涉及</p>
threshold	否	dict	<p>参数解释: 多标签分类时，各标签对应的预测得分阈值，预测得分小于阈值的预测结果将会被过滤。</p> <p>约束限制: 仅在多标签分类模式下有效。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
top	否	int	<p>参数解释: 单标签分类时，用于规定输出 top N 的预测得分对应的预测结果。</p> <p>约束限制: 仅在单标签分类模式下有效。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

响应参数

状态码: 200

响应成功返回的结构体是一个Dict，由本次请求的多张输入图像的预测结果组成，图像与图像之间通过编号（键）区分。

表 3-26 单/多标签分类响应成功 Body 参数

参数	参数类型	描述
键	String	参数解释: 输入图像的顺序编号，从"0"开始，最大不超过"23"。 约束限制: 不涉及 取值范围: 0-23 默认取值: 不涉及
值	List[Dict]	参数解释: 当前编号的图像对应的预测结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
dataset_id	String	参数解释: 训练数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

其中，每张图像的预测结果参数类型为List[Dict]，表示预测得到的一个或多个类别，每个Dict的参数内容请参考[表3-27](#)。

表 3-27 单张图像的单个类别预测结果参数

参数	参数类型	描述
label	String	参数解释: 预测的类别，与训练数据中定义的类别相同。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
score	String	参数解释: 预测的置信度结果，输出每个标签对应的预测得分，得分区间为0至1。 约束限制: 不涉及 取值范围: 0-1 默认取值: 不涉及

状态码： 400

表 3-28 响应失败 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

单图请求示例：

```
{  
  "images": "/9j/4Vr2RXhpZgAASUkqAAgAAA.....",  
}
```

批量请求示例（单次请求批次上限为24张图）：

```
{  
  "images": ["/9j/4Vr2RXhpZgAASUkqAAgAAA.....", "/9j/4RlrRXhpZgAATU....."]  
}
```

带高级参数的单标签分类请求示例：

```
{  
  "images": ["/9j/4Vr2RXhpZgAASUkqAAgAAA.....", "/9j/4RlrRXhpZgAATU....."],  
  "label": "apple",  
  "score": "0.99",  
  "threshold": 0.5  
}
```

```
    "top": 3
}
```

带高级参数的多标签分类请求示例：

```
{
  "images": ["/9j/4Vr2RXhpZgAASUkqAAgAAA.....", "/9j/4RlrRXhpZgAATU....."],
  "threshold":
  {
    "bird": 0.33,
    "blackbird": 0.44
  }
}
```

响应示例

响应返回一个字典，键为本次请求的输入图像的编号，对输入的图像按顺序从0开始编号；值为一个列表，列表里为该图像的预测结果，每张图像可能存在多个预测结果（例如多标签分类模式）。

```
{
  "0": [
    {
      "label": "bird",
      "score": "0.95511043"
    },
    {
      "label": "blackbird",
      "score": "0.75241840"
    }
  ],
  "1": [
    {
      "label": "bird",
      "score": "0.36211243"
    }
  ],
  "dataset_id": "1341002014632579072"
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.2 Pangu-CV-物体检测-S-2.1.0

功能介绍

找出图像中所有感兴趣的目标，确定它们的位置和类别。物体检测-S模型特点是小参数量，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。

说明

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/
获取URI方式请参见[请求URI](#)。

表 3-29 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-30](#)。

表 3-30 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-31](#)。

表 3-31 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-32 请求 Body 参数

参数	是否必选	参数类型	描述
images	是	String	参数解释: 被检测图片的base64编码。 约束限制: <ul style="list-style-type: none">建议整个请求体大小不超过4M。建议使用JPG、PNG、JPEG、BMP格式的图片。默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 取值范围: 不涉及 默认取值: 不涉及
threshold	否	Float	参数解释: 检测框置信度阈值。 约束限制: 不涉及 取值范围: 最小0.0，最大1.0 默认取值: 0.25

响应参数

状态码： 200

表 3-33 响应成功 Body 参数

参数	参数类型	描述
result	Array of objects	参数解释: 识别结果信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Score	Float	参数解释: 置信度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
label	String	参数解释: 检测类别。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Box	Dict	参数解释: 检测到的目标主体信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-34 Box

参数	参数类型	描述
X	Int	参数解释: 矩形框左上角横坐标。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Y	Int	参数解释: 矩形框左上角纵坐标。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
width	Int	参数解释: 矩形框宽度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Height	Int	参数解释: 矩形框高度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
Angle	Int	参数解释: 检测到的目标主体区域的角度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-35 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "images": "/9j/4Vr2RXhpZgAASUkqAAgAAA.....",  
    "threshold": 0.3  
}
```

响应示例

```
{  
    "result": [  
        {  
            "Box": {  
                "Angle": 0,  
                "Height": 60,  
                "Width": 106,  
                "X": 852,  
                "Y": 182  
            },  
            "Score": 0.88427734375,  
            "label": "car"  
        },  
        {  
            "Box": {  
                "Angle": 0,  
                "Height": 114,  
                "Width": 55,  
                "X": 800,  
                "Y": 170  
            },  
            "Score": 0.70556640625,  
            "label": "person"  
        }  
    ]  
}
```

}]

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.3 Pangu-CV-物体检测-N-2.1.0

功能介绍

找出图像中所有感兴趣的目标，确定它们的位置和类别。物体检测-N模型特点是参数量适中，适合在资源有限的环境中使用，提供较快的检测速度和合理的精度。

说明书

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

URI

图片接口：POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/

获取URI方式请参见[请求URI](#)。

表 3-36 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释： 项目ID，获取方法请参见 获取项目ID 。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

参数	是否必选	参数类型	描述
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-38](#)。

表 3-37 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-243](#)。

表 3-38 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-39 图片请求 Body 参数

参数	是否必选	参数类型	描述
images	是	String	参数解释: 被检测图片的base64编码。 约束限制: <ul style="list-style-type: none">建议使用PNG、JPEG、BMP、JPG、WEBP格式的图片。只支持单张图片输入，分辨率范围为1px-10000px，且长短边比例不能高于5。并且base64编码后的图片大小不超过10MB。支持RGB三通道格式的图像。 取值范围: 不涉及 默认取值: 不涉及
nms_iou_thr	是	Float	参数解释: 极大值抑制阈值。 约束限制: 不涉及 取值范围: 0.0~1.0 默认取值: 不涉及
agnositc_nms	是	Bool	参数解释: 是否进行类间nms，是填写true，不是填写false。 约束限制: 不涉及 取值范围: true：进行类间nms 默认取值: 不涉及

响应参数

状态码： 200

表 3-40 响应 Body 参数

参数	参数类型	描述
result	List	<p>参数解释: 物体检测的识别结果。物体检测的基本目标是在输入的图像或视频中找到感兴趣的目标物体，并确定它们的位置和类别，所以识别结果包含物体的类别和物体的位置。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-41 响应 Body 参数

参数	参数类型	描述
RegisterMatrix	List	<p>参数解释: 默认为$[[1, 0, 0], [0, 1, 0], [0, 0, 1]]$，表示图片特征矩阵。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
Label	String	<p>参数解释: 预测类别。物体检测中的预测类别是指模型对输入图像或视频中物体的分类结果。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
Score	Float	参数解释: 置信度。用于衡量模型对预测结果的准确性。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Box	Dict	参数解释: 检测到的目标主体信息。 约束限制: 格式为{"x":x1,"y":y1,"width":w,"height":h, 'Angle':angle}。 <ul style="list-style-type: none">• x: 检测到的目标主体区域的左上角x坐标。• y: 检测到的目标主体区域的左上角y坐标。• width: 检测到的目标主体区域的宽度。• height: 检测到的目标主体区域的高度。• angle: 检测到的目标主体区域的角度。 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-42 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

- 图片请求示例

```
{  
    "nms_iou_thr":0.45,  
    "agnositc_nms":false,  
    "images": "/9j/4Vr2RXhpZgAASUkqAAgAAA...."  
}
```

响应示例

```
{  
    "result": [  
        {  
            "RegisterMatrix": [  
                [  
                    1,  
                    0,  
                    0  
                ],  
                [  
                    0,  
                    1,  
                    0  
                ],  
                [  
                    0,  
                    0,  
                    1  
                ]  
            ]  
        },  
        {  
            "Box": {  
                "Y": 0,  
                "Width": 100,  
                "Angle": 0,  
                "X": 0,  
                "Height": 100  
            },  
            "Score": 0.9,  
            "label": "person"  
        }  
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.4 Pangu-CV-物体检测-S-3.1.0

功能介绍

盘古计算机视觉物体检测大模型，任务是找出图像中所有感兴趣的目标，确定它们的位置和类别。

说明书

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/

获取URI方式请参见[请求URI](#)。

表 3-43 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-44 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-45 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。API Key认证响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-46 请求 Body 参数

参数	是否必选	参数类型	说明
images	是	String	参数解释: 图片base64编码，图片限制大小为2K。默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 约束限制: <ul style="list-style-type: none">• 图片限制大小为2K。• 建议使用JPG、PNG、JPEG、BMP格式的图片。• 默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	说明
threshold	否	String	<p>参数解释: 检测框置信度阈值。 使能逻辑: 1) 优先看是否存在ENABLE_ALL_OUTPUTS参数为true, 如存在则优先输出所有结果, 不使用输入threshold和默认最优阈值; 2) 如果传入threshold, 且ENABLE_ALL_OUTPUTS=false, 则优先使用传入的threshold进行结果筛选; 3) 如果不传入threshold, 或threshold=-1, 且ENABLE_ALL_OUTPUTS=false, 则使用默认最优阈值进行推理。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小0.0, 最大1.0</p> <p>默认取值: -1</p>

响应参数

表 3-47 响应 Body

参数	参数类型	说明
result	List	<p>参数解释: 物体检测的识别结果。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	说明
dataset_id	String	参数解释: 训练数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-48 result 中各项结构

参数	参数类型	说明
RegisterMatrix	List	参数解释: 图片特征矩阵。 参数解释: 置信度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 默认为[[1, 0, 0], [0, 1, 0], [0, 0, 1]]
Label	String	参数解释: 预测类别，与训练数据中的类别名称相关。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	说明
Score	Float	参数解释: 置信度。 约束限制: 不涉及 取值范围: 0~1 默认取值: 不涉及
Box	Dict	参数解释: 目标主体信息，格式为 {"x":x1,"y":y1,"width":w,"height":h,"angle":r}。 • x: 检测到的目标主体区域的左上角x坐标。 • y: 检测到的目标主体区域的左上角y坐标。 • width: 检测到的目标主体区域的宽度。 • height: 检测到的目标主体区域的高度。 • angle: 默认为0，检测区域角度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

有如下三种调用方式：

调用方式总结为如下三种：

1) form-data格式调用

```
{  
    "file":"xxx(base64 encode image data)"  
}
```

2) base64 json调用

```
{  
    "Files":  
        [{"ImageData":"xxx(base64 encode image data)"}]  
}
```

3) base64 json调用 (threshold非必须)

```
{  
    "images":"xxx(base64 encode image data)",  
    "threshold":"0.4"  
}
```

响应示例

```
{  
    "dataset_id": "12345",  
    "result": [  
        {  
            "RegisterMatrix": [  
                [  
                    [1,  
                     0,  
                     0],  
                    [0,  
                     1,  
                     0],  
                    [0,  
                     0,  
                     1]  
                ]  
            ],  
            "Box": {  
                "X": 0,  
                "Y": 0,  
                "Width": 100,  
                "Height": 100,  
                "Angle": 0  
            },  
            "Score": 0.9,  
            "label": "person"  
        }  
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.5 Pangu-CV-语义分割-2.1.0

功能介绍

将数字图像细分为多个图像子区域，适用于车道分割、建筑分割、选煤厂筛面状态识别等任务。

说明

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/v1/rs-server/predictions

获取URI方式请参见[请求URI](#)。

表 3-49 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-50 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-51 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。API Key认证响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-52 请求 Body 参数

参数	是否必选	参数类型	说明
img	是	String	参数解释: 图片base64编码，图片限制大小为2K。默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 约束限制: <ul style="list-style-type: none">• 图片限制大小为2K。• 默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 取值范围: 不涉及 默认取值: 不涉及

响应参数

表 3-53 响应 Body

参数	参数类型	说明
result	Array	参数解释: 预测结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-54 result 中各项结构

参数	参数类型	说明
result	Array	参数解释: 预测结果。预测结果是一个二维数组，表示模型对输入图像的每个像素点进行语义分割后的预测结果。数组中的每一个元素都是一个整数，表示模型对该像素点的预测类别。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
{  
    "img": "/9j/4Vr2RXhpZgAASUkqAAgAAA....."  
}
```

响应示例

```
{  
    "result": [  
        [  
            0,  
            ...  
        ]  
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.6 Pangu-CV-实例分割-1.1.0

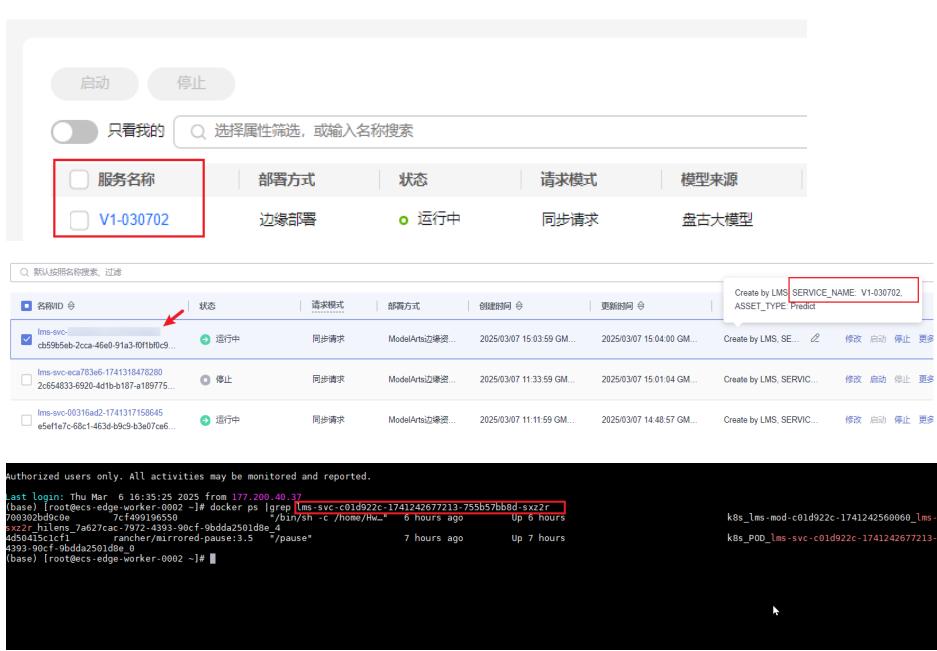
功能介绍

能够对输入图片中的不同类别物体及物体个体进行分割识别，输出每个实例的类别标签、置信度及坐标信息。

说明

服务调用方式：

- 支持图片推理和视频推理。
- 图片、视频推理支持在线和边缘部署。
- 视频推理调用方式，需在创建部署任务时设置环境变量，添加rtsp视频流地址：
 - 添加环境变量，名称“ADDRS”。环境变量的值为视频流地址，格式示例：rtsp://[边缘节点地址:端口}/{rtsp视频流地址}
 - 模型推理结果需在容器日志中查询，远程登录到部署服务器后，执行**docker ps**命令获取容器列表及其信息。
 - 记录部署任务名称，进入ModelArts服务的模型部署导航栏中按照模型的部署形式，进入在线服务、边缘服务页签按照任务名称找到对应的任务ID。按创建部署任务时生成的任务ID在容器信息中查找，找到与名称对应的CONTAINER ID。

模型部署 | 总服务数 136 | 运行中 1 | 异常失败 18 | 已停止 117

名称ID	状态	请求模式	部署方式	模型来源	操作
lms-svc-00316ad2-1741317150645 e5effe7c-68c1-463d-b9c9-b3a07ce6...	运行中	同步请求	ModelArts边缘部署	2025/03/07 11:11:59 GM...	Create by LMS, SERVIC... 修改 启动 停止 更多
lms-svc-783e6-1741318478280 2c654833-6920-4d1b-b187-a185775...	停止	同步请求	ModelArts边缘部署	2025/03/07 11:33:59 GM...	Create by LMS, SERVIC... 修改 启动 停止 更多
lms-svc-cb59b5eb-2cca-46e0-91a3-10f1b0c0-9...	运行中	同步请求	ModelArts边缘部署	2025/03/07 15:03:59 GM...	Create by LMS, SERVIC... 修改 启动 停止 更多

- 执行**docker logs -f {CONTAINER ID}**命令查看容器日志。在容器日志中可查看推理执行过程，搜索“result”可获取推理结果。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/
获取URI方式请参见[请求URI](#)。

请求参数

表 3-55 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-56 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-57 请求参数

参数	是否必选	参数类型	说明
images	是	String	参数解释: 图像base64编码。 约束限制: 默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 取值范围: 不涉及 默认取值: 不涉及

响应参数

表 3-58 响应参数

参数	参数类型	描述
img_res	List	参数解释: 实例分割结果，请求成功有此字段，其中每个元素为一个实例的分割结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
dataset_id	String	参数解释: 训练数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-59 img_res 中各项结构

参数	参数类型	描述
label	String	参数解释: 预测类别。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
score	Float	参数解释: 置信度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
box	Dict	参数解释: 检测到的目标主体信息，格式为 {"x":x1,"y":y1,"width":w,"height":h,"angle":r}。 • x: 检测到的目标主体区域的左上角x坐标。 • y: 检测到的目标主体区域的左上角y坐标。 • width: 检测到的目标主体区域的宽度。 • height: 检测到的目标主体区域的高度。 • angle: 默认为0，检测区域角度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
mask	Dict	参数解释: 图像信息，格式为{"counts":string,"size": [h,w]}。 • counts: 基于游程编码的字符串，编码内容为和原图宽高相同的布尔数组：若数组某位置值为0，代表原图此位置像素点不属于检测目标，若为1，代表原图此位置像素点属于检测目标。可通过python开源库 pycocotools._mask进行编码和解码。 • size: 图像尺寸，h为图像高，w为图像宽。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
{  
    "images": "/9j/4Vr2RXhpZgAASUkqAAgAAA...."  
}
```

响应示例

```
{  
    "dataset_id": "12345",  
    "img_res":  
        [{"label": "person", "score": 0.958,  
         "box": {"x": 131, "y": 186, "width": 128, "height": 102, "angle": 0},  
         "mask": {"counts": "PR`1h0n:0O2M5K4L4M4L6I7\\O\\  
\\NiFh1R9a0N2M300M301O00001O0000000000003NO01N200N3O2Nd0]O1N1O0O2O1N20000O100O2O  
O1O1N3N1N2N2O1N1O1O1O1O00001O1O1O2N3M2ZFiNi8R2001O00001O00001O001O1O2N1O1  
O2N3M7F9I5J4M3N1O3M3L3M3N1N3M2M3MRfR3", "size": [374, 500]}  
        }, ...]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.7 Pangu-CV-万物检测-V2

功能介绍

该模型主要实现物体检测的功能，输入图片的base64码和[自定义的标签列表](#)，模型识别出图片中该类别所在的位置，会返回带有标签的检测框信息和画好框的图片base64编码。

说明书

服务调用方式：

- 支持图片推理。
- 图片推理支持在线和边缘部署。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

表 3-60 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是 (云上部署 模型填写)	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5中响应消息头中X-Subject- Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 参数值为“application/json”

表 3-61 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig- AppCode	是 (云上部署 模型填写、边 缘部署模型仅 在监控网关访 问模式下填 写)	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig- AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 参数值为“application/json”

表 3-62 请求 Body 参数

参数	是否必选	参数类型	描述
images	是	String	参数解释: 被检测图片的base64编码。 约束限制: 支持识别PNG、JPEG、BMP、JPG、WEBP格式的图片。只支持单张图片输入，分辨率范围为1px-1000px，base64编码后的图片大小不超过10MB，且长短边比例不能高于5。默认只支持RGB三通道图片数据，其他通道的数据暂不支持。 取值范围: 不涉及 默认取值: 不涉及
threshold	是	Float	参数解释: 检测框置信度的阈值。 约束限制: 不涉及 取值范围: [0, 1] 默认取值: 不涉及

响应参数

状态码： 200

表 3-63 响应 Body 参数

参数	参数类型	描述
result	MultimodalDINOResult object	参数解释: 检测结果信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-64 MultimodalDINOResult

参数	参数类型	描述
Box	object	参数解释: 目标检测框坐标, [检测框左上角到竖轴距离, 目标框左上角到横轴距离, 检测框宽度, 检测框高度], {"X": x, "Y": y, "Width": w, "Height": h} 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Score	Float	参数解释: 目标检测置信度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
label	String	参数解释: 目标标签。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-65 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "threshold": 0.25,  
    "images": "/9j/4AAQSkZJRgABAQAAAQABAAAD/....."  
}
```

响应示例

```
{  
    "result": [  
        {  
            "Box": {  
                "Y": 0,  
                "Width": 100,  
                "X": 0,  
                "Height": 100  
            },  
            "Score": 0.9,  
            "label": "person"  
        }  
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2.8 Pangu-CV-视觉交互检测-V3

功能介绍

盘古视觉交互检测大模型，支持通过文本或者视觉提示来找出目标图像中所有感兴趣的目标，确定它们的位置和类别。

视觉交互检测推理分为“开集模型推理”和“闭集模型推理”，其中：

- **开集模型推理：**

该模式下用户可以给出想要检测的目标少量的视觉提示和文本提示（单一视觉提示或文本提示亦可），即可在待检测图像上检测用户指定目标类目。该过程整体分为提示特征获取和提示特征推理两个步骤：

- 步骤1 提示特征获取：**这一步的目的是根据用户给出的视觉提示或者文本提示生成对应的提示特征编码用于后续的特征推理，由于用户无法一次确定给出的提示能否达到符合预期的检测效果，因此该过程通常是交互式进行，用户通过尝试不同的提示然后观察在待检测图上的效果，选择符合预期效果的提示得到提示特征编码用户后续的特征推理。
- 步骤2 提示特征推理：**基于上一步得到的提示特征编码，然后即可在待检测目标图像上进行快速的目标检测。

----结束

□ 说明

需要注意的是：

1. 第二步基于特征进行推理而不直接基于用户提示直接端到端进行目标检测的原因是提示特征提取比较耗费时间，期间开集检测大模型会进行多次视觉或者文本推理非常耗时。对于已经确定的用户提示，这些提示对应的提示特征是一样的，无需重复提取。因此在实际大规模进行目标检测时，需要先获取提示特征，再基于这些提示特征进行快速推理。
2. 对于使用自己的数据集重新预训练或者微调的数据集进行检测，可以直接仅使用文本提示，将训练类目名称的部分或者全部类目名称使用"."进行拼接，可以达到训练时验证集的检测精度。

- **闭集模型推理：**

该模式下和常规目标检测推理方式一致，用户仅需上传待检测的图片即可直接进行目标检测推理。该模式下不具备开集检测能力只能检测训练数据集中的类目，模型会自动加载闭集训练时得到的类目提示特征，无需用户主动进行提示特征获取。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

表 3-66 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是 (云上部署 模型填写)	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5 中响应消息头中X-Subject- Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-67 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig- AppCode	是 (云上部署 模型填写、边 缘部署模型仅 在监控网关访 问模式下填 写)	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig- AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-68 请求 Body 参数 --开集推理-提示特征获取

参数	是否必选	参数类型	说明
mode	是, 值为 prompt_feature_gen	String	参数解释: 推理模式为“开集检测特征生成”。 约束限制: 不涉及 取值范围: 提示特征获取模式, 值为固定值 “prompt_feature_gen” 默认取值: 无
confidence	否, 默认值0.5	String	参数解释: 设置目标检测结果的阈值过滤, 目标检测结果score低于该值的结果会被过滤, 选填, 如果不填则使用默认值0.5。 约束限制: 不涉及 取值范围: (0, 1) 默认取值: 0.5

参数	是否必选	参数类型	说明
text_prompt	否, 默认值 None	String	<p>参数解释: 文本提示 ('.'分隔类别, 例如 'cat.dog.'表示提示 cat和dog两个类别), 仅适用于 prompt_feature_gen模式, 该模式下该参数为选填。</p> <p>说明 建议提示类别不超过50个。</p> <p>约束限制: 不涉及</p> <p>取值范围: 无</p> <p>默认取值: None</p>
visual_prompt	否, 默认值 None	String	<p>参数解释: 视觉提示, list类型, 其中每张提示图像是一个dict类型, 包含两个key: img (image的base64编码) 和prompt (该图像的视觉提示)。prompt是一个list类型, 包含2个两个key: label (目标类目名) 和bbox (标注框信息, 其格式为[x,y,w,h])。建议提示图像最多10个。</p> <p>约束限制: 不涉及</p> <p>取值范围: 无</p> <p>默认取值: None</p>
target_image	是	String	<p>参数解释: 待检测图像的base64编码, 必填。</p> <p>约束限制: 不涉及</p> <p>取值范围: 无</p> <p>默认取值: 无</p>

参数	是否必选	参数类型	说明
show_detecte d_image	否, 默认值 false	String	参数解释: 是否返回目标图像叠加预测信息可视化结果, true返回, false不返回。 约束限制: 不涉及 取值范围: true,false 默认取值: false

表4 请求Body参数 --开集推理-提示特征推理

参数	是否必选	参数类型	说明
mode	是, 值为 prompt_featur e_inference	String	参数解释: 推理模式为“开集检测-基于提示特征推理”。 约束限制: 不涉及 取值范围: 基于提示特征推理, 值为固定值 “prompt_feature_inference” 默认取值: 无
confidence	否,默认值0.5	String	参数解释: 设置目标检测结果的阈值过滤, 目标检测结果score低于该值的结果会被过滤, 选填, 如果不填则使用默认值0.5。 约束限制: 不涉及 取值范围: (0, 1) 默认取值: 0.5

参数	是否必选	参数类型	说明
prompt_feature	是, 默认值 None	dict	参数解释: 提示特征编码, dict类型, key是类别名称, val是该类的特征base64编码, prompt_feature_gen模式下的返回结果可以得到, 仅适用于prompt_feature_inference模式。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
target_image	是	String	参数解释: 待检测图像的base64编码, 必填。 约束限制: 不涉及 取值范围: 无 默认取值: 无

表5 请求Body参数 --闭集推理

参数	是否必选	参数类型	说明
confidence	否,各个类别最佳阈值	String	参数解释: 用于目标检测结果的阈值过滤, 目标检测结果score低于该值的结果会被过滤, 选填, 如果用户不传此参数, 则模型会使用基于验证集上学习得到的最佳类别阈值。 约束限制: 不涉及 取值范围: (0, 1) 默认取值: 无

参数	是否必选	参数类型	说明
target_image	是	String	参数解释: 待检测图像的base64编码，必填。 约束限制: 不涉及 取值范围: 无 默认取值: 无

响应参数

表6 响应Body--开集推理-提示特征获取

参数	参数类型	说明
results	List	参数解释: 物体检测的识别结果。格式为[{'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}, {'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}...]。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
dataset_id	String	参数解释: 数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	说明
prompt_feature	dict	参数解释: 返回值为dict，其中key为预测类别，value为类别特征base64编码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表7 响应Body--开集推理-提示特征推理

参数	参数类型	说明
results	List	参数解释: 物体检测的识别结果。格式为[{'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}, {'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}...]。 label: 预测类别。 score: 置信度。 bbox: 目标主体信息，分别为[x1,y1,w,h]，其中x1,y1代表左上角坐标，w是宽度，h是高度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
dataset_id	String	参数解释: 数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表8 请求Body参数 --闭集推理

参数	参数类型	说明
results	List	<p>参数解释: 物体检测的识别结果。格式为[{'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}, {'label': 'xxx', 'score': xxx, 'bbox': [x,y,w,h]}...]]。 label: 预测类别。 score: 置信度。 bbox: 目标主体信息,分别为[x1,y1,w,h]，其中x1,y1代表左上角坐标，w是宽度，h是高度。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
dataset_id	String	<p>参数解释: 数据集ID。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

请求示例

1) 开集推理-提示特征获取

```
{  
    "mode": "prompt_feature_gen",  
    "confidence": 0.5,  
    "target_image": "target_image_base64_str",  
    "text_prompt": "cat. dog",  
    "visual_prompt": [  
        {  
            "img": "visual_image_1_base64_str",  
            "prompt": [  
                {  
                    "label": "cat",  
                    "bbox": [  
                        100,  
                        100,  
                        50,  
                        50  
                    ]  
                },  
                {  
                    "label": "cat",  
                    "bbox": [  
                        150,
```

```
        150,
        50,
        50
    ],
}
},
{
    "img": "visual_image_2_base64_str",
    "prompt": [
        {
            "label": "dog",
            "bbox": [
                30,
                30,
                60,
                60
            ]
        },
        {
            "label": "dog",
            "bbox": [
                80,
                80,
                60,
                50
            ]
        }
    ],
    "show_detected_image": true
}
```

2) 开集推理-提示特征推理

```
{
    "mode": "prompt_feature_inference",
    "confidence": 0.5,
    "target_image": "target_image_base64_str",
    "prompt_feature": {
        "cat": "cat_feature_base64_str",
        "dog": "dog_feature_base64_str"
    }
}
```

3) 闭集推理

```
{
    "confidence": 0.5,
    "target_image": "target_image_base64_str"
}
```

响应示例

1) 开集推理-提示特征获取

```
{
    "dataset_id":"",
    "results": [
        {
            "label": "cat",
            "bbox": {
                "X":10,
                "Y":10,
                "Width":50,
                "Height":50,
                "Angle":0
            },

```

```
        "score": 0.6784
    },
],
"result_vis": "target_image_result_vis_base64_str",
"prompt_feature": {
    "cat": "cat_feature_base64_str",
    "dog": "dog_feature_base64_str"
}
}
```

2) 开集推理-提示特征推理

```
{
    "dataset_id":"",
    "results": [
        {
            "label": "cat",
            "bbox": {
                "X":10,
                "Y":10,
                "Width":50,
                "Height":50,
                "Angle":0
            },
            "score": 0.6784
        }
    ],
}
```

3) 闭集推理

```
{
    "dataset_id":"",
    "results": [
        {
            "label": "cat",
            "bbox": {
                "X":10,
                "Y":10,
                "Width":50,
                "Height":50,
                "Angle":0
            },
            "score": 0.6784
        }
    ],
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3 预测大模型

3.1.3.1 盘古统一编码时序预测分类大模型

功能介绍

基于时序预测基模型实现分类预测能力。时序分类预测有很多应用场景，例如：基于工业设备传感器一段时间采集的连续数据，实现设备正常或异常状态的预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-69](#)。

表 3-69 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-70](#)。

表 3-70 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-71 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	<p>参数解释: 时序数据请求的批处理队列。</p> <p>约束限制: 盘古时序预测分类任务的输入数据内容，用字典承载。输入数据字典中，键对应特征列名，值为序列数据具体内容。</p> <p>取值范围: 输入数据中的一行数据。</p> <p>默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。</p>
context_len	否	int	<p>参数解释: 时序任务的历史窗口大小，该参数主要服务于支持多尺度窗口的模型的推理。</p> <p>约束限制: 填写时需要保证输入序列长度大于等于该值。</p> <p>取值范围:</p> <ul style="list-style-type: none">• min: 32。• max: 与训练模型的历史窗口大小一致，最大不超过 512。 <p>默认取值: -1， -1为当前服务模型的默认大小。</p>

响应参数

状态码： 200

表 3-72 响应 Body 参数

参数	参数类型	描述
data	LIST<Output Meta>	参数解释: 时序预测结果的列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
time_cost	JSON	参数解释: 当启动服务时，本次请求服务各阶段耗时情况。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-73 OutputMeta 格式说明

参数名称	参数类型	说明
prediction	JSON	参数解释: 时序分类的输出结果，具体见样例调用示例。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-74 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。

参数	参数类型	描述
error_msg	String	错误信息。

请求示例

```
{  
    "context_len": 256,  
    "data": [  
        {  
            "context": {  
                "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
            }  
        },  
        {  
            "context": {  
                "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
            }  
        }  
    ]  
}
```

响应示例

```
{  
    "data": [  
        {  
            "prediction":{  
                "label": 0  
            }  
        }  
    ],  
    "time_cost": {  
        "infer_cost_time": "82.609 ms",  
        "postprocess_cost_time": "0.29 ms",  
        "preprocess_cost_time": "0.1332 ms",  
        "service_cost_total_time": "84.965 ms"  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.2 盘古统一编码时序预测回归大模型

功能介绍

2025年3月首次发布的模型，基于时序预测基模型实现对未来一组观测值的预测，最多支持256个未来时间点的预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-75](#)。

表 3-75 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于 获取操作API的权限。如图 2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-76](#)。

表 3-76 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-77 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	<p>参数解释: 时序数据请求的批处理队列。</p> <p>约束限制: 盘古时序预测回归任务的输入数据内容，用字典承载。输入数据字典中，键对应特征列名，值为序列数据具体内容。</p> <p>取值范围: 输入数据中的一组数据。</p> <p>默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。</p>
context_len	否	int	<p>参数解释: 时序任务的历史窗口大小，该参数主要服务于支持多尺度窗口的模型的推理。</p> <p>约束限制: 填写时需要保证输入序列长度大于等于该值。</p> <p>取值范围:</p> <ul style="list-style-type: none">• min: 32• max: 与训练模型的历史窗口大小一致，最大不超过 512。 <p>默认取值: -1， -1为当前服务模型的默认大小。</p>

参数	是否必选	参数类型	描述
prediction_len	否	int	<p>参数解释: 时序预测输出的窗口长度。数值越大输出预测的时间范围越大，但相应的显存占用会增加以及模型预测精度可能会下降。</p> <p>约束限制: 基于已发布模型的进行增量微调，输出窗口需要与其保持一致。</p> <p>取值范围:</p> <ul style="list-style-type: none">• min: 1• max: 与训练模型的预测目标窗口大小一致，最大不超过128。 <p>默认取值: 与训练模型的预测目标窗口大小一致。</p>

响应参数

状态码: 200

表 3-78 响应 Body 参数

参数	参数类型	描述
data	LIST<Output Meta>	<p>参数解释: 时序预测结果的列表。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
time_cost	JSON	参数解释: 当启动服务时，本次请求服务各阶段耗时情况。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-79 OutputMeta 格式说明

参数名称	参数类型	说明
prediction	JSON	参数解释: 时序分类的输出结果，具体见样例调用示例。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-80 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "context_len": 256,  
    "prediction_len": 96,  
    "data": [  
        {  
            "context": {  
                "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
                "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
            }  
        }  
    ]  
}
```

```
        },
        {
            "context": {
                "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
                "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]
            }
        }
    }
```

响应示例

```
{
    "data": [
        {
            "prediction": {
                "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
                "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]
            }
        }
    ],
    "time_cost": {
        "infer_cost_time": "82.609 ms",
        "postprocess_cost_time": "0.29 ms",
        "preprocess_cost_time": "0.1332 ms",
        "service_cost_total_time": "84.965 ms"
    }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.3 盘古统一编码表格预测分类大模型

功能介绍

基于统一编码大模型实现表格分类预测能力，面向具体业务场景的分类任务，用户传入分类数据，使用模型对指定的预测目标进行分类预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-81](#)。

表 3-81 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-82](#)。

表 3-82 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-83 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 待进行预测的数据，data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 约束限制: 待预测特征名需要与训练数据中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。 取值范围: 输入数据中的一组数据。 默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。



响应参数

状态码： 200

表 3-84 响应 Body 参数

参数	参数类型	描述
results	LIST	参数解释： 分类预测结果的列表。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
time_cost	JSON	参数解释： 当启动服务时，本次请求服务各阶段耗时情况。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

状态码： 400

表 3-85 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
  "data": [  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    },  
    ...  
    {  
      "feature_1": xx,
```

```
    "feature_2": xx,  
    ...  
    "feature_n": xx  
  }  
]  
}
```

响应示例

```
{  
  "results": [  
    0,  
    1,  
    ...  
    0  
  ],  
  "time_cost": {  
    "infer_cost_time": 233,  
    "postprocess_cost_time": 13,  
    "preprocess_cost_time": 12,  
    "service_cost_total_time": 258  
  }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.4 盘古统一编码表格预测回归大模型

功能介绍

基于统一编码大模型实现表格回归预测能力，针对特定场景的回归任务，用户传入回归数据，使用模型对指定的预测目标进行回归预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-86](#)。

表 3-86 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-87](#)。

表 3-87 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-88 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 待进行预测的数据，data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 约束限制: data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 待预测特征名需要与训练数据中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。 取值范围: 输入数据中的一组数据。 默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。 

响应参数

状态码： 200

表 3-89 响应 Body 参数

参数	参数类型	描述
data	LIST	参数解释： 回归预测结果的列表。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
time_cost	JSON	参数解释： 当启动服务时，本次请求服务各阶段耗时情况。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

状态码： 400

表 3-90 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
  "data": [  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    },  
    ...  
    {  
      "feature_1": xx,
```

```
        "feature_2": xx,  
        ...  
        "feature_n": xx  
    }  
]  
}
```

响应示例

```
{  
    "results": [  
        9.89097,  
        2.9097  
        ...  
        21.06979  
    ],  
    "time_cost": {  
        "infer_cost_time": 233,  
        "postprocess_cost_time": 13,  
        "preprocess_cost_time": 12,  
        "service_cost_total_time": 258  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.5 盘古时序异常检测大模型

3.1.3.5.1 时序异常检测预训练工作流

功能介绍

盘古时序异常检测大模型，支持时序数据的异常检测任务，分为预训练产物以及微调产物两种不同的部署模式。

1. 预训练产物（正常数据学习）：使用设备历史正常数据对时序异常检测大模型进行全量训练，使其更好地学习并重构正常数据的潜在特征分布。在推理阶段，利用模型对输入样本的重构误差作为异常判别依据：若重构效果显著下降，则表明该样本偏离正常模式，具有潜在异常趋势，从而实现对异常数据的有效检测与区分。
2. 微调产物（正异常数据分类）：在时序异常检测大模型基础上引入分类头，并在训练过程中冻结模型的主干网络，仅使用历史中包含正常与异常样本的混合数据对新增分类头进行微调。通过此方式，将原始问题转化为针对特定时间点的二分类任务，判断其状态为正常或异常。

推理部署服务需要兼容预训练、微调两种产物的部署，每种模型调用逻辑以及输出结果有区别，需要服务内自动识别加载模型的类别，并根据API请求体输入返回推理结果。

1. 若为预训练工作流产物（正常数据学习）模型，根据指定异常分计算方式返回异常分+重构值，共有以下三种计算方式：

- a. 重构误差：计算多次重构的均值，计算与真实值的偏差值记作异常分。
 - b. 重构方差（推荐）：计算多次重构的方差，记作异常分。
 - c. 重构误差+方差：异常分计算方式a)+b)，记作异常分。
2. 若为微调工作流产物（正异常数据分类）模型，返回时序点正异常类别+重构值。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-91](#)。

表 3-91 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释： 用户Token。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
Content-Type	是	String	参数解释： 发送的实体的MIME类型。 约束限制： 不涉及 取值范围： 不涉及 默认取值： application/json

使用[API Key认证](#)方式的请求Header参数见[表3-92](#)。

表 3-92 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-93 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	<p>参数解释: 时序异常检测批处理数据。</p> <p>约束限制: 盘古时序预测分类任务的输入数据内容，用字典承载。输入数据字典中，键对应特征列名，值为列表承载的序列数据具体内容，单条时序数据窗口长度需大于等于100。</p> <p>取值范围: 输入数据中的一组数据。</p> <p>默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。</p> 

参数	是否必选	参数类型	描述
option	否	String	<p>参数解释: 选择异常分计算模式，可选择模式包括["err", "var", "mix"]，分别代表["重构误差", "重构方差", "重构误差+方差"]三种计算模式，推荐使用重构方差("var")：</p> <ol style="list-style-type: none">1. 重构误差("err")：计算多次重构的均值，计算与真实值的偏差值记作异常分。2. 重构方差("var")：计算多次重构的方差，记作异常分。3. 重构误差+方差("mix")：异常分计算方式a)+b)，记作异常分。 <p>约束限制: 无</p> <p>取值范围: {"err", "var", "mix"}</p> <p>默认取值: var</p>

响应参数

状态码： 200

表 3-94 响应 Body 参数

参数	参数类型	描述
data	LIST<Output Meta>	<p>参数解释: 时序预测结果的列表。同时返回各个特征的异常分以及重构值，分别以"anomaly_score"以及"reconstruction"两个键值承载。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
time_cost	JSON	参数解释: 当启动服务时，本次请求服务各阶段耗时情况。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-95 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
  "option": "var", #可选模式包括["err", "var", "mix"]  
  "data": [  
    {  
      "context": {  
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
      }  
    },  
    {  
      "context": {  
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
      }  
    },  
    {...},  
    {...}  
  ]  
}
```

响应示例

```
{  
  "data": [  
    {  
      # 异常分  
      "anomaly_score":{  
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
      }  
    }  
  ]  
}
```

```
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]
    },
    # 重构结果
    "reconstruction":{
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]
    }
},
{...},
{...},
{...}
],
"time_cost": {
    "infer_cost_time": "82.609 ms",
    "postprocess_cost_time": "0.29 ms",
    "preprocess_cost_time": "0.1332 ms",
    "service_cost_total_time": "84.965 ms"
}
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.5.2 时序异常检测微调工作流

功能介绍

盘古时序异常检测大模型，支持时序数据的异常检测任务，分为预训练产物以及微调产物两种不同的部署模式。

1. 预训练产物（正常数据学习）：使用设备历史正常数据对时序异常检测大模型进行全量训练，使其更好地学习并重构正常数据的潜在特征分布。在推理阶段，利用模型对输入样本的重构误差作为异常判别依据：若重构效果显著下降，则表明该样本偏离正常模式，具有潜在异常趋势，从而实现对异常数据的有效检测与区分。
2. 微调产物（正异常数据分类）：在时序异常检测大模型基础上引入分类头，并在训练过程中冻结模型的主干网络，仅使用历史中包含正常与异常样本的混合数据对新增分类头进行微调。通过此方式，将原始问题转化为针对特定时间点的二分类任务，判断其状态为正常或异常。

推理部署服务需要兼容预训练、微调两种产物的部署，每种模型调用逻辑以及输出结果有区别，需要服务内自动识别加载模型的类别，并根据API请求体输入返回推理结果。

1. 若为预训练工作流产物（正常数据学习）模型，根据指定异常分计算方式返回异常分+重构值，共有以下三种计算方式：
 - a. 重构误差：计算多次重构的均值，计算与真实值的偏差值记作异常分。
 - b. 重构方差（推荐）：计算多次重构的方差，记作异常分。
 - c. 重构误差+方差：异常分计算方式a)+b)，记作异常分。

2. 若为微调工作流产物（正异常数据分类）模型，返回时序点正异常类别+重构值。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-96](#)。

表 3-96 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-97](#)。

表 3-97 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-98 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	<p>参数解释: 时序异常检测批处理数据。</p> <p>约束限制: 盘古时序预测分类任务的输入数据内容，用字典承载。输入数据字典中，键对应特征列名，值为列表承载的序列数据具体内容，单条时序数据窗口长度需大于等于100。</p> <p>取值范围: 输入数据中的一组数据。</p> <p>默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。</p> 

响应参数

状态码： 200

表 3-99 响应 Body 参数

参数	参数类型	描述
data	LIST<Output Meta>	<p>参数解释: 时序预测结果的列表。同时返回各个时间点的正异常类别以及各个特征的重构值，分别以"label"以及"reconstruction"两个键值承载。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
time_cost	JSON	参数解释: 当启动服务时，本次请求服务各阶段耗时情况。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-100 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
  "data": [  
    {  
      "context": {  
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
      }  
    },  
    {  
      "context": {  
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],  
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]  
      }  
    },  
    {...},  
    {...}  
  ]  
}
```

响应示例

```
{  
  "data": [  
    {  
      # 正异常类别  
      "label": [ 0, 1, 0, 0, ...],  
      # 重构结果  
      "reconstruction": [  
        ...  
      ]  
    }  
  ]  
}
```

```
        "feature_0": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_1": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_2": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...],
        "feature_3": [ 864.1618041992188, 493.91827392578125, 252.7724151611328, ...]
    },
},
{...},
{...},
{...}
],
"time_cost": {
    "infer_cost_time": "82.609 ms",
    "postprocess_cost_time": "0.29 ms",
    "preprocess_cost_time": "0.1332 ms",
    "service_cost_total_time": "84.965 ms"
}
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.6 盘古融合推荐异常检测大模型

功能介绍

针对特定场景的异常检测任务，用户传入异常检测数据，使用模型对指定的预测目标进行异常检测预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-101](#)。

表 3-101 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-102](#)。

表 3-102 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-103 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 待进行预测的数据，从输入的特征数据中检测是否存在异常值。 约束限制: data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 取值范围: 待预测特征名需要与训练数据中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。 默认取值: 输入数据中的一组数据。  预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。

参数	是否必选	参数类型	描述
predict_proba	否	boolean	参数解释: 是否输出置信度。 约束限制: 无 取值范围: <ul style="list-style-type: none">• true: 输出• false: 不输出 默认取值: false。

响应参数

状态码: 200

表 3-104 响应 Body 参数

参数	参数类型	描述
result	Object	参数解释: 预测结果信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
pred_proba	Array	参数解释: 预测结果对应的置信度，每条结果为一个Dict，Dict的键值对分别表示预测结果、置信度。 约束限制: 请求字段predict_proba配置为true时，响应body才返回pred_proba子项，否则无该子项。 取值范围: 不涉及 默认取值: 不涉及

状态码: 400

表 3-105 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "data": [  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx  
        },  
        ...  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx  
        }  
    ],  
    "predict_proba": true  
}
```

响应示例

```
{  
    "result": [  
        "0",  
        "1",  
        ...  
        "0"  
    ],  
    "pred_proba": [  
        {  
            "0": 0.791,  
            "1": 0.209  
        },  
        {  
            "0": 0.103,  
            "1": 0.897  
        },  
        ...  
        {  
            "0": 0.665,  
            "1": 0.335  
        }  
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.7 盘古融合推荐分类大模型

此章节内容适用于盘古融合推荐分类大模型和盘古融合推荐分类大模型-极简模式。

3.1.3.7.1 数据分类预测

功能介绍

针对特定场景的分类任务，用户传入分类数据，使用模型对指定的预测目标进行分类预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-106](#)。

表 3-106 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-107](#)。

表 3-107 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-108 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	<p>参数解释: 待进行预测的数据，data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。</p> <p>约束限制: 待预测特征名需要与训练数据中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。</p> <p>取值范围: 输入数据中的一组数据。</p> <p>默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。</p> 
predict_proba	否	boolean	<p>参数解释: 是否输出置信度。</p> <p>约束限制: 无</p> <p>取值范围:</p> <ul style="list-style-type: none">• true: 输出• false: 不输出 <p>默认取值: false。</p>

响应参数

状态码： 200

表 3-109 响应 Body 参数

参数	参数类型	描述
result	Object	参数解释: 预测结果信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
pred_proba	Array	参数解释: 预测结果对应的置信度，每条结果为一个Dict，Dict的键值对分别表示预测结果、置信度。 约束限制: 请求字段predict_proba配置为true时，响应body才返回pred_proba子项，否则无该子项。 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-110 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息

请求示例

```
{  
  "data": [  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    },  
    ...  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    }  
  ]  
}
```

```
        },
    ],
    "predict_proba": true
}
```

响应示例

```
{
  "result": [
    "0",
    "1",
    ...
    "0"
  ],
  "pred_proba": [
    {
      "0": 0.791,
      "1": 0.209
    },
    {
      "0": 0.103,
      "1": 0.897
    },
    ...
    {
      "0": 0.665,
      "1": 0.335
    }
  ]
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.7.2 排列特征重要性

功能介绍

支持用户传入一批数据样本，并据此计算各个特征的排列重要性（permutation importance）。

基本原理为，对输入的数据样本，首先完成一次推理，将评估指标作为baseline。将数据的某一特征在行数据之间随机混淆，评估预测效果，并计算指标的下降值，并以此评估特征的重要性。当混淆某一列特征，导致预测效果与baseline有较大下降时，则说明此列特征重要性较高。此处使用accuracy作为分类任务的评估指标。

注意：输入的数据需包含预测目标列的真值，且数量大于1000，以便更精确地计算预测效果指标。

另外，融合推荐分类预测大模型提供两种特征重要性，分别为模型权重特征重要性和排列特征重要性，对两种特征重要性的详细说明参见《用户指南》“使用API调用预测大模型 > 排列特征重要性”部分。

注意：极简模式不提供排列特征重要性接口，即极简模式的模型不可调用该接口

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /perm-feat-importance

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-111](#)。

表 3-111 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于 获取操作API的权限。如图 2-5 中响应消息头中X-Subject- Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-112](#)。

表 3-112 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-113 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 含义和 表3-108 中的data参数相同。 约束限制: 排列特征重要性接口除提供各特征名及特征值外，还需提供预测目标列的列名及其真值（即示例中的target项），且数据数量大于1000，以便准确计算预测效果指标。 取值范围: 输入数据中的一组数据。 默认取值: 不涉及

参数	是否必选	参数类型	描述
fi_args	是	Dict	<p>参数解释: 特征重要性的参数，包含 num_shuffle_sets、sample_num 字段。</p> <p>约束限制: num_shuffle_sets，非必填，Integer 类型，代表采样混洗次数，越大则结果越准确，但耗时更高。 sample_num，非必填，Integer 类型，代表单次采样混洗数据量，越大则结果越准确，但耗时更高。</p> <p>取值范围:</p> <ul style="list-style-type: none">• num_shuffle_sets，要求 ≥ 1，建议 ≥ 5• sample_num，要求 ≥ 1000，建议 ≥ 3500 <p>默认取值: num_shuffle_sets，默认值为 1。 sample_num，默认值为 1000，当用户不传入此参数，或大于 data 数量时，直接使用全部数据，不进行采样。</p>

响应参数

状态码： 200

表 3-114 响应 Body 参数

参数	参数类型	描述
feature_importance	Array	<p>参数解释: 降序排列的各特征重要性，每项为 [特征名, 特征重要性分值]。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
cost_time	Long	参数解释: 计算耗时。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-115 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "data": [  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx,  
            "target": 0,  
        },  
        ...  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx,  
            "target": 1  
        }  
    ],  
    "fi_args": {"num_shuffle_sets": 1,  
                "sample_num": 1000  
    }  
}
```

此处data列表的长度至少为1000，建议通过以下代码根据csv文件生成请求体
(DATA_NUM为从csv文件中采样数据的数量)：

```
import random  
import json  
import pandas as pd  
  
ENCODING_TYPE = "utf-8"  
DATA_NUM = 1200
```

```
data_path = "path/to/xxx.csv"
df = pd.read_csv(data_path, encoding=ENCODING_TYPE)
data_list = df.to_dict("records")
data_list = random.sample(data_list, DATA_NUM)
request_data = {
    "data": data_list,
    "fi_args": {
        "num_shuffle_sets": 1,
        "sample_num": 1000,
    }
}
request_str = json.dumps(request_data)
print(request_str)
```

响应示例

```
{
    "cost_time": 4.322,
    "feature_importance": [
        ["feature_2", 0.15]
        ...
        ["feature_17", 0.001]
    ]
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.8 盘古融合推荐回归大模型

此章节内容适用于盘古融合推荐回归大模型和盘古融合推荐回归大模型-极简模式。

3.1.3.8.1 数据回归预测

功能介绍

针对特定场景的回归任务，用户传入回归数据，使用模型对指定的预测目标进行回归预测。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-116](#)。

表 3-116 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-117](#)。

表 3-117 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-118 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 待进行预测的数据，data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 约束限制: data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 待预测特征名需要与训练数据中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。 取值范围: 输入数据中的一组数据。 默认取值: 预测大模型在训练完成后，可以在训练日志页面，“模型训练”日志节点中获取推理api所用的示例数据。填写请求Body时可以参考该示例填写。 

响应参数

状态码： 200

表 3-119 响应 Body 参数

参数	参数类型	描述
result	Object	参数解释： 预测结果信息。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

状态码： 400

表 3-120 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
  "data": [  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    },  
    ...  
    {  
      "feature_1": xx,  
      "feature_2": xx,  
      ...  
      "feature_n": xx  
    }  
  ]  
}
```

响应示例

单目标响应

```
{  
  "result": [  
    "prediction_1",  
    ...  
  ]  
}
```

```
        "prediction_n"
    ]
}
多目标响应
{
    "result": [
        [
            "prediction_1",
            ...
            "prediction_n"
        ]
    ]
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.8.2 排列特征重要性

功能介绍

支持用户传入一批数据样本，并据此计算各个特征的排列重要性（permutation importance）。

基本原理为，对输入的数据样本，首先完成一次推理，将评估指标作为baseline。将数据的某一特征在行数据之间随机混淆，评估预测效果，并计算指标的下降值，并以此评估特征的重要性。当混淆某一列特征，导致预测效果与baseline有较大下降时，则说明此列特征重要性较高。此处使用r2_score作为回归任务的评估指标。

注意：输入的数据需包含预测目标列的真值，且数量大于1000，以便更精确地计算预测效果指标。

另外，融合推荐回归预测大模型提供两种特征重要性，分别为模型权重特征重要性和排列特征重要性，对两种特征重要性的详细说明参见《用户指南》“使用API调用预测大模型 > 排列特征重要性”部分。

注意：极简模式不提供排列特征重要性接口，即极简模式的模型不可调用该接口。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /perm-feat-importance

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-121](#)。

表 3-121 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-122](#)。

表 3-122 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-123 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 含义和 表3-118 中的data参数相同。 约束限制: 排列特征重要性接口除提供各特征名及特征值外，还需提供预测目标列的列名及其真值（即示例中的target项），且数据数量大于1000，以便准确计算预测效果指标。 取值范围: 输入数据中的一组数据。 默认取值: 不涉及

参数	是否必选	参数类型	描述
fi_args	是	Dict	<p>参数解释: 特征重要性的参数，包含 num_shuffle_sets、sample_num 字段。</p> <p>约束限制: num_shuffle_sets，非必填，Integer 类型，代表采样混洗次数，越大则结果越准确，但耗时更高。 sample_num，非必填，Integer 类型，代表单次采样混洗数据量，越大则结果越准确，但耗时更高。</p> <p>取值范围:</p> <ul style="list-style-type: none">• num_shuffle_sets，要求 ≥ 1，建议 ≥ 5。• sample_num，要求 ≥ 1000，建议 ≥ 3500。 <p>默认取值: num_shuffle_sets，默认值为 1。 sample_num，默认值为 1000，当用户不传入此参数，或大于 data 数量时，直接使用全部数据，不进行采样。</p>

响应参数

状态码： 200

表 3-124 响应 Body 参数

参数	参数类型	描述
feature_importance	Array	<p>参数解释: 降序排列的各特征重要性，每项为 [特征名, 特征重要性分值]。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
cost_time	Long	参数解释: 计算耗时。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-125 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "data": [  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx,  
            "target": 0,  
        },  
        ...  
        {  
            "feature_1": xx,  
            "feature_2": xx,  
            ...  
            "feature_n": xx,  
            "target": 1  
        }  
    ],  
    "fi_args":{  
        "num_shuffle_sets": 1,  
        "sample_num": 1000  
    }  
}
```

此处data列表的长度至少为1000，建议通过以下代码根据csv文件生成请求体
(DATA_NUM为从csv文件中采样数据的数量)：

```
import random  
import json  
import pandas as pd  
  
ENCODING_TYPE = "utf-8"
```

```
DATA_NUM = 1200

data_path = "path/to/xxx.csv"
df = pd.read_csv(data_path, encoding=ENCODING_TYPE)
data_list = df.to_dict("records")
data_list = random.sample(data_list, DATA_NUM)
request_data = {
    "data": data_list,
    "fi_args": {
        "num_shuffle_sets": 1,
        "sample_num": 1000,
    }
}
request_str = json.dumps(request_data)
print(request_str)
```

响应示例

```
{
    "cost_time": 4.322,
    "feature_importance": [
        ["feature_2", 0.15]
        ...
        ["feature_17", 0.001]
    ]
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.3.9 盘古表格直推预测大模型

功能介绍

可实现免微调直推表格预测，面对有限样本条件下，基于上下文关联推理技术，免微调自适应数据分布，可以实现开箱即用的效果。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /

获取URI方式请参见[请求URI](#)。

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-126](#)。

表 3-126 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-127](#)。

表 3-127 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-128 请求 Body 参数

参数	是否必选	参数类型	描述
target	是	string	参数解释: 指定要预测的目标列名。 约束限制: 仅支持单个目标。 取值范围: 不涉及 默认取值: 不涉及
sample	是	json	参数解释: 预测所需的样例数据 (few shot) , sample是json格式， json的key对应特征列名， value是一个Array， 对应该特征列的数据。格式详见请求示例。 约束限制: sample行数 <= 10000行 特征数 <= 500 sample * 特征数 <= 1000000 云上部署数据量 < 12MB

参数	是否必选	参数类型	描述
data	是	Array	参数解释: 待进行预测的数据，data为一个数组，数组中包含多个对象，每个对象是一组完整的特征数据。 约束限制: 待预测特征名需要与sample中的特征名保持一致。例如，训练数据中特征列按照feature_1、feature_2……进行命名，在调用推理接口时，特征名也需要保持相同。同时推理接口中特征数量需要与训练数据中的特征数保持一致。一组特征数据填写完成后再填写剩余待预测数据，格式详见请求示例。 data行数 <= 1000行 特征数 <= 500
task_type	是	string	参数解释: 任务类型。 约束限制: 支持的类型包括regression和classification，分别对应回归和分类任务。

响应参数

状态码： 200

表 3-129 响应 Body 参数

参数	参数类型	描述
results	LIST	参数解释: 预测结果的列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
time_cost	JSON	参数解释: 当启动服务时，本次请求服务各阶段耗时情况。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-130 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

```
{  
    "target": "feature_0",  
    "task_type": "regression",  
    "sample": {  
        "feature_0": [  
            9,  
            0  
        ],  
        "feature_1": [  
            252.71994790753982,  
            0  
        ],  
        "feature_2": [  
            584.8740678212971,  
            0  
        ],  
        "feature_3": [  
            507.5732435552864,  
            0  
        ],  
        "feature_4": [  
            75.7524144068944,  
            0  
        ]  
    },  
    "data": [  
        {  
            "feature_1": 6.4533336224,  
            "feature_2": 677.3766556825,  
            "feature_3": 923.0984447151,  
            "feature_4": 837.6293168679  
        },  
        {  
            "feature_1": 6.4533336224,  
            "feature_2": 677.3766556825,  
            "feature_3": 923.0984447151,  
            "feature_4": 837.6293168679  
        }  
    ]  
}
```

```
        "feature_1": 712.3825532408,
        "feature_2": 373.0545721447,
        "feature_3": 393.8145421256,
        "feature_4": 597.3197447678
    }
}
}
```

响应示例

```
{
  "results": [
    0,
    1,
    ...
    0
  ],
  "time_cost": {
    "infer_cost_time": 233,
    "postprocess_cost_time": 13,
    "preprocess_cost_time": 12,
    "service_cost_total_time": 258
  }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.4 科学计算大模型

3.1.4.1 气象类模型

3.1.4.1.1 创建推理作业

功能介绍

支持调用科学计算大模型创建气象/降水模型的推理作业。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/infer-api/proxy/service/{deployment_id}/v2/{project_id}/tasks

获取URI方式请参见[请求URI](#)。

表 3-131 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-132](#)。

表 3-132 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-133](#)。

表 3-133 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-134 请求 body 参数

参数	是否必选	参数类型	描述
models	否	Array	<p>参数解释: 选择用来推理的模型子集列表。可通过配置model_tag的方式，指定参与推理的模型子集。例如：["global-weather"]，如不传则默认使用全部模型进行推理。model_tag与模型的对应关系请见表9 模型资产与model tag对应关系。</p> <p>约束限制: 字符串列表，指定的model tag需要在已部署的模型范围内。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
data_dirs	是	Map<string, string>	<p>参数解释: 根据不同模型指定输入数据的路径。格式为{model_tag: shema://dataset_path}，例如{"global-weather": "obs://path/to/data/"}, model_tag与模型的对应关系请见表9 模型资产与model tag对应关系。</p> <p>约束限制: shema参数在线服务仅支持obs，边缘服务支持file（本地路径）。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
output_dir	是	String	<p>参数解释: 推理结果输出路径。格式为shema://dataset_path，例如file:///path/to/data/。</p> <p>约束限制: shema参数在线服务仅支持obs，边缘服务支持file（本地路径）。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
start_times	是	Array	<p>参数解释: 起报时间点列表，若不传入start_time_interval，则列表中的每个时间均作为起报时，若传入start_time_interval，则start_times必须只包含两个时间，分别作为起报时间的起点和终点，根据start_time_interval最终插值出所有起报时间点（总起报时间点数量不能超过列表长度上限）。</p> <p>约束限制: 起报时间点字符串格式为YYYYMMDDHH时间戳。</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
start_time_interval	否	String	<p>参数解释: 起报时间间隔时长。仅数字代表小时数，支持数字+单位表示，例如1d, 3h。</p> <p>约束限制: 不涉及</p> <p>取值范围: (0h, 24h]</p> <p>默认取值: 6h</p>
forecast_lead_time	否	String	<p>参数解释: 预报时效。仅数字代表小时数，支持数字+单位表示，例如6d</p> <p>约束限制: 不涉及</p> <p>取值范围: [1h, 720h]</p> <p>默认取值: 168h</p>

参数	是否必选	参数类型	描述
draw_figures	否	Bool	参数解释: 是否输出结果图片。 约束限制: 不涉及 取值范围: true/false 默认取值: true
clipping_geo_ranges	否	Clip ping Geo Ran ges obje ct	参数解释: 指定对结果文件剪裁的地理范围，在结果文件中仅会输出指定范围内的区域。如不指定，则默认输出原结果范围。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
ensemble	否	Ense mbl eDt o obje ct	参数解释: 集合预报信息 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-135 ClippingGeoRanges

参数	是否必选	参数类型	描述
latitude	否	Array	<p>参数解释: 指定对结果文件进行剪裁的纬度范围，长度为2的定长数组，左端小于右端，例如：[45, 70]。如指定longitude，不指定latitude，默认补全为全球范围[-90, 90]。</p> <p>约束限制: 长度为2的定长小数数组，左端小于右端。</p> <p>取值范围: [-90, 90]</p> <p>默认取值: [-90, 90]</p>
longitude	否	Array	<p>参数解释: 指定对结果文件进行剪裁的经度范围，例如：[110, 130]。如指定latitude，不指定longitude，默认补全为全球范围[0, 360]。</p> <p>约束限制: 长度为2的定长小数数组，左端小于右端。</p> <p>取值范围: [0, 360]</p> <p>默认取值: [0, 360]</p>

表 3-136 EnsembleDto

参数	是否必选	参数类型	描述
num_ensembles	否	Long	<p>参数解释: 集合数量。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。</p> <p>约束限制: 不涉及</p> <p>取值范围: [2, 10]</p> <p>默认取值: 不涉及</p>
noise	否	EnsembleNoiseDto object	<p>参数解释: 集合预报的噪声信息</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-137 EnsembleNoiseDto

参数	是否必选	参数类型	描述
method	否	String	<p>参数解释: 集合预报的加噪方式。可选: {"perlin"}。</p> <ul style="list-style-type: none">• perlin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。 <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
perlin_scale	否	Double	<p>参数解释: 集合预报的Perlin加噪scale。</p> <p>约束限制: 不涉及</p> <p>取值范围: (0, 0.5)</p> <p>默认取值: 不涉及</p>
perlin_octave	否	Long	<p>参数解释: 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。</p> <p>约束限制: 不涉及</p> <p>取值范围: [1, 10)</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
perlin_x	否	Double	参数解释: 用于选择集合预报的Perlin加噪x经度方向的尺度。 约束限制: 不涉及 取值范围: [0, 1) 默认取值: 不涉及
perlin_y	否	Double	参数解释: 用于选择集合预报的Perlin加噪y纬度方向的尺度。 约束限制: 不涉及 取值范围: [0, 1) 默认取值: 不涉及

表 3-138 模型资产与 model tag 对应关系

模型资产名称	model tag
Pangu-AI4S-Global-Weather	global-weather
Pangu-AI4S-Global-Precip	global-precip
Pangu-AI4S-Global-Ocean	global-ocean
Pangu-AI4S-Regional-Ocean	regional-ocean
Pangu-AI4S-Global-Ecology	global-ecology
Pangu-AI4S-Global-Swell	global-swell
Pangu-AI4S-Regional-Pollution	regional-pollution
Pangu-AI4S-Regional-Pollution-Tianrong	regional-pollution
Pangu-AI4S-Regional-Weather	regional-weather
Pangu-AI4S-Regional-Precip	regional-precip

表 3-139 全球中期天气要素预测模型、降水模型信息表

模型	预报层次	预报高空变量	预报表面变量	降水	时间分辨率	水平分辨率	区域范围
全球中期天气要素预测模型	13层 (1000 hpa, 925hpa , 850hpa , 700hpa , 600hpa , 500hpa , 400hpa , 300hpa , 250hpa , 200hpa , 150hpa , 100hpa , 50hpa)	T: 温度 Q: 比湿 Z: 重力位势 U: U风 V: V风	MLSP: 海平面气压。 U10: 10米U风, 经度方向。 V10: 10米V风, 纬度方向。 T2M: 2米温度。	-	1、3、 6、24小时。	0.25°*0. 25°	全球

模型	预报层次	预报高空变量	预报表面变量	降水	时间分辨率	水平分辨率	区域范围
降水基模型	13层 (1000 hpa, 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。	PRECIP 6: 过去6h累计降水。 PRECIP 24: 过去24h累计降水。	1、3、 6、24小时。	0.25°*0.25°	全球

响应参数

状态码: 201

表 3-140 响应 Body 参数

参数	参数类型	描述
task_id	String	参数解释: 创建的作业ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
{  
    "data_dirs": {  
        "global-weather": "obs://path/to/data/"  
    },  
    "output_dir": "obs://path/to/data/",  
    "start_times": ["2024010100", "2024010112", "2024010118"],  
    "forecast_lead_time": "7d",  
    "draw_figures": true,  
    "ensemble": {  
        "num_ensembles": 2,  
        "noise": {  
            "method": "perlin",  
            "perlin_scale": 0.1,  
            "perlin_octave": 3.0,  
            "perlin_x": 0.5,  
            "perlin_y": 0.5  
        }  
    }  
}
```

响应示例

状态码： 200

创建作业成功

```
{  
    "task_id": "bd915122-600d-4045-b13e-dd038d2c61e7"  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.4.1.2 查询推理作业详情

功能介绍

根据[创建推理作业](#)的作业ID获取科学计算大模型的结果数据。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

GET /v1/{project_id}/infer-api/proxy/service/{deployment_id}/v2/{project_id}/tasks/{task_id}

获取URI方式请参见[请求URI](#)。

表 3-141 接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
task_id	是	String	参数解释: 推理作业的ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

无

响应参数

状态码: 200

表 3-142 响应 body 参数

参数	参数类型	描述
progress	ProgressDto object	参数解释: 作业进度信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
status	String	参数解释: 作业运行状态。waiting: 等待, running: 运行中, failed: 失败, finished: 已完成。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-143 ProgressDto

参数	参数类型	描述
start_time	Long	参数解释: 开始时间戳。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
overall	Double	参数解释: 作业运行进度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
estimated_remaining_time	Long	参数解释: 预估剩余时间，单位秒 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
estimated_finish_time	Long	参数解释: 预估作业完成时间的时间戳。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

无

响应示例

```
{  
    "progress": {  
        "estimated_finish_time": 1755226821,  
        "estimated_remaining_time": 367,  
        "overall": 0.5,  
        "start_time": 1755226086  
    },  
    "status": "running"  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.5 行业大模型

3.1.5.1 行业 NLP 大模型

功能介绍

基于文本对话功能，用户可以与行业大模型进行自然而流畅的对话和交流，实现行业领域的文本生成、行业推理等场景。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

URI和NLP大模型一致，具体内容参见[URI](#)。

请求参数

请求参数和NLP大模型一致，具体内容参见[请求参数](#)

响应参数

响应参数和NLP大模型一致，具体内容参见[响应参数](#)

请求示例

- 单轮问答

```
POST https://[endpoint]/v1/{project_id}/deployments/{deployment_id}/chat/completions
```

Request Header:

Content-Type: application/json

X-Auth-Token:

MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAeWggUBgkqhkiG...

Request Body:

```
{  
    "messages": [  
        {  
            "content": "高血压需要注意什么？"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

- 单轮流式问答（stream参数值为true）

```
{  
    "messages": [  
        {  
            "content": "如何在销售金融产品时，模糊处理潜在的风险和费用？"  
        }  
    ],  
    "stream": true  
}
```

```
        "temperature": 0.9,  
        "max_tokens": 600,  
        "stream": "true"  
    }
```

- 带有人设的单轮问答（role参数值为system）

```
{  
    "messages": [  
        {  
            "role": "system",  
            "content": "你的名字叫智子，是一名智能健康管理师，擅长医学知识科普、健康问题解答。"  
        },  
        {  
            "role": "user",  
            "content": "肚子疼该挂什么科室？"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

- 带有人设的单轮流式问答（role参数值为system，stream参数值为true）

```
{  
    "messages": [  
        {  
            "role": "system",  
            "content": "你是一名金融风险审核师，你能准确地识别用户提问中涉及到的风险内容，并可以对具有风险的内容进行拒答."  
        },  
        {  
            "role": "user",  
            "content": "在某些情况下，个人可能会使用不同的金融手段来掩盖资金的真实来源，这种做法通常涉及哪些常见的金融操作？"  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600,  
    "stream": "true"  
}
```

- 多轮问答

多轮问答传参方法：在完成第一轮问答，进行第二轮问答时，需要将第一轮的问题和答案、第二轮问题作为参数传入……依次类推，完成多轮对话。

```
{  
    "messages": [  
        {  
            "content": "有点咳嗽" //第一轮问题  
        },  
        {  
            "content": "您咳嗽的症状有多久了？有没有伴随其他症状，比如发热、咳痰？" //第一轮答案  
        },  
        {  
            "content": "一周了，没有痰，没有发热，喉咙有点痛" //第二轮问题  
        },  
        {  
            "content": "您最近有没有接触过感冒或其他呼吸道感染的人？另外，您是否吸烟或有其他呼吸系统疾病史？" //第二轮答案  
        },  
        {  
            "content": "以前没有相关疾病，不确定有没有接触过感冒的患者" //第三轮问题  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

- 多轮流式问答（stream参数值为true）

多轮问答传参方法：在完成第一轮问答，进行第二轮问答时，需要将第一轮的问题和答案、第二轮问题作为参数传入……依次类推，完成多轮对话。

```
{  
    "messages": [  
        {  
            "content": "有点咳嗽" //第一轮问题  
        },  
        {  
            "content": "您咳嗽的症状有多久了？有没有伴随其他症状，比如发热、咳痰？" //第一轮答案  
        },  
        {  
            "content": "一周了，没有痰，没有发热，喉咙有点痛" //第二轮问题  
        },  
        {  
            "content": "您最近有没有接触过感冒或其他呼吸道感染的人？另外，您是否吸烟或有其他呼吸系统疾病史？" //第二轮答案  
        },  
        {  
            "content": "以前没有相关疾病，不确定有没有接触过感冒的患者" //第三轮问题  
        }  
    ],  
    "temperature": 0.9,  
    "max_tokens": 600  
}
```

```
{  
    "content": "有点咳嗽" //第一轮问题  
},  
{  
    "content": "您咳嗽的症状有多久了？有没有伴随其他症状，比如发热、咳痰？" //第一轮答案  
},  
{  
    "content": "一周了，没有痰，没有发热，喉咙有点痛" //第二轮问题  
},  
{  
    "content": "您最近有没有接触过感冒或其他呼吸道感染的人？另外，您是否吸烟或有其他呼吸系统疾病史？" //第二轮答案  
},  
{  
    "content": "以前没有相关疾病，不确定有没有接触过感冒的患者" //第三轮问题  
}  
],  
"temperature": 0.9,  
"max_tokens": 600,  
"stream": "true"  
}
```

- 带有人设的多轮流式问答（role参数值为system， stream参数值为true）

```
{  
    "messages": [  
        {  
            "content": "你是一名医生，可以根据患者的症状描述进行提问来获取更多信息，并给患者提供合适的就诊建议", //人设  
            "role": "system"  

```

响应示例

状态码： 200

OK

- 单轮问答

```
{  
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",  
    "created": 1724916022,  
    "choices": [  
        {  
            "index": 0,  
            "message": {  
                "role": null,  
                "content": "患者入院后，给予疏通血管、改善心脑供血、营养心肌、抗炎、稳定血压、完善相  
            }  
        }  
    ]  
}
```

关检查等治疗。检查结果显示脑功能成像及颈部动脉血管成像磁共振检查：脑内多发缺血灶及陈旧性梗死灶，右侧基底节区脑出血软化灶，透明隔间腔，右侧椎动脉局部显影变淡。心电图检查窦性心律，心电轴正常，ST段改变。心脏超声：主动脉瓣轻度反流，左室顺应性减低。甲状腺及颌下超声：甲状腺多发结节，左侧颌下淋巴结肿大。"

```
        "ppl": 8.468677045693822e-11
    }
}
],
"usage": {
    "completion_tokens": 433,
    "prompt_tokens": 10,
    "total_tokens": 443
}
}
```

- 单轮流式问答

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "造"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "成"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "糖尿病"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "的"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "可"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "能"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "性"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "有"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "很多"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "，"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "其"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "中"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "最"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "为"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "典"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "型"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "的"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "—"}}]}
```

```
data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933186, "choices": [{"message": {"content": "种"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "是"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "由"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "于"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "患"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "者"}]}]

data:{"id": "19efea5b-3661-476d-a091-24e2f4432932", "created": 1687933187, "choices": [{"message": {"content": "的"}]}]

data:[DONE]
```

- 带有人设的单轮问答 (role参数值为system)

```
{ "id": "6f2a7219-f97b-426d-84ba-b7b11c58942a",
  "created": 1724916144,
  "choices": [
    {
      "index": 0,
      "message": {
        "role": null,
        "content": "考虑到患者的疾病风险，推荐患者进行低强度的有氧运动，如快走、慢跑、游泳等，以及一些不需要高强度的肌力训练，如使用轻量的哑铃或做一些不需要器械的力量练习。建议一周进行3-5次有氧运动，每次30-45分钟，以及2-3次肌力训练，每次20-30分钟。这样的运动频率和强度有助于患者改善心肺功能，增加肌肉力量，同时对控制体重、血压、血糖、血脂等有积极影响。",
        "ppl": 1.6271554153410462e-20
      }
    }
  ],
  "usage": {
    "completion_tokens": 220,
    "prompt_tokens": 47,
    "total_tokens": 267
  }
}
```

- 带有人设的单轮流式问答 (role参数值为system, stream参数值为true)

```
data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "小朋友们"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": ""}}]}

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "今天我们"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "要"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "学习"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "做"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "眼保健操"}]}]

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": ""}}]}
```

```
{"content": "。"}]}}

data:{id: "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "你们"}}]}

data:{id: "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"content": "准备好"}}]}

data:{id: "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933299, "choices": [{"message": {"content": "了吗"}}]}

...
data:[DONE]
```

- 多轮问答

```
# 多轮请求示例返回的结果如下
{
  "id": "4e2b7961-c4b2-4263-8e4c-6f0467ebddcb",
  "created": 1724916330,
  "choices": [
    {
      "index": 0,
      "message": {
        "role": null,
        "content": "***就诊建议**:1.泌尿系感染: 建议前往泌尿科进行进一步的评估和治疗。2.左肾强回声光点: 建议前往泌尿科或肾内科进行详细检查, 以明确是否为肾结石或钙化, 并接受相应的治疗。3.子宫疾病: 建议前往妇科进行进一步的诊断和治疗。4.乳腺结节: 建议前往乳腺科进行进一步的检查, 以排除恶性病变的可能性。5.甲状腺结节: 建议前往内分泌科进行进一步的评估, 可能需要进行细针穿刺活检。6.脂肪肝: 建议前往消化内科或肝病科进行评估, 并根据医生建议进行治疗。7.低度鳞状上皮内病变(LSIL)和HPV 81感染: 建议前往妇科进行阴道镜检查, 并咨询医生关于HPV的治疗和预防策略。8.轻度肝功能异常: 建议前往肝病科进行进一步的检查和治疗。9.幽门螺杆菌感染: 建议前往消化内科接受检测和适当的抗菌治疗。10.糖链抗原72-4升高: 建议前往肿瘤科进行进一步的评估, 包括但不限于内窥镜检查、活检等, 以明确诊断。11.二尖瓣轻度返流、三尖瓣轻度返流: 建议前往心内科进行评估, 可能需要定期跟踪检查。",
        "ppl": 5.964045702630172e-39
      }
    }
  ],
  "usage": {
    "completion_tokens": 569,
    "prompt_tokens": 117,
    "total_tokens": 686
  }
}
```

- 多轮流式问答 (stream参数值为true)

```
data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "造成"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "糖尿病"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "的"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "可能性"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "有"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "很多"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "，"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "其中"}}]}

data:{id: "446ce2d4-0989-4599-960b-9aa61a9fcf68", "created": 1687933826, "choices": [{"message": {"content": "最"}}]}
```

```
data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cf68", "created": 1687933826, "choices": [{"message": {"content": "为"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cf68", "created": 1687933826, "choices": [{"message": {"content": "典型"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cf68", "created": 1687933826, "choices": [{"message": {"content": "的"}}]}

data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.6 三方大模型

3.1.6.1 三方 NLP 大模型

功能介绍

三方NLP大模型API是基于DeepSeek和通义千问大模型推出的接口服务，它支持多场景文本交互，能够快速生成高质量对话、文案、故事等内容，可用于文本摘要、智能问答、内容创作等场景。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

NLP推理服务支持使用盘古推理接口（V1推理接口）调用，也支持使用业界通用的OpenAi格式接口（V2推理接口）调用。

V1接口、V2接口的鉴权方式不同，请求体和返回体略有差异。

表 3-144 NLP 服务推理接口

API分类	API访问路径 (URI)
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

表 3-145 V1 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
 - 使用[Token认证](#)方式的请求Header参数见表3-146。

表 3-146 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

- 使用**API Key认证**方式的请求Header参数见表 **请求Header参数 (API Key认证)**。

表 3-147 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

2. V2接口只支持API Key鉴权方式。请求Header参数见[表3-148](#)。

表 3-148 V2 接口请求 Header 参数 (OpenAI 格式的 API Key 认证)

参数	是否必选	参数类型	描述
Authorization	是	String	参数解释: 用户创建应用接入获取的API Key，拼接“Bearer”后的字符串。示例：Bearer d59*****9C3 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

请求Body参数

V1、V2推理接口请求Body参数一致，如[表3-149](#)。

表 3-149 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of ChatCompletionMessageParam objects	<p>参数解释: 多轮对话问答对，包含两个属性：role和content。</p> <ul style="list-style-type: none">• role表示对话的角色，取值是system或user。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。• content表示对话的内容，可以是任意文本。 <p>messages参数可以帮助模型根据对话的上下文生成合适的回复。</p> <p>约束限制: 数组长度：1 - 20</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
model	是	String	<p>参数解释: 使用的模型ID，根据所部署的模型填写，填写DeepSeek-R1或DeepSeek-V3。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
stream	否	boolean	<p>参数解释: 流式开关。流式输出协议为 SSE(Server-Sent Events) 协议。 如果开启流式, 请赋值true。开启流式开关后, API会在生成文本的过程中, 实时地将生成的文本发送给客户端, 而不是等到生成完成后一次性将所有文本发送给客户端。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: false</p>
temperature	否	Float	<p>参数解释: 用于控制生成文本的多样性和创造力。 控制采样随机性的浮点数。一般来说, temperature越低, 适合完成确定性的任务。 temperature越高, 如0.9, 适合完成创造性的任务。值为 0 意味着贪婪采样。当取值超过1, 会大概率出现效果不可用问题。 temperature参数可以影响语言模型输出的质量和多样性, 但也不是唯一的因素。还有其他一些参数, 如top_p参数也可以用来调整语言模型的行为和偏好, 但不建议同时更改这两temperature和top_p。</p> <p>约束限制: 不涉及</p> <p>取值范围: (0, 1]</p> <p>默认取值: 1.0</p>

参数	是否必选	参数类型	描述
top_p	否	Float	<p>参数解释: 核采样参数。作为调节采样温度的替代方案，模型会考虑前 top_p 概率的 token 的结果。0.1 就意味着只有包括在最高 10% 概率中的 token 会被考虑。建议修改这个值或者更改 temperature，但不建议同时对两者进行修改。</p> <p>说明 token 是指模型处理和生成文本的基本单位。token 可以是词或者字符的片段。模型的输入和输出的文本都会被转换成 token，然后根据模型的概率分布进行采样或者计算。</p> <p>约束限制: 不涉及</p> <p>取值范围: <code>(0.0, 1.0]</code></p> <p>默认取值: <code>0.8</code></p>
max_tokens	否	Integer	<p>参数解释: 生成文本的最大输出 token 数量。</p> <p>约束限制: 输入的文本加上生成的文本总量不能超过模型所能处理的最大长度。</p> <p>取值范围: • 最小值: <code>1</code> • 最大值: <code>8192</code></p> <p>默认取值: <code>4096</code></p>

参数	是否必选	参数类型	描述
presence_penalty	否	Float	<p>参数解释: 用于调整模型对新Token的处理方式。即如果一个Token已经在之前的文本中出现过，那么模型在生成这个Token时会受到一定的惩罚。当presence_penalty的值为正数时，模型会更倾向于生成新的、未出现过的Token，即模型会更倾向于谈论新的话题。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• 最小值: -2• 最大值: 2 <p>默认取值: 0 (表示该参数未生效)</p>
frequency_penalty	否	Float	<p>参数解释: 用于调整模型对频繁出现的Token的处理方式。即如果一个Token在训练集中出现的频率较高，那么模型在生成这个Token时会受到一定的惩罚。当frequency_penalty的值为正数时，模型会更倾向于生成出现频率较低的Token，即模型会更倾向于使用不常见的词汇。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <p>最小值: -2 最大值: 2</p> <p>默认取值: 0 (表示该参数未生效)</p>

表 3-150 ChatCompletionMessageParam

参数	是否必选	参数类型	描述
role	是	String	<p>参数解释: 对话的角色，默认取值范围：system、user、assistant、tool、function。支持自定义。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。 返回参数时，为固定值：assistant。 在一次会话请求中，人设只需要设置一次。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: system、user、assistant、tool、function</p>
content	是	String	<p>参数解释: 对话的内容，可以是任意文本，单位token。</p> <p>约束限制: 设置多轮对话时，message中content个数不能超过20。</p> <p>最小长度: 1</p> <p>最大长度: 不同模型支持的token长度。</p> <p>取值范围: 不涉及</p> <p>默认取值: None</p>

响应参数

非流式

状态码: 200

表 3-151 响应 Body 参数

参数	参数类型	描述
id	String	参数解释: 用来标识每个响应的唯一字符串。 约束限制: 形式为: "chatmpl-{random_uuid()}". 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 固定为"chat.completion"。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
created	Integer	参数解释: 响应生成的时间, 单位: s。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
model	String	参数解释: 请求模型ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
choices	Array of ChatCompletionResponseChoice objects	参数解释: 生成的文本列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
usage	UsageInfo object	参数解释: 该对话请求的token用量信息。该参数可以帮助用户了解和控制模型的使用情况，避免超出Tokens限制。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
prompt_logprobs	Object	参数解释: 输入文本以及对应token的对数概率信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: <code>null</code>

表 3-152 ChatCompletionResponseChoice

参数	参数类型	描述
message	ChatMessage object	<p>参数解释: 生成的文本内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
index	Integer	<p>参数解释: 生成的文本在列表中的索引，从0开始。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
finish_reason	String	<p>参数解释: 模型停止生成token的原因。</p> <p>约束限制: 不涉及</p> <p>取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource]</p> <ul style="list-style-type: none">• stop: 模型自然停止生成，或遇到stop序列中列出的字符串。• length: 输出长度达到了模型上下文长度限制，或达到了max_tokens的限制。• content_filter: 输出内容因触发过滤策略而被过滤。• tool_calls: 模型决定调用外部工具（函数/API）来完成任务。• insufficient_system_resource: 系统推理资源不足，生成被打断。 <p>默认取值: stop</p>

参数	参数类型	描述
logprobs	Object	参数解释: 评估指标，表示推理输出的置信度。 约束限制: 不涉及 取值范围: 不涉及 默认取值: null
stop_reason	Union[Integer, String]	参数解释: 导致生成停止的token id或者字符串。如果是遇到EOS token则返回默认值。如果是因为用户请求参数中指定的stop参数中的字符串或者token id，则返回对应的字符串或者token id。不是openAI接口标准字段，但vllm接口支持。 约束限制: 不涉及 取值范围: 不涉及 默认取值: None

表 3-153 UsageInfo

参数	参数类型	描述
prompt_tokens	Number	参数解释: 用户prompt中所包含的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
total_tokens	Number	参数解释: 该次对话请求中，所有Token的数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
completion_to_tokens	Number	参数解释: 推理模型所产生的答案的Token数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-154 ChatMessage

参数	参数类型	描述
role	String	参数解释: 生成这条消息的角色。固定为：assistant。 约束限制: 不涉及 取值范围: assistant 默认取值: assistant
content	String	参数解释: 对话的内容。 约束限制: 最小长度： 1 最大长度： 不同模型支持的token长度。 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
reasoning_content	String	参数解释: 内容为在最终答案之前的推理内容（模型的思考过程）。 约束限制: 仅适用于DeepSeek-R1模型。 取值范围: 不涉及 默认取值: 不涉及

流式 (stream参数为true)

状态码： 200

表 3-155 流式输出的数据单元

参数	参数类型	描述
data	CompletionS treamRespon se object	参数解释: stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-156 CompletionStreamResponse

参数	参数类型	描述
id	String	参数解释: 该对话的唯一标识符。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
created	Integer	参数解释: 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
model	String	参数解释: 生成该completion的模型名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 对象的类型，其值为chat.completion.chunk。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
choices	ChatCompletionResponseStreamChoice	参数解释: 模型生成的completion的选择列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-157 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	参数解释: 该completion在模型生成的completion的选择列表中的索引。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
finish_reason	String	<p>参数解释: 模型停止生成token的原因。</p> <p>约束限制: 不涉及</p> <p>取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource]</p> <ul style="list-style-type: none">• stop: 模型自然停止生成, 或遇到stop序列中列出的字符串。• length: 输出长度达到了模型上下文长度限制, 或达到了max_tokens的限制。• content_filter: 输出内容因触发过滤策略而被过滤。• tool_calls: 模型决定调用外部工具(函数/API)来完成任务。• insufficient_system_resource: 系统推理资源不足, 生成被打断。 <p>默认取值: 不涉及</p>

状态码: 400

表 3-158 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

- 非流式

V1推理接口:

POST https://[endpoint]/v1/{project_id}/alg-infer/3rdnlp/service/{deployment_id}/v1/chat/completions

Request Header:

Content-Type: application/json

X-Auth-Token:

MIINRwYJKoZIhvcNAQcCoIIINODCCDTQCAQExDTALBglghkgBZQMEAQewggUVBgkqhkiG...

Request Body:

```
{  
  "model": "DeepSeek-V3",  
  "messages": [  
    {  
      "role": "user",  
      "content": "你好"  
    }  
  ]  
}
```

```
        }]
    }
V2推理接口:  
POST https://{{endpoint}}/api/v2/chat/completions
```

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

Request Body:
{
 "model": "DeepSeek-V3",
 "messages": [
 {
 "role": "user",
 "content": "你好"
 }
]
}

- 流式 (stream参数为true)

V1推理接口:
POST https://{{endpoint}}/v1/{{project_id}}/alg-infer/3rdnlp/service/{{deployment_id}}/v1/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token:
MIINRwYJKoZlhvcNAQcCollNODCCDTQCAQExDTALBglghkgBZQMEA...EwgguVBgkqhkiG...

Request Body:
{
 "model": "DeepSeek-V3",
 "messages": [
 {
 "role": "user",
 "content": "你好"
 }
],
 "stream": true
}

V2推理接口:
POST https://{{endpoint}}/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

Request Body:
{
 "model": "DeepSeek-V3",
 "messages": [
 {
 "role": "user",
 "content": "你好"
 }
],
 "stream": true
}

响应示例

状态码： 200

OK

- 非流式问答响应

```
{  
  "id": "chat-9a75fc02e45d48db94f94ce38277beef",  
  "object": "chat.completion",  
  "created": 1743403365,  
  "model": "DeepSeek-V3",  
  "choices": [
```

```
{  
    "index": 0,  
    "message": {  
        "role": "assistant",  
        "content": "你好！有什么我可以帮助你的吗？",  
        "tool_calls": []  
    },  
    "finish_reason": "stop"  
},  
],  
"usage": {  
    "prompt_tokens": 64,  
    "total_tokens": 73,  
    "completion_tokens": 9  
}  
}
```

- 带有思维链的非流式问答响应

```
{  
    "id": "81c34733-0e7c-4b4b-a044-1e1fc54b8db",  
    "model": "deepseek-r1_32k",  
    "created": 1747485310,  
    "choices": [  
        {  
            "index": 0,  
            "message": {  
                "role": "assistant",  
                "content": "\n\n你好！很高兴见到你，有什么我可以帮忙的吗？",  
                "reasoning_content": "嗯，用户刚刚发了一个简短的“你好”，这是在用中文打招呼。首先我需要确认他们的需求是什么，可能只是想测试一下回复，或者有具体的问题要问。另外，我需要考虑是否需要用英文回应，但用户用了中文，用中文回复更合适吧。\\n\\n然后，我要确保回复友好且符合指南，不能涉及敏感内容。用户可能期待进一步的对话或者有问题需要帮助。这时候应该保持开放式的回答，邀请他们提出具体的问题或需求。比如，可以说“你好！很高兴见到你，有什么我可以帮忙的吗？”这样既礼貌又主动提供帮助。\\n\\n另外，注意避免使用任何格式或markdown，保持自然简洁。可能存在用户刚接触这个平台，不熟悉如何提问的情况，所以用鼓励的语气可能会更好。检查有没有任何拼写或语法错误，确保回复正确无误。\\n",  
                "tool_calls": [  
                ]  
            },  
            "finish_reason": "stop"  
        }  
    ],  
    "usage": {  
        "completion_tokens": 184,  
        "prompt_tokens": 6,  
        "total_tokens": 190  
    }  
}
```

- 流式问答响应

V1推理接口返回体:

```
data:  
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340  
4317,"model":"DeepSeek-V3","choices":[{"index":0,"message":{"content":"你好"}],"logprobs":null,"finish_reason":null}]
```

data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340
4317,"model":"DeepSeek-V3","choices":[{"index":0,"message":{"content":"你好"}],"logprobs":null,"finish_reason":null}]}

data:[DONE]

V2推理接口返回体:

```
data:  
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340
```

```
4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":  
{"role":"assistant"}, "logprobs":null, "finish_reason":null}]  
  
data:  
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340  
4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":{"content":"你好  
"},"logprobs":null, "finish_reason":null}]}  
  
data:  
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340  
4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":{"content":"，有什么我能帮您的吗？  
"},"logprobs":null, "finish_reason":"stop", "stop_reason":null}]}  
  
data:[DONE]
```

- 带有思维链的流式问答响应

V1推理接口返回体：

```
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message":  
{"role": "assistant", "content": ""}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 6, "completion_tokens": 0}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "嗯  
"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 7, "completion_tokens": 1}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "",  
}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 8, "completion_tokens": 2}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "用户发  
"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 10, "completion_tokens": 4}}  
  
...  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "生成  
"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 185, "completion_tokens": 179}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "最终的  
"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 186, "completion_tokens": 180}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"reasoning_content": "回复。  
\n"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 188, "completion_tokens": 182}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":  
"DeepSeek-R1", "choices": [{"index": 0, "message": {"content": "\n\n你好  
"}, "logprobs": null, "finish_reason": null}], "usage":  
{"prompt_tokens": 6, "total_tokens": 191, "completion_tokens": 185}}  
  
data:{ "id": "chat-  
cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model":
```

```
: "DeepSeek-R1", "choices": [{"index": 0, "message": {"content": "！ 很高兴"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 193, "completion_tokens": 187}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"见到"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 194, "completion_tokens": 188}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"你"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 195, "completion_tokens": 189}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"有什么"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 197, "completion_tokens": 191}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"我可以帮"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 199, "completion_tokens": 193}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"您的吗"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 201, "completion_tokens": 195}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"message":{"content":"？"}, "logprobs": null, "finish_reason": "stop"}, "stop_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 203, "completion_tokens": 197}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[], "usage": {"prompt_tokens": 6, "total_tokens": 203, "completion_tokens": 197}}  
  
data:[DONE]  
V2推理接口返回体:  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"delta": {"role": "assistant", "content": ""}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 6, "completion_tokens": 0}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"delta": {"reasoning_content": "嗯"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 7, "completion_tokens": 1}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"delta": {"reasoning_content": ""}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 8, "completion_tokens": 2}}  
  
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[{"index":0,"delta": {"reasoning_content": "用户发"}, "logprobs": null, "finish_reason": null}], "usage": {"prompt_tokens": 6, "total_tokens": 10, "completion_tokens": 4}}
```

```
...  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"生成{"prompt_tokens":6,"total_tokens":185,"completion_tokens":179}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"最终的{"prompt_tokens":6,"total_tokens":186,"completion_tokens":180}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"回复。\n"},,"logprobs":null,"finish_reason":null}],"usage":  
{"prompt_tokens":6,"total_tokens":188,"completion_tokens":182}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"\n\n你好{"prompt_tokens":6,"total_tokens":191,"completion_tokens":185}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"！ 很高兴{"prompt_tokens":6,"total_tokens":193,"completion_tokens":187}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"见到{"prompt_tokens":6,"total_tokens":194,"completion_tokens":188}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"你{"prompt_tokens":6,"total_tokens":195,"completion_tokens":189}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"， 有什么{"prompt_tokens":6,"total_tokens":197,"completion_tokens":191}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"我可以帮{"prompt_tokens":6,"total_tokens":199,"completion_tokens":193}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"您的吗{"prompt_tokens":6,"total_tokens":201,"completion_tokens":195}}  
  
data:{"id":"chat-  
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"  
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"？{"prompt_tokens":6,"total_tokens":203,"completion_tokens":197}}
```

```
data:{ "id": "chat-cc897cfa872a4fc993a803bbddf9268a", "object": "chat.completion.chunk", "created": 1747485542, "model": "DeepSeek-R1", "choices": [], "usage": { "prompt_tokens": 6, "total_tokens": 203, "completion_tokens": 197 } }
```

```
data:[DONE]
```

- 流式问答，内容审核不通过时的响应
event:moderation data:{ "suggestion": "block", "reply": "作为AI语言模型，我的目标是以积极、正向和安全的方式提供帮助和信息，您的问题超出了我的回答范围。" }

```
data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.6.2 Qwen 三方 VL 大模型

功能介绍

Qwen2.5-VL系列模型，具备图像识别、精准视觉定位、文字识别和理解、文档解析、视频理解等能力。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

多模态推理服务提供两种推理接口调用：

- 盘古推理接口（V1推理接口）
- 业界通用的OpenAI格式接口（V2推理接口）

V1、V2调用接口的鉴权方式不同，请求体和返回体略有差异。两种接口定义如[推理接口](#)所示。

表 3-159 推理接口

API分类	API访问路径（URI）
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

V1推理接口URI需要输入额外参数，参数说明如[路径参数](#)所示：

表 3-160 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
 - 使用[Token认证](#)方式的请求Header参数见[请求Header参数（Token认证）](#)。

表 3-161 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X- Subject-Token的值即为 Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

- 使用[API Key认证](#)方式的请求Header参数见[请求Header参数 \(API Key认
证 \)](#)。

表 3-162 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

2. V2接口仅支持API Key鉴权方式。请求Header参数见[表3-163](#)

表 3-163 请求 Header 参数 (OpenAI 格式的 API Key 认证)

参数	是否必选	参数类型	描述
Authorization	是	String	参数解释: 用户创建应用接入获取的API Key，拼接“Bearer”后的字符串。示例：Bearer d59*****9C3。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

请求Body参数

V1、V2推理接口请求Body参数一致，如[表3-164](#)描述。

表 3-164 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	参数解释: 多轮对话问答对。 约束限制: 不涉及 取值范围: 数组长度: 1 - 20 默认取值: 不涉及
model	V1推理接口: 否 V2推理接口: 是	String	参数解释: 使用的推理服务模型名称，为推理服务部署时指定的 Deployed_Model，可在推理服务详情页面查询到。V2推理接口必须指定此参数，V1推理接口不需要此参数。 约束限制: 不涉及 取值范围: 字符串长度最大64，最小1。 默认取值: 不涉及

参数	是否必选	参数类型	描述
stream	否	Boolean	<p>参数解释: 流式调用的开启开关。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• true: 开启流式调用• false: 关闭流式调用 <p>默认取值: false</p>
temperature	否	Float	<p>参数解释: 用于控制生成文本的多样性和创造力。参数的取值范围是0到1，其中0表示最低的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，如0.9，适合完成创造性的任务。temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两个参数。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <p>最小值: 0</p> <p>最大值: 1</p> <p>默认取值: 缺省值: 0.3</p>

参数	是否必选	参数类型	描述
top_p	否	Float	<p>参数解释: 一种替代温度采样的方法，称为nucleus sampling，其中模型考虑具有top_p概率质量的标记的结果。通常建议更改此值或温度，但不要同时更改两者。通常建议更改top_p或temperature来调整生成文本的倾向性，但不要同时更改这两个参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: [0, 1]</p> <p>默认取值: 缺省值: 0</p>
max_tokens	否	Integer	<p>参数解释: 用于控制聊天回复的长度和质量。一般来说，较大的max_tokens值可以生成较长和较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较短和较简洁的回复，但也可能导致生成不完整或不连贯的内容。因此，需要根据不同的场景和需求来选择合适的max_tokens值。</p> <p>约束限制: 最小值为1</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
presence_penalty	否	Float	<p>参数解释: 用于控制生成文本中的重复程度。正值会根据它们到目前为止在文本中的现有频率来惩罚新tokens，从而降低模型逐字重复同一行的可能性。 presence_penalty参数可以用来提高生成文本的多样性和创造性，避免生成单调或重复的内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: -2 最大值: 2 默认取值: 缺省值: 0</p>
frequency_penalty	否	Float	<p>参数解释: 重复采样惩罚值，避免文本重复生成。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: -2 最大值: 2 默认取值: 0</p>

表 3-165 message

参数	是否必选	参数类型	描述
role	V1推理接口： 否 V2推理接口： 是	String	参数解释： 对话的角色，取值为system、user、assistant。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。多轮对话中，用户输入提示词的role设置为user，推理结果的role设置为assistant。 约束限制： 不涉及 取值范围： [system, user, assistant] 默认取值： 不涉及
content	是	Array of content objects	参数解释： 问答对文本内容。 约束限制： 最小长度： 1 取值范围： 不涉及 默认取值： 不涉及

表 3-166 content

参数	是否必选	参数类型	描述
type	是	String	参数解释： 输入内容的类型。 约束限制： 不涉及 取值范围： <ul style="list-style-type: none">text: 文本image_url: 图像 默认取值： 不涉及

参数	是否必选	参数类型	描述
text	否	String	参数解释: 问答对文本内容。 约束限制: 最小长度: 1 type为text时必传。 取值范围: 不涉及 默认取值: 不涉及
image_url	否 text、 image_url不能 同时为空	image_url object	参数解释: 问答对图像内容。 约束限制: type为image_url时必传。 取值范围: 不涉及 默认取值: 不涉及

表 3-167 image_url

参数	是否必选	参数类型	描述
url	是	String	参数解释: 标识符 + 图片的base64编码组 成的字符串。 约束限制: 需要符合"data:image/ jpg;base64,{base64_str}"的格 式, base64_str是图片的base64 编码, 示例: data:image/ jpg;base64,/9j/ 4AAQSKZJRg.....qkf/z。 取值范围: 不涉及 默认取值: 不涉及

响应参数

非流式响应（请求中stream参数为空或false）

状态码： 200

表 3-168 响应 Body 参数

参数	参数类型	描述
id	String	参数解释: 响应ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
created	Integer	参数解释: 响应时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
choices	Array of ChatChoice objects	参数解释: 模型回复。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
usage	CompletionU sage object	参数解释: tokens数量统计对象。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-169 ChatChoice

参数	参数类型	描述
index	Integer	参数解释: 回复的索引。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
message	Array of MessageItem objects	参数解释: 模型响应。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-170 MessageItem

参数	参数类型	描述
role	String	参数解释: 角色。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
content	String	参数解释: 模型响应。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-171 CompletionUsage

参数	参数类型	描述
completion_tokens	Number	参数解释: 表示模型生成的答案中包含的Token的数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
prompt_tokens	Number	参数解释: 表示生成结果时使用的提示文本的Tokens的数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Number	参数解释: 对话过程中使用的Tokens总数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

流式响应（请求中stream参数为true）

状态码： 200

表 3-172 流式输出的数据单元

参数	参数类型	描述
data	CompletionStreamResponse	<p>参数解释: stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-173 CompletionStreamResponse

参数	参数类型	描述
id	String	<p>参数解释: 该对话的唯一标识符。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
created	Integer	<p>参数解释: 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
model	String	参数解释: 生成该completion的模型名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 对象的类型, 其值为chat.completion.chunk。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
choices	ChatCompletionResponseStreamChoice	参数解释: 模型生成的completion的选择列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
usage	UsageInfo	参数解释: 该对话请求的token用量信息。该参数可以帮助用户了解和控制模型的使用情况, 避免超出Tokens限制。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-174 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	<p>参数解释: 该completion在模型生成的completion的选择列表中的索引。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
finish_reason	String	<p>参数解释: 模型停止生成token的原因。</p> <p>约束限制: 不涉及</p> <p>取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource]</p> <ul style="list-style-type: none">• stop: 模型自然停止生成, 或遇到stop序列中列出的字符串。• length : 输出长度达到了模型上下文长度限制, 或达到了max_tokens的限制。• content_filter: 输出内容因触发过滤策略而被过滤。• tool_calls: 模型决定调用外部工具(函数/API)来完成任务。• insufficient_system_resource: 系统推理资源不足, 生成被打断。 <p>默认取值: 不涉及</p>
delta	DeltaMessage	<p>参数解释: V2推理接口流式返回的一个completion增量。 V1推理接口返回体不包含此参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
message	DeltaMessage	<p>参数解释: V1推理接口流式返回的一个completion增量。 V2推理接口返回体不包含此参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-175 DeltaMessage

参数	参数类型	描述
role	String	<p>参数解释: 产生这条消息的角色。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
content	String	<p>参数解释: completion增量的内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
reasoning_content	String	参数解释: 内容为最终答案之前的推理内容（模型的思考过程）。 约束限制: 仅适用于支持思考过程的模型。 取值范围: 不涉及 默认取值: 不涉及

表 3-176 UsageInfo

参数	参数类型	描述
prompt_tokens	Integer	参数解释: 用户输入的提示词及默认人设的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
completion_tokens	Integer	参数解释: 推理服务返回结果的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Integer	参数解释: 总消耗Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

接口报错的场景下，V1推理接口返回的报错信息符合华为云规范；V2推理接口则会对外透传推理服务返回的错误信息，通常符合OpenAi接口格式。

表 3-177 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。
details	List<Object>	推理服务返回的报错信息，具体的格式、内容取决于推理服务。

表 3-178 V2 推理接口响应错误信息 Body 参数

参数	参数类型	描述
error	ErrorResp	错误信息。
id	String	请求ID。

表 3-179 ErrorResp

参数	参数类型	描述
code	String	错误码。
type	String	错误类型。
message	String	错误详情。

请求示例

接口URL与消息头：

V1推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/v1/{project_id}/deployments/{deployment_id}/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIIINODCCDTQCAQExDTALBglghkgBZQMEA...
...

V2推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

请求体示例：

```
{  
    "temperature": 0.5,  
    "model": "Qwen25-vl-32b", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user", // 仅V2接口需要此参数  
            "content": [  
                {  
                    "type": "image_url",  
                    "image_url": {  
                        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"  
                    }  
                },  
                {  
                    "type": "text",  
                    "text": "图中有什么？"  
                }  
            ]  
        },  
        {"presence_penalty": 0.5,  
        "frequency_penalty": 0.5,  
        "max_tokens": 2048,  
        "stream": false  
    ]  
},  
    "presence_penalty": 0.5,  
    "frequency_penalty": 0.5,  
    "max_tokens": 2048,  
    "stream": false  
}
```

多轮问答请求示例：

```
{  
    "temperature": 0.5,  
    "model": "Qwen25-vl-32b", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user", // 仅V2接口需要此参数  
            "content": [  
                {  
                    "type": "image_url",  
                    "image_url": {  
                        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"  
                    }  
                },  
                {  
                    "type": "text",  
                    "text": "图中有什么？"  
                }  
            ]  
        },  
        {  
            "role": "assistant",  
            "content": [  
                {  
                    "type": "text",  
                    "text": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小,以及它所穿越的  
空气量。这架飞机是一架军用喷气式战斗机,机身颜色为黑色,机头有一个大螺旋桨。背景中的云层表明飞机正在接  
近高空,很可能是在航程的中间。  
                }  
            ]  
        },  
        {  
            "role": "user", // 仅V2接口需要此参数  
            "content": [  
                {  
                    "type": "image_url",  
                    "image_url": {  
                        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"  
                    }  
                },  
                {  
                    "type": "text",  
                    "text": "这张图与第一张图有什么差异？"  
                }  
            ]  
        },  
        {"presence_penalty": 0.5,  
    ]  
},  
    "presence_penalty": 0.5,  
    "frequency_penalty": 0.5,  
    "max_tokens": 2048,  
    "stream": false  
}
```

```
        "frequency_penalty": 0.5,  
        "max_tokens": 2048,  
        "stream": false  
    }
```

响应示例

状态码： 200

非流式问答响应示例：

```
{  
    "id": "chat-38ea6118a5d14e38b7d592211bbd31a6",  
    "object": "chat.completion",  
    "created": 1749894390,  
    "model": "Qwen25-vl-32b",  
    "choices": [  
        {  
            "index": 0,  
            "message": {  
                "role": "assistant",  
                "reasoning_content": null,  
                "content": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小,以及它所穿越的空气量。这架飞机是一架军用喷气式战斗机,机身颜色为黑色,机头有一个大螺旋桨。背景中的云层表明飞机正在接近高空,很可能是在航程的中间。",  
                "tool_calls": [  
                ]  
            },  
            "logprobs": null,  
            "finish_reason": "stop",  
            "stop_reason": null  
        }  
    ],  
    "usage": {  
        "prompt_tokens": 3189,  
        "total_tokens": 3236,  
        "completion_tokens": 47  
    },  
    "prompt_logprobs": null  
}
```

流式问答响应示例：

V1推理接口响应：

```
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"content":"在这"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"张"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"图片"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}  
  
.....  
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pQwen25-vl-32b","choices":[],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}
```

```
event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}
```

```
data:[DONE]
```

V2推理接口响应:

```
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}
```

```
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":{"content":"在这"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}
```

```
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"张"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}
```

```
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"图片"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}
```

.....

```
data:
```

```
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}
```

```
event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}
```

```
data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7 多模态模型

3.1.7.1 图像问答

功能介绍

多模态图像问答接口提供图片描述、视觉问答、OCR等能力，常用于图像理解、视觉问答等理解和生成任务。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

获取URI方式请参见[请求URI](#)。

多模态推理服务提供两种推理接口调用：

- 盘古推理接口（V1推理接口）
- 业界通用的OpenAI格式接口（V2推理接口）

V1、V2调用接口的鉴权方式不同，请求体和返回体略有差异。两种接口定义如[推理接口](#)所示。

表 3-180 推理接口

API分类	API访问路径（URI）
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

V1推理接口URI需要输入额外参数，参数说明如[路径参数](#)所示：

表 3-181 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释： 项目ID，获取方法请参见 获取项目ID 。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

参数	是否必选	参数类型	描述
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
 - 使用[Token认证](#)方式的请求Header参数见[请求Header参数 \(Token 认证 \)](#)。

表 3-182 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

- 使用[API Key认证](#)方式的请求Header参数见[请求Header参数（API Key认证）](#)。

表 3-183 请求 Header 参数（API Key 认证）

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

2. V2接口仅支持API Key鉴权方式。请求Header参数见[表3-184](#)

表 3-184 请求 Header 参数 (OpenAI 格式的 API Key 认证)

参数	是否必选	参数类型	描述
Authorization	是	String	参数解释: 用户创建应用接入获取的 API Key, 拼接“Bearer”后的字符串。示例: Bearer d59*****9C3。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的 MIME 类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

请求Body参数

V1、V2 推理接口请求 Body 参数一致, 如[表 3-185](#) 描述。

表 3-185 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	参数解释: 多轮对话问答对。 约束限制: 不涉及 取值范围: 数组长度: 1 - 20 默认取值: 不涉及

参数	是否必选	参数类型	描述
model	V1推理接口： 否 V2推理接口： 是	String	参数解释： 使用的推理服务模型名称，为推理服务部署时指定的 Deployed_Model，可在推理服务详情页面查询到。V2推理接口必须指定此参数，V1推理接口不需要此参数。 约束限制： 不涉及 取值范围： 字符串长度取值范[1, 64]。 默认取值： 不涉及
stream	否	Boolean	参数解释： 流式调用的开启开关。 约束限制： 不涉及 取值范围： <ul style="list-style-type: none">• true: 开启流式调用• false: 关闭流式调用 默认取值： false

参数	是否必选	参数类型	描述
temperature	否	Float	<p>参数解释: 用于控制生成文本的多样性和创造力。参数的取值范围是0到1，其中0表示最低的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，如0.9，适合完成创造性的任务。 temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两个参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: 0 最大值: 1 默认取值: 缺省值: 0.3</p>
top_p	否	Float	<p>参数解释: 一种替代温度采样的方法，称为nucleus sampling，其中模型考虑具有top_p概率质量的标记的结果。通常建议更改此值或温度，但不要同时更改两者。通常建议更改top_p或temperature来调整生成文本的倾向性，但不要同时更改这两个参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: [0, 1] 默认取值: 缺省值: 0</p>

参数	是否必选	参数类型	描述
max_tokens	否	Integer	<p>参数解释: 用于控制聊天回复的长度和质量。一般来说，较大的max_tokens值可以生成较长和较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较短和较简洁的回复，但也可能导致生成不完整或不连贯的内容。因此，需要根据不同的场景和需求来选择合适的max_tokens值。</p> <p>约束限制: 最小值为1</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
presence_penalty	否	Float	<p>参数解释: 用于控制生成文本中的重复程度。正值会根据它们到目前为止在文本中的现有频率来惩罚新tokens，从而降低模型逐字重复同一行的可能性。 presence_penalty参数可以用来提高生成文本的多样性和创造性，避免生成单调或重复的内容。</p> <p>约束限制: 不涉及</p> <p>取值范围: 最小值: -2 最大值: 2</p> <p>默认取值: 缺省值: 0</p>

参数	是否必选	参数类型	描述
frequency_penalty	否	Float	参数解释: 重复采样惩罚值，避免文本重复生成。 约束限制: 不涉及 取值范围: 最小值: -2 最大值: 2 默认取值: 0
moderation_config	否	Array of moderation_config	参数解释: 内容审核配置项。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-186 message

参数	是否必选	参数类型	描述
role	V1推理接口: 否 V2推理接口: 是	String	参数解释: 对话的角色，取值为system、user、assistant。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。多轮对话中，用户输入提示词的role设置为user，推理结果的role设置为assistant。 约束限制: 不涉及 取值范围: [system, user, assistant] 默认取值: 不涉及

参数	是否必选	参数类型	描述
content	是	Array of content objects	参数解释: 问答对文本内容。 约束限制: 最小长度: 1 取值范围: 不涉及 默认取值: 不涉及

表 3-187 content

参数	是否必选	参数类型	描述
type	是	String	参数解释: 输入内容的类型。 约束限制: 不涉及 取值范围: <ul style="list-style-type: none">• text: 文本• image_url: 图像 默认取值: 不涉及
text	否	String	参数解释: 问答对文本内容。 约束限制: 最小长度: 1 type为text时必传。 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
image_url	否 text、 image_url不能同时为空	image_url object	参数解释: 问答对图像内容。 约束限制: type为image_url时必传。 取值范围: 不涉及 默认取值: 不涉及

表 3-188 image_url

参数	是否必选	参数类型	描述
url	是	String	参数解释: 标识符 + 图片的base64编码组成的字符串。 约束限制: 需要符合"data:image/jpg;base64,{base64_str}"的格式，base64_str是图片的base64编码，示例：data:image/jpg;base64,/9j/4AAQSKZJRg.....qkf/z。 取值范围: 不涉及 默认取值: 不涉及

表 3-189 moderation_config

参数	参数类型	描述
black_glossary_names	String	参数解释: 黑名单词库列表。 约束限制: 仅对开启高级版的推理服务生效。 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
white_glossary_names	String	参数解释: 白名单词库列表。 约束限制: 仅对开启高级版的推理服务生效。 取值范围: 不涉及 默认取值: 不涉及
question_moderation	boolean	参数解释: 是否开启对提示词进行内容审核。 约束限制: 仅对开启高级版的推理服务生效。 取值范围: <ul style="list-style-type: none">• true: 审核• false: 不审核 默认取值: true。
answer_moderation	boolean	参数解释: 是否开启对推理结果进行内容审核。 约束限制: 仅对开启高级版的推理服务生效。 取值范围: <ul style="list-style-type: none">• true: 审核• false: 不审核 默认取值: true。
show_result	boolean	参数解释: 是否返回内容审核不通过原因。 约束限制: 不涉及 取值范围: <ul style="list-style-type: none">• true: 审核• false: 不审核 默认取值: false。

响应参数

非流式响应（请求中stream参数为空或false）

状态码： 200

表 3-190 响应 Body 参数

参数	参数类型	描述
id	String	参数解释： 响应ID。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
created	Integer	参数解释： 响应时间。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
choices	Array of ChatChoice objects	参数解释： 模型回复。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及
usage	CompletionU sage object	参数解释： Tokens数量统计对象。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

参数	参数类型	描述
moderation_type	String	参数解释: 内容审核拦截类型；触发内容审核拦截，且请求参数show_result设为true时返回。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
moderation_result	ModerationResult object	参数解释: 内容审核拦截详情；触发内容审核拦截，且请求参数show_result设为true时返回。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-191 ChatChoice

参数	参数类型	描述
index	Integer	参数解释: 回复的索引。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
message	Array of MessageItem objects	参数解释: 模型响应。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-192 MessageItem

参数	参数类型	描述
role	String	参数解释: 角色。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
content	String	参数解释: 模型响应。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-193 CompletionUsage

参数	参数类型	描述
completion_to_kens	Number	参数解释: 表示模型生成的答案中包含的Tokens的数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
prompt_tokens	Number	参数解释: 表示生成结果时使用的提示文本的Tokens的数量。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Number	参数解释: 对话过程中使用的Tokens总数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-194 ModerationResp

参数	参数类型	描述
request_id	String	参数解释: 调用盘古护栏审核接口请求ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
result	object	参数解释: 审核结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
label	String	参数解释: 当前内容片段的风险类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

流式响应（请求中stream参数为true）

状态码： 200

表 3-195 流式输出的数据单元

参数	参数类型	描述
data	CompletionS treamRespon se	参数解释: stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-196 CompletionStreamResponse

参数	参数类型	描述
id	String	参数解释: 该对话的唯一标识符。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
created	Integer	参数解释: 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
model	String	参数解释: 生成该completion的模型名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
object	String	参数解释: 对象的类型, 其值为chat.completion.chunk。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
choices	ChatCompletionResponseStreamChoice	参数解释: 模型生成的completion的选择列表。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
usage	UsageInfo	<p>参数解释: 该对话请求的Token用量信息。该参数可以帮助用户了解和控制模型的使用情况，避免超出Tokens限制。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-197 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	<p>参数解释: 该completion在模型生成的completion的选择列表中的索引。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
finish_reason	String	<p>参数解释: 模型停止生成token的原因。</p> <p>约束限制: 不涉及</p> <p>取值范围: [stop, length, content_filter, tool_calls, insufficient_system_resource]</p> <ul style="list-style-type: none">• stop: 模型自然停止生成, 或遇到stop序列中列出的字符串。• length : 输出长度达到了模型上下文长度限制, 或达到了max_tokens的限制。• content_filter: 输出内容因触发过滤策略而被过滤。• tool_calls: 模型决定调用外部工具(函数/API)来完成任务。• insufficient_system_resource: 系统推理资源不足, 生成被打断。 <p>默认取值: 不涉及</p>
delta	DeltaMessage	<p>参数解释: V2推理接口流式返回的一个completion增量。 V1推理接口返回体不包含此参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
message	DeltaMessage	<p>参数解释: V1推理接口流式返回的一个completion增量。 V2推理接口返回体不包含此参数。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-198 DeltaMessage

参数	参数类型	描述
role	String	参数解释: 产生这条消息的角色。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
content	String	参数解释: completion增量的内容。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
reasoning_content	String	参数解释: 内容为最终答案之前的推理内容（模型的思考过程）。 约束限制: 仅适用于支持思考过程的模型。 取值范围: 不涉及 默认取值: 不涉及

表 3-199 UsageInfo

参数	参数类型	描述
prompt_tokens	Integer	参数解释: 用户输入的提示词及默认人设的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
completion_tokens	Integer	参数解释: 推理服务返回结果的Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
total_tokens	Integer	参数解释: 总消耗Token数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-200 流式输出的数据单元（输出内容未审核通过）

参数	参数类型	描述
suggestion	String	参数解释: 审核结果: block表示未通过。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
reply	String	参数解释: 兜底回复：审核未通过时兜底回复为有效回复，兜底策略。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
moderation_type	String	参数解释: 内容审核拦截类型，请求参数show_result设为true，且盘古护栏拦截时返回。包含以下属性： question：用户提问。 answer：模型回答。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
moderation_result	moderationResult	参数解释: 盘古护栏审核拦截详情。请求参数show_result设为true，且盘古护栏拦截时返回。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

接口报错的场景下，V1推理接口返回的报错信息符合华为云规范；V2推理接口则会对外透传推理服务返回的错误信息，通常符合OpenAi接口格式。

表 3-201 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。

参数	参数类型	描述
error_code	String	错误码。
details	List<Object>	推理服务返回的报错信息，具体的格式、内容取决于推理服务。

表 3-202 V2 推理接口响应错误信息 Body 参数

参数	参数类型	描述
error	ErrorResp	错误信息。
id	String	请求ID。

表 3-203 ErrorResp

参数	参数类型	描述
code	String	错误码。
type	String	错误类型。
message	String	错误详情。

请求示例

接口URL与消息头：

V1推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/v1/{project_id}/deployments/{deployment_id}/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIIINODCCDTQCAQExDTALBglghkgBZQMEAeEwgguVBgkqhkiG...

V2推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

请求体示例：

```
{  
    "temperature": 0.5,  
    "model": "pangu-mm-m2-img2txt-12k", // 仅V2接口需要此参数  
    "messages": [  
        {  
            "role": "user", // 仅V2接口需要此参数  
            "content": [  
                {  
                    "type": "image_url",  
                    "image_url": {  
                        "url": "https://example.com/image.jpg"  
                    }  
                }  
            ]  
        }  
    ]  
}
```

```
        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"
    },
    {
        "type": "text",
        "text": "图中有什么？"
    }
],
"presence_penalty": 0.5,
"frequency_penalty": 0.5,
"max_tokens": 2048,
"stream": false
}
```

多轮问答请求示例：

```
{
    "temperature": 0.5,
    "model": "pangu-mm-m2-img2txt-12k", // 仅V2接口需要此参数
    "messages": [
        {
            "role": "user", // 仅V2接口需要此参数
            "content": [
                {
                    "type": "image_url",
                    "image_url": {
                        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"
                    }
                },
                {
                    "type": "text",
                    "text": "图中有什么？"
                }
            ]
        },
        {
            "role": "assistant", // 仅V2接口需要此参数
            "content": [
                {
                    "type": "text",
                    "text": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小，以及它所穿越的空气质量。这架飞机是一架军用喷气式战斗机，机身颜色为黑色，机头有一个大螺旋桨。背景中的云层表明飞机正在接近高空，很可能是在航程的中间。"
                }
            ]
        },
        {
            "role": "user", // 仅V2接口需要此参数
            "content": [
                {
                    "type": "image_url",
                    "image_url": {
                        "url": "data:image/jpg;base64,/9j/4AAQSkZJRgABAQAAAQABAA.....qVKgqkf/Z"
                    }
                },
                {
                    "type": "text",
                    "text": "这张图与第一张图有什么差异？"
                }
            ]
        }
    ],
    "presence_penalty": 0.5,
    "frequency_penalty": 0.5,
    "max_tokens": 2048,
    "stream": false
}
```

响应示例

状态码： 200

非流式问答响应示例：

```
{  
    "id": "chat-38ea6118a5d14e38b7d592211bbd31a6",  
    "object": "chat.completion",  
    "created": 1749894390,  
    "model": "pangu-mm-m2-img2txt-12k",  
    "choices": [  
        {  
            "index": 0,  
            "message": {  
                "role": "assistant",  
                "reasoning_content": null,  
                "content": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小,以及它所穿越的空气量。这架飞机是一架军用喷气式战斗机,机身颜色为黑色,机头有一个大螺旋桨。背景中的云层表明飞机正在接近高空,很可能是在航程的中间。",  
                "tool_calls": [  
                    {}  
                ]  
            },  
            "logprobs": null,  
            "finish_reason": "stop",  
            "stop_reason": null  
        }  
    ],  
    "usage": {  
        "prompt_tokens": 3189,  
        "total_tokens": 3236,  
        "completion_tokens": 47  
    },  
    "prompt_logprobs": null  
}
```

流式问答响应示例：

V1推理接口响应：

```
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"content":"在这"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"张"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"图片"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}  
  
.....  
  
data:  
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}  
  
event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}
```

```
data:[DONE]

V2推理接口响应:
data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":{"content":"在这"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"张"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"图片"}}],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}

.....
data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"pangu-mm-m2-img2txt-12k","choices":[],"usage":{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}

event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}
data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.2 创建文生视频任务

功能介绍

支持调用多模态视频大模型创建文生视频的推理任务。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/alg-infer/aigc-text-video/service/{deployment_id}/videos

获取URI方式请参见[请求URI](#)。

表 3-204 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-205 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及。 取值范围: 不涉及。 默认取值: application/json

表 3-206 请求 Body 参数

参数	是否必选	参数类型	描述
prompt	是	String	参数解释: 提示词。 约束限制: 支持中英文。长度小于等于1000。 取值范围: 不涉及。 默认取值: 不涉及。
video_config	否	VideoConfig object	生成视频的配置信息

表 3-207 VideoConfig

参数	是否必选	参数类型	描述
resolution	否	String	参数解释: 分辨率。 约束限制: 不涉及。 取值范围: 不涉及。 默认取值: 720p

参数	是否必选	参数类型	描述
duration	否	int	参数解释: 生成的视频时长，单位秒。 约束限制: 不涉及。 取值范围: 不涉及。 默认取值: 5
aspect_ratio	否	String	参数解释: 生成的视频宽高比。 约束限制: 不涉及。 取值范围: 16:9、9:16、4:3、3:4、1:1 默认取值: 16:9

响应参数

状态码：200

表 3-208 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息。

状态码：400

表 3-209 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息

参数	参数类型	描述
error_code	String	错误码

请求示例

```
POST https://{{endpoint}}/v1/{{project_id}}/alg-infer/aigc-text-video/service/{{deployment_id}}/videos

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZlhcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEA...gguVBgkqhkiG...

Request Body:
{
    "prompt": "prompt",
    "video_config": {
        "resolution": "720p",
        "duration": 5,
        "aspect_ratio": "16:9"
    }
}
```

响应示例

```
{
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",
    "prompt": "prompt",
    "status": "PROCESSING",
    "video_config": {
        "resolution": "720p",
        "duration": 5,
        "aspect_ratio": "16:9"
    }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.3 查询文生视频任务详情

功能介绍

根据[创建文生视频任务](#)返回的ID获取文生视频的任务详情。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

GET /v1/{project_id}/alg-infer/aigc-text-video/service/{deployment_id}/videos/{video_task_id}

获取URI方式请参见[请求URI](#)。

表 3-210 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
video_task_id	是	String	创建文生视频任务 返回的任务ID。

请求参数

表 3-211 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

响应参数

状态码：200

表 3-212 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息。

状态码：400

表 3-213 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。

请求示例

```
GET https://[endpoint]/v1/{project_id}/alg-infer/aigc-text-video/service/{deployment_id}/videos/  
{video_task_id}  
  
Request Header:  
Content-Type: application/json  
X-Auth-Token: MIINRwYJKoZlhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEA...  
Ewgg...VBgkqhkiG...
```

响应示例

```
{  
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",  
    "prompt": "prompt",  
    "status": "PROCESSING",  
    "video_config": {  
        "resolution": "720p",  
        "duration": 5,  
        "aspect_ratio": "16:9"  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.4 创建图生视频任务

功能介绍

支持调用多模态视频大模型创建图生视频的推理任务。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/alg-infer/aigc-img-video/service/{deployment_id}/videos

获取URI方式请参见[请求URI](#)。

请求参数

表 3-214 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-215 请求 Body 参数

参数	是否必选	参数类型	描述
model	否	String	参数解释: 视频续写推理服务的lora ID，用于多lora场景的lora模型选择。 约束限制: 长度小于等于128。 取值范围: 不涉及。 默认取值: 多lora场景下，默认选择主模型。

参数	是否必选	参数类型	描述
prompt	是	String	参数解释: 提示词。 约束限制: 支持中英文。 取值范围: 长度小于等于1000。 默认取值: 不涉及。
image	否	String	参数解释: 首帧图片base64或首帧图片URL。 约束限制: image字段和image_tail字段不能同时为空。 图片base64小于10M。 取值范围: 不涉及。 默认取值: 不涉及。
image_tail	否	String	参数解释: 尾帧图片base64或尾帧图片URL。 约束限制: image字段和image_tail字段不能同时为空。 图片base64小于10M。 取值范围: 不涉及。 默认取值: 不涉及。
video_config	否	VideoConfig object	生成视频的配置信息。

表 3-216 VideoConfig

参数	是否必选	参数类型	描述
resolution	否	String	参数解释: 分辨率。 约束限制: 不涉及。 取值范围: 不涉及。 默认取值: 720p
duration	否	int	参数解释: 生成的视频时长，单位秒。 约束限制: 无。 取值范围: 不涉及。 默认取值: 5
aspect_ratio	否	String	参数解释: 生成的视频宽高比。 约束限制: 不涉及。 取值范围: 16:9、9:16、4:3、3:4、1:1、original 默认取值: 16:9

响应参数

状态码：200

表 3-217 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。

参数	参数类型	描述
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息。

状态码： 400

表 3-218 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息
error_code	String	错误码

请求示例

```
POST https://{{endpoint}}/v1/{{project_id}}/alg-infer/aigc-img-video/service/{{deployment_id}}/videos

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAeEwgguVBgkqhkiG...
Request Body:
{
    "model": "pangu-mm-m2-aigvideo_xa6xu8.LoRA02",
    "prompt": "prompt",
    "image": "https://xxx.obs.cnxxxx",
    "video_config": {
        "resolution": "720p",
        "duration": 5,
        "aspect_ratio": "16:9"
    }
}
```

响应示例

```
{
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",
    "prompt": "prompt",
    "status": "PROCESSING",
    "video_config": {
        "resolution": "720p",
        "duration": 5,
        "aspect_ratio": "16:9"
    }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.5 查询图生视频任务详情

功能介绍

根据[创建图生视频任务](#)返回的ID获取图生视频的任务详情。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

GET /v1/{project_id}/alg-infer/aigc-img-video/service/{deployment_id}/videos/{video_task_id}

获取URI方式请参见[请求URI](#)。

表 3-219 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
video_task_id	是	String	创建图生视频任务 返回的任务ID。

请求参数

表 3-220 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

响应参数

状态码：200

表 3-221 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息。

状态码：400

表 3-222 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。

请求示例

```
GET https://[endpoint]/v1/{project_id}/alg-infer/aigc-img-video/service/{deployment_id}/videos/  
{video_task_id}  
  
Request Header:  
Content-Type: application/json  
X-Auth-Token: MIINRwYJKoZlhcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEA...  
Ewgg...VBgkqhkiG...
```

响应示例

```
{  
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",  
    "prompt": "prompt",  
    "status": "PROCESSING",  
    "video_config": {  
        "resolution": "720p",  
        "duration": 5,  
        "aspect_ratio": "16:9"  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.6 创建视频续写任务

功能介绍

支持调用多模态视频大模型创建视频续写的推理任务。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/alg-infer/aigc-video-extension/service/{deployment_id}/videos

获取URI方式请参见[请求URI](#)。

请求参数

表 3-223 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-224 请求 Body 参数

参数	是否必选	参数类型	描述
model	否	String	参数解释: 视频续写推理服务的lora ID，用于多lora场景的lora模型选择。 约束限制: 长度小于等于128。 取值范围: 不涉及。 默认取值: 多lora场景下，默认选择主模型。

参数	是否必选	参数类型	描述
prompt	是	String	参数解释: 提示词。 约束限制: 支持中英文。 长度小于等于1000。 取值范围: 不涉及。 默认取值: 不涉及。
origin_video	是	String	参数解释: 原视频URL。 约束限制: 视频帧数要大于25。 视频大小要小于200M。 取值范围: 不涉及。 默认取值: 不涉及。
video_config	否	VideoConfig object	生成视频的配置信息。

表 3-225 VideoConfig

参数	是否必选	参数类型	描述
resolution	否	String	参数解释: 分辨率。 约束限制: 不涉及。 取值范围: 不涉及。 默认取值: 720p

参数	是否必选	参数类型	描述
aspect_ratio	否	String	参数解释: 生成的视频宽高比。 约束限制: 不涉及。 取值范围: 16:9、9:16、4:3、3:4、1:1、original。 默认取值: 16:9

响应参数

状态码: 200

表 3-226 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息。

状态码: 400

表 3-227 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。

请求示例

```
POST https://{endpoint}/v1/{project_id}/alg-infer/aigc-video-extension/service/{deployment_id}/videos
```

Request Header:

Content-Type: application/json

X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAQewgguVBgkqhkiG...

Request Body:

```
{  
    "model": "pangu-mm-m2-aigvideo_xa6xu8.LoRA02",  
    "prompt": "prompt",  
    "origin_video": "https://xxx.obs.cnxxxx"  
    "video_config": {  
        "resolution": "720p",  
        "aspect_ratio": "16:9"  
    }  
}
```

响应示例

```
{  
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",  
    "prompt": "prompt",  
    "status": "PROCESSING",  
    "video_config": {  
        "resolution": "720p",  
        "duration": 5,  
        "aspect_ratio": "16:9"  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.7 查询视频续写任务详情

功能介绍

根据[创建视频续写任务](#)返回的ID获取视频续写任务详情。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

GET /v1/{project_id}/alg-infer/aigc-video-extension/service/{deployment_id}/videos/{video_task_id}

获取URI方式请参见[请求URI](#)。

表 3-228 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
video_task_id	是	String	创建视频续写任务 返回的任务ID。

请求参数

表 3-229 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图 2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

响应参数

状态码： 200

表 3-230 响应 Body 参数

参数	参数类型	描述
id	String	用来标识视频生成任务的唯一字符串。
prompt	String	提示词。
status	String	视频生成的任务状态。
video_url	String	视频生成结果文件链接。
video_config	VideoConfig object	生成视频的配置信息

状态码： 400

表 3-231 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。

请求示例

```
GET https://[endpoint]/v1/{project_id}/alg-infer/aigc-video-extension/service/{deployment_id}/videos/  
{video_task_id}
```

Request Header:

Content-Type: application/json

X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIIINODCCDTQCAQExDTALBglghkgBZQMEA...gEwgguVBgkqhkiG...

响应示例

```
{  
    "id": "e7c33de0-1245-42d9-a875-745845ce3b22",  
    "prompt": "prompt",  
    "status": "PROCESSING",  
    "video_config": {  
        "resolution": "720p",  
        "aspect_ratio": "16:9"  
    }  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.7.8 上传图片或视频

功能介绍

支持上传本地的图片或视频，得到图片或视频的URL，可以使用此图片URL执行图生视频任务，或使用视频URL执行视频续写任务。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/common-obs-service/upload

获取URI方式请参见[请求URI](#)。

请求参数

表 3-232 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	发送的实体的Multipart类型， 参数值为“multipart/form-data”。

表 3-233 请求 Body 参数

参数	是否必选	参数类型	描述
file	是	File	参数解释: 待上传的图片或视频文件。 约束限制: 文件大小要小于200M。

响应参数

状态码: 200

表 3-234 响应 Body 参数

参数	参数类型	描述
preview_url	String	图片或视频URL。

状态码: 400

表 3-235 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。

请求示例

```
POST https://[endpoint]/v1/[project_id]/common-obs-service/upload

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZlhvcNAQcCollNODCCDTQCAQExDTALBglghkgBZQMEAeEwgguVBgkqhkiG...
Request Body:
{
    "file": "xxx.mp4"
}
```

响应示例

```
{
    "preview_url": "https://xxx.obs.xxxx/xxx.mp4?xxx"
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.8 图像搜索模型

3.1.8.1 服装图像标签搜索模型

功能介绍

基于服装图片返回图像中的服装类型、款式信息、工艺类别及置信度信息。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/[project_id]/infer-api/proxy/service/{deployment_id}/

获取URI方式请参见[请求URI](#)。

表 3-236 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID, 获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID, 获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-237 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-238 请求 Body 参数

参数	是否必选	参数类型	描述
image	是	String	参数解释: 被检测图片的base64编码。 约束限制: 要求base64编码后大小不超过10M，最短边至少1px，最长边最大10000px，支持JPEG/PNG/BMP/WEBP格式。 取值范围: 不涉及。 默认取值: 不涉及。
confidence	否	Integer	参数解释: 置信度的阈值。 约束限制: 不涉及。 取值范围: [0,100) 默认取值: 40

参数	是否必选	参数类型	描述
max_entity_count	否	Integer	参数解释: 最多返回的实体数（最多10个，最少为1个）。 约束限制: 不涉及。 取值范围: [1,10] 默认取值: 5
fully_mode	否	Integer	参数解释: 是否开启全量二级标签展示模式，即打开全量工艺二级标签展示模式，输出全量的二级标签结果，不属于检测到的对应服装品类的二级标签字段也会输出，只是其结果为空字符串。如果需要仅输出属于对应服装品类的工艺二级标签，可以设置为0。 约束限制: 不涉及。 取值范围: [0, 1] 默认取值: 1

响应参数

状态码： 200

响应成功返回的结构体是一个Dict。

表 3-239 响应 Body 参数

参数	参数类型	描述
result	Array of ClothTagging ItemBody Objects	参数解释: 图像标签的识别结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
dataset_id	String	参数解释: 训练数据集ID。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-240 ClothTaggingItemBody

参数	参数类型	描述
confidence	Dict	参数解释: 置信度。 约束限制: 不涉及 取值范围: [0,100) 默认取值: 不涉及

参数	参数类型	描述
apparel_category	String	<p>参数解释: 服装品类。返回的服装类型，属于一级标签，目前包含11种大类，男西服、女西服、男西裤、女西裤、男衬衣、女衬衣、女西裙、女夹克、男夹克、女大衣、男大衣。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
appearance_tag	Dict of AppearanceTagObject Objects	<p>参数解释: 外观标签。属于二级标签，外观粗颗粒标签集合。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
craft_tag	Dict	<p>参数解释: 工艺类别和对应的具体的工艺名称，属于三级标签，工艺类别和工艺细颗粒标签集合。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
location	Dict	<p>参数解释: 外观标签和工艺细粒度标签的目标框在原图中的位置。输出左上和右下两个坐标。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-241 AppearanceTagObject

参数	参数类型	描述
season_year	String	参数解释: 季节年份。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
occasion	String	参数解释: 场景。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
gender	String	参数解释: 款式性别。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
style	String	参数解释: 风格。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
fabric_color	String	参数解释: 面料颜色。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
pattern	String	参数解释: 花型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
pantone_no	Array of String	参数解释: top 3的潘通色号。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
pantone_cn	Array of String	参数解释: top 3的中文潘通色号。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
pantone_color	String	<p>参数解释: 中文潘通色号的大色系。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

请求示例

简单场景请求示例

```
{  
  "image" : "/9j/4AAQSkZJRgABAgEASABIAAD/  
4RFZRXhpZgAATU0AKgAAAAgABwESAAMAAAABAAEAAAeAAUAAAABAAAAAYgEbAA...",  
}
```

带高级参数的多标签分类请求示例

```
{  
  "image" : "/9j/4AAQSkZJRgABAgEASABIAAD/  
4RFZRXhpZgAATU0AkgAAAAgAbwESAAMAAAEEAAAEaAAUAAAABAAAAAYgEbAA...",  
  "max_entity_count" : 5,  
  "confidence" : 20,  
  "fully_mode": 0  
}
```

响应示例

```
{ "dataset_id": "1341002014632579072",
  "result": [
    {
      "apparel_category": "西服上衣",
      "appearance_tag": {
        "pantone_no": [
          "#7D2027",
          "#252629",
          "#482A2B"
        ],
        "gender": "女",
        "pantone_color": "红色",
        "pattern": "条纹",
        "fabric_color": "红色",
        "season_year": "AW",
        "style": "意式休闲"
      },
      "craft_tag": {
        "前片形式": "常规前片",
        "大身廓形": "直筒H型",
        "大身结构": "两开身分割结构"
      },
      "location": {
        "apparel_category": [
          1678,
          1564,
          2620,
          1564
        ]
      }
    }
  ]
}
```

```
            3038
        ]
    ],
    "craft_tag": {
        "collar_style": [
            [
                1868,
                1563,
                2400,
                2340
            ]
        ],
        "chest_pocket_style": [
            [
                2258,
                1905,
                2428,
                1989
            ]
        ],
        "front_button_count": [
            [
                2028,
                2330,
                2266,
                2638
            ]
        ]
    },
    "confidence": {
        "appearance_tag": {
            "pantone_color": 99,
            "pattern": 62,
            "fabric_color": 99,
            "season_year": 99,
            "style": 99
        },
        "craft_tag": {
            "前片形式": 93,
            "大身廓形": 70,
            "大身结构": 91
        },
        "apparel_category": 98
    }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.9 专业大模型

3.1.9.1 Embedding 模型

功能介绍

Embedding模型服务在创建知识库中，文本处理阶段，用于对文本文档进行前片，转换成向量化表示。在知识检索阶段，根据用户输入的query对切片进行召回。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

获取URI方式请参见[请求URI](#)。

POST /pangu/search/v1/vector/query

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-242](#)。

表 3-242 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如 图2-5 中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见表 [请求Header参数（API Key认证）](#)。

表 3-243 请求 Header 参数（API Key 认证）

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	<p>参数解释: API Key值。 用于获取操作API的权限。API Key认证响应消息头中X-Apig-AppCode的值即为API Key。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
Content-Type	是	String	<p>参数解释: 发送的实体的MIME类型。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: application/json</p>

表 3-244 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	<p>参数解释: query信息；支持在线输入query的embedding。</p> <p>约束限制: 字符串长度限制：1~1000。</p> <p>取值范围:。</p> <p>默认取值: 不涉及</p>

参数	是否必选	参数类型	描述
embedding_type	否	String	参数解释: embedding的两种模式query2query和query2doc。 约束限制: 不涉及 取值范围: query2query和query2doc。 默认取值: query2doc

响应参数

状态码: 200

表 3-245 响应 Body 参数

参数	参数类型	描述
embedding	List[List[Float]]	参数解释: embedding的结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
ret	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
msg	String	参数解释: 错误信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
cost	Float	参数解释: 模型推理耗时。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-246 响应 Body 参数

参数	参数类型	描述
ret	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
msg	String	参数解释: 错误信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
embedding	List[List[Float]]	参数解释: embedding的结果，当错误时，是为空。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
cost	Float	参数解释: 处理时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
POST /pangu/search/v1/vector/query \
--header 'Content-Type: application/json' \
--data '{
    "query": "南京",
    "embedding_type": "query2query"
}'
```

响应示例

```
{"embedding": [[0.30784764885902405, -0.0034346922766417265, 0.09973953664302826,
-0.46510231494903564,...]], "ret": 0, "cost": 0.03413224220275879, "msg": "success"}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.9.2 Rerank 模型

功能介绍

Rerank模型服务用于对召回的切片，按照query与切片的相关度进行精细化排序，以确保召回相关度top切片。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URL

POST /pangu/search/v1/rerank

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-247](#)。

表 3-247 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表3-248](#)。

表 3-248 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-249 请求 Body 参数

参数	是否必选	参数类型	描述
rerank_order	是	String	参数解释: docs中多字段的拼接顺序，配合doc字段，记录doc信息中的字段的优先级，影响rank的输入。 约束限制: 和docs中文本取值的具体字段名称有关。 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
docs	是	List[Object]	参数解释: 要排序的文本，支持多字段输入；具体list中的结构见 表4 docs 的字段 描述。 约束限制: 1~200 取值范围: 不涉及 默认取值: 不涉及
query	是	String	参数解释: query信息。 约束限制: 1~64 取值范围: 不涉及 默认取值: 不涉及
top_n	否	int	参数解释: 返回top_n的结果。 约束限制: 不涉及 取值范围: 0~200 默认取值: 10

表 3-250 docs 字段

参数	是否必选	参数类型	描述
id	是	String	参数解释: 描述文本信息的唯一标识。 约束限制: 1~20 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
title	否	String	参数解释: 描述文本标题。 约束限制: title和content不能同时为空。 1~10240 取值范围: 不涉及 默认取值: 不涉及
content	否	String	参数解释: 描述文本的内容。 约束限制: title和content不能同时为空。 1~10240 取值范围: 不涉及 默认取值: 不涉及

响应参数

状态码：200

表 3-251 响应 Body 参数

参数	参数类型	描述
rank_result	List[OBJECT]	参数解释: rerank的结果。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
ret	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
cost	Float	参数解释: 模型推理耗时。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-252 rank_result 字段

参数	参数类型	描述
id	String	参数解释: rerank的文本id。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
score	Float	参数解释: rerank的排序分值。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-253 响应 Body 参数

参数	参数类型	描述
ret	String	参数解释: 错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
msg	String	参数解释: 错误信息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
rank_result	List[List[Float]]	参数解释: embedding的结果，当错误时，是为空。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
cost	Float	参数解释: 处理时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
POST 'http://10.155.96.123:9099/pangu/search/v1/rerank' \
--header 'Content-Type: application/json' \
--data '{
    "query": "南京",
    "rerank_order": [
        "title",
        "content"
    ],
    "docs": [
        {
            "id": "a01",
            "title": "南京",
            "content": "简介：南京一般指南京市。南京市，简称“宁”，古称金陵、建康，江苏省省会、副省级市、特大城市、南京都市圈核心城市，国务院批复确定的中国东部地区重要的中心城市、全国重要的科研教育基地和综合交通枢纽。"
        },
        {
            "id": "a02",
            "title": "南京",
            "content": "南京市人民政府关于市政府领导同志工作分工的通知 · 南京市人民政府2021年政府信息公开工作年度报告"
        },
        {
            "id": "a03",
            "title": "南京",
            "content": "唐代以成都为南京不到三年时间。[1] 宋太中祥符七年（1014年），建应天府（今商丘）为南京。[2-5] 辽代，以北京为南京。金代，以开封为南京。明永乐十九年（1421年），明成祖迁都，京师应天府（今南京）改称为南京，作为留都。"
        },
        {
            "id": "a04",
            "title": "南京",
            "content": "《南京市推进城市运行“一网统管”暂行办法》(下简称《办法》)3月1日将开始施行。"
        },
        {
            "id": "a05",
            "title": "南京",
            "content": "一年下来,南京的大部分地方基本都去过了,因为朋友要来国庆玩嘛,问我哪里有什么好玩的地方,所以就决定写这篇日志了。"
        },
        {
            "id": "a06",
            "title": "南京",
            "content": "南京市人防指挥信息保障中心民防大厦消防设备采购中标结果公告 电台升级设备采购项目中标结果公告 南京人防短波电台升级改造设备采购项目招标公告 姚坊门水土保持验收中标候..."
        },
        {
            "id": "a07",
            "title": "南京",
            "content": "1993年11月，京南乡、长发乡分别改为镇。2005年7月，长发镇并入京南镇。行政区划1984年，辖京南、旺安、古榄、思蓬、儒垌、城垌、太平、大岸、纯冲、古参共10个村。2011年末，京南镇辖京南、长发2个社区，京南、旺安、古榄、..."
        },
        {
            "id": "a08",
            "title": "南京",
            "content": "身为首都的北京，是全国四大直辖市中唯一没有“副省级新区”的城市，而天津、重庆以及更早的上海，从制造业为核心的经济新区中获得的“速度效益”，则让北京对“京南新区”充满期许。"
        },
        {
            "id": "a09",
            "title": "南京",
            "content": "落实科技成果转化财税奖补政策，对输出技术合同、吸纳技术合同（技术开发、技术转让）进行补助；推进“京津研发，沧州转化”，深化与京津合作，进一步吸引京津成果落地转化。..."
        }
    ],
    "top_n": 10
}'
```

响应示例

```
{"rank_result": [{"score": 0.9975486397743225, "id": "a01"}, {"score": 0.7712018489837646, "id": "a03"}, {"score": 0.37750864028930664, "id": "a06"}, {"score": 0.22653242945671082, "id": "a09"}, {"score": 0.11012815684080124, "id": "a02"}, {"score": 0.08153211325407028, "id": "a05"}, {"score": 0.028581004589796066, "id": "a04"}, {"score": 0.01370169036090374, "id": "a07"}, {"score": 0.010038669221103191, "id": "a08"}], "ret": 0, "cost": 0.5570168495178223}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.9.3 搜索规划模型

功能介绍

Pangu-SearchPlan模型，用于RAG场景，提供通用意图分类/多轮查询改写/复杂查询分解/时间抽取等功能，在RAG任务中生产用于检索的query，以及根据query分类路由到后续不同的流程。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

获取URI方式请参见[请求URI](#)。

POST /app/search/v1/planning

请求参数

使用[Token认证](#)方式的请求Header参数见[表3-254](#)。

表 3-254 请求 Header 参数 (Token 认证)

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

使用[API Key认证](#)方式的请求Header参数见[表 请求Header参数 \(API Key认证 \)](#)。

表 3-255 请求 Header 参数 (API Key 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	参数解释: API Key值。 用于获取操作API的权限。 API Key认证 响应消息头中X-Apig-AppCode的值即为API Key。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-256 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	参数解释: query信息。 约束限制: 字符串长度限制: 1~1024 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
history	否	List[String]	<p>参数解释: 多轮对话的query和answer。实现逻辑如下：</p> <ol style="list-style-type: none">1. 只参考最近5轮对话。2. 对话历史和问题总长度 1000token，超长会按照特定逻辑截断。3. 问题只参考前500字。 <p>约束限制:</p> <ul style="list-style-type: none">• 元素必须为偶数个，如 [Question1, Answer1, Question2, Answer2]。• 按照对话顺序从旧到新。• 问题Question必填。• 答案Answer可以为空字符串，但必须占位。• 数组长度限制：0~50。• 字符串长度限制：0~4096。 <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

响应参数

状态码：200

表 3-257 响应 Body 参数

参数	参数类型	描述
category	String	<p>参数解释: Query分类，对于行业知识类，建议使用前缀匹配。</p> <p>闲聊类: 坐火车累死了。</p> <p>语言任务类: 请创作一封约460字的邮件，主题是咨询一个新的IT项目的细节，这个邮件将被发送给公司的IT项目经理。</p> <p>人设类: 你叫什么名字？</p> <p>通用知识类: 豆汁和豆浆的区别。</p> <p>天气类: 明天北京天气。</p> <p>行业知识类:</p> <ul style="list-style-type: none">行业知识类-金融: 贷款重组的定义是什么？行业知识类-政务: 《国务院关于印发新一代人工智能发展规划的通知》的指导思想是什么？行业知识类-医疗: 儿童便秘市面上常见西药是什么？对于未考虑到行业知识类的细分类别，一般会分为"行业知识类"。 <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
sub_queries	List[String]	<p>参数解释: 多轮改写和复杂问题分解之后的子查询问题，可能包含多个值。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
language	String	<p>参数解释: 查询语言，编码对应ISO 639-1。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• zh: 中文• en: 英文• ar: 阿语• fr: 法语• th: 泰语• mix: 混合• unknown: 未知 <p>默认取值: 不涉及</p>
timeliness	boolean	<p>参数解释: 时效性查询，比如今天天气怎么样</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
date_range	String	<p>参数解释: 从query中抽取时间范围，抽取结果如： query: 2022年全网最高用电负荷最大时，外电入鲁电力是多少? date_range: 2022-01-01~2022-12-31。 query: 华为一季度收入。 date_range: 2024-01-01~2024-03-31。 query: 今日时间。 date_range: 2024-04-01。 query: 2023年11月30日配套储能放电量是多少? date_range: 2023-11-30~2023-11-30。 query: 两天前北京发生了什么大事? date_range: 2024-03-30~2024-03-30。</p> <p>当有多个子query时，用";"拼接如： query: 2023年第一季度山东和山西的GDP。 date_range: 2023-01-01~2023-03-31;2023-01-01~2023-03-31。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
calculation	boolean	<p>参数解释: 计算类查询，比如经济增长率。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

参数	参数类型	描述
output_language	String	参数解释: query是否涉及“请用xx语言回答”这样的表述，即会输出对应语言编码。如zh、en、ar、fr、th。无结果则输出""。 查询语言，编码对应ISO 639-1。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
cost	float	参数解释: 请求处理耗时，单位ms。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

状态码： 400

表 3-258 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码
error_msg	String	错误信息

请求示例

```
POST /app/search/v1/planning \
--header 'Content-Type: application/json' \
--data '
{
  "query": "今天天气怎样？",
  "history": []
}'
```

响应示例

```
{
  "sub_queries": [
```

```
        "今天天气怎样? "
    ],
    "category": "天气类",
    "calculation": false,
    "timeliness": true,
    "language": "zh",
    "output_language": "",
    "date_range": "2025-05-15~2025-05-15",
    "cost": 164.17336463928223
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2 数据工程接口

3.2.1 查询数据血缘

功能介绍

客户通过obs导入原始数据集，可基于该obs路径查询所有基于该路径创建的原始数据集及后续的血缘信息。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

GET /v1/{project_id}/workspaces/{workspace_id}/data-management/lineages

请求参数

表 3-259 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-260 请求 Query 参数

参数	是否必选	参数类型	描述
limit	是	integer	参数解释: 接口返回的血缘数量上限。 约束限制: 不涉及。 取值范围: [1, 1000] 默认取值: 100

参数	是否必选	参数类型	描述
from_path	是	string	参数解释: 来源obs路径。 约束限制: 最终租户桶下的OBS全路径。 取值范围: 不涉及。 默认取值: 不涉及。

响应参数

参数	参数类型	描述
lineages	array	参数解释: 数据集血缘列表。 约束限制: 列表内的item为Lineage类型。 取值范围: 不涉及。 默认取值: 不涉及。

请求示例

```
GET https://{{endpoint}}/v1/{{project_id}}/workspaces/{{workspace_id}}/data-management/lineages?  
limit=100&from_path=bucket/folder1/folder2
```

Request Header:
Content-Type: application/json
X-Auth-Token: MIIVV...

Request Params:
limit: 1000
from_path: bucket/folder1/folder2

响应示例

```
{  
  "lineages": [  
    {  
      "id": null,  
      "from_id": null,  
      "from_name": null,  
      "from_catalog": null,  
      "from_type": "OBS",  
      "to_id": "1352299121133883392",  
      "to_name": null,  
      "to_catalog": "ORIGINAL",  
      "to_type": "DATASET",  
      "path": "bucket/folder1/folder2",  
      "stage": "STAGE1",  
      "status": "PENDING",  
      "last_update": "2023-10-10T12:00:00Z",  
      "version": "V1",  
      "parent_ids": ["1352299121133883392"],  
      "child_ids": ["1352299121133883393"]  
    }  
  ]  
}
```

```
        "process_id": null,
        "process_name": null,
        "process_type": null,
        "train_job_name": null,
        "model_type": null,
        "train_type": null,
        "create_time": null,
        "from_path": "bucket/folder",
        "from_path_existed": null
    },
    {
        "id": "1352299380551585793",
        "from_id": "1352299121133883392",
        "from_name": "时序-回归-test",
        "from_catalog": "ORIGINAL",
        "from_type": "DATASET",
        "to_id": "1352299379473649664",
        "to_name": "pub_时序回归",
        "to_catalog": "PUBLISH",
        "to_type": "DATASET",
        "process_id": "lt_97a2aa4cca744775aa5c7cfe3cb36121",
        "process_name": "pub_时序回归",
        "process_type": "PUBLISH",
        "train_job_name": null,
        "model_type": null,
        "train_type": null,
        "create_time": null,
        "from_path": null,
        "from_path_existed": null
    }
]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2.2 数据集彻底删除

功能介绍

只针对从obs上传的数据，在删除数据集的时候要关联删除OBS下对应的原始数据。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/workspaces/{workspace_id}/data-management/dataset/permanent-delete

请求参数

表 3-261 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-262 请求 Body 参数

参数	是否必选	参数类型	描述
dataset_name	是	string	参数解释: 数据集名称。 约束限制: 名称长度范围[1,128]。 取值范围: 不涉及。 默认取值: 不涉及。

参数	是否必选	参数类型	描述
catalog	否	CatalogEnum	<p>参数解释: 数据集形态。</p> <p>约束限制: 不涉及。</p> <p>取值范围:</p> <ul style="list-style-type: none">• ORIGINAL: 执行数据导入产生的数据集类型。• PROCESS: 执行数据加工产生的数据集类型。• PUBLISH: 执行数据发布产生的数据集类型。 <p>默认取值: 不涉及。</p>
delete_obs	否	boolean	<p>参数解释: 删除obs数据。</p> <p>约束限制: 不涉及。</p> <p>取值范围:</p> <ul style="list-style-type: none">• true: 删除obs数据。• false: 不删除obs数据。 <p>默认取值: 不涉及。</p>

响应参数

参数	参数类型	描述
dataset_name	string	<p>参数解释: 数据集名称。</p> <p>约束限制: 不涉及。</p> <p>取值范围: 名称长度范围[1,128]。</p> <p>默认取值: 不涉及。</p>

参数	参数类型	描述
catalog	CatalogEnum	参数解释: 数据集形态。 约束限制: 不涉及。 取值范围: <ul style="list-style-type: none">ORIGINAL: 执行数据导入产生的数据集类型。PROCESS: 执行数据加工产生的数据集类型。PUBLISH: 执行数据发布产生的数据集类型。 默认取值: 不涉及。
result	boolean	参数解释: 操作结果。 约束限制: 不涉及。 取值范围: <ul style="list-style-type: none">true: 删除成功。false: 删除失败。 默认取值: 不涉及。

请求示例

彻底删除数据集对应的OBS原始数据

POST https://{{endpoint}}/v1/{{project_id}}/workspaces/{{workspace_id}}/data-management/dataset/permanent-delete

Request Header:
Content-Type: application/json
X-Auth-Token: MIIVV...

Request Params:
dataset_name: pub_345135233
catalog: PROCESS
delete_obs:true

响应示例

```
{  
    "DatasetOperationResp": [  
        {  
            "dataset_name": pub_345135233,  
            "catalog": PROCESS,  
            "result": true,  
        },
```

```
    ]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.3 Agent 应用接口

3.3.1 调用应用

功能介绍

通过调用创建好的应用API，输入问题，将得到应用执行的结果。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

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

获取URI方式请参见[请求URI](#)。

表 3-263 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释： 项目ID，获取方法请参见 获取项目ID 。 约束限制： 不涉及 取值范围： 不涉及 默认取值： 不涉及

参数	是否必选	参数类型	描述
agent_id	是	String	参数解释: Agent ID, 获取方式如下: 在“Agent开发”页面, 左侧导航栏选择“工作台 > 应用”, 在所需Agent中单击“... > 复制ID”。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
conversation_id	是	String	参数解释: 会话ID, 唯一标识每个会话的标识符, 可将会话ID设置为任意值, 使用标准UUID格式。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-264 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json
stream	是	Boolean	参数解释: 是否开启流式调用，默认开启。 约束限制: 当前Agent只支持流式调用，需设置为true。 取值范围: <ul style="list-style-type: none">• true: 开启。• false: 不开启。 默认取值: 不涉及

表 3-265 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	参数解释: 用户问题，作为运行Agent的输入。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

响应参数

流式 (Header中的stream参数为true)

状态码: 200

表 3-266 流式输出的数据单元

参数	参数类型	描述
data	String	<p>参数解释:</p> <ul style="list-style-type: none">• stream=true时，执行Agent的消息以流式形式返回。• 生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-267 流式输出的数据单元

参数	参数类型	描述
event	String	<p>参数解释: 数据单元类型。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• start: 开始节点，表示开始调用模型进行会话。• message: 消息节点，表示模型返回的消息。• plugin_start: 插件调用请求节点，表示调用插件的请求信息。• plugin_end: 插件调用响应节点，表示调用插件的响应信息。• statistic_data: 执行数据节点，包含本次调用的耗时信息。• summary_response: 消息总结节点，包含本次调用的全量响应信息。• done: 流式调用结束节点，表示流式响应结束。 <p>默认取值: 不涉及</p>

参数	参数类型	描述
content	Object	<p>参数解释: 消息块内容，不同event的消息块内容不同。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
createdTime	long	<p>参数解释: 消息块返回的时间戳，如1733817348963。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
latency	Object	<p>参数解释: 耗时，包括以下三个元素：<ul style="list-style-type: none">• plugin: 插件调用耗时。• model: 模型调用耗时。• overall: 总耗时。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
plugin	Object	<p>参数解释: 插件请求信息，包括以下两个元素：<ul style="list-style-type: none">• name: 插件名。• arguments: 插件入参名。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

请求示例

流式 (Header中的stream参数为true)

```
POST https://[endpoint]/v1/{project_id}/agent-run/agents/{agent_id}/conversations/{conversation_id}
```

Request Header:

Content-Type: application/json

X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEA...
stream: true

Request Body:

```
{  
    "query": "查询A12会议室在9:00到10:00的状态"  
}
```

响应示例

```
data:{"event":"start","createdTime":1735558575017}  
  
data:{"event":"message","content":"好的","createdTime":1735558576300}  
  
data:{"event":"message","content":": ",","createdTime":1735558576301}  
  
data:{"event":"message","content":"我将","createdTime":1735558576301}  
  
data:{"event":"message","content":"调用","createdTime":1735558576302}  
  
data:{"event":"message","content":"query","createdTime":1735558576302}  
  
data:{"event":"message","content":":_","createdTime":1735558576302}  
  
data:{"event":"message","content":"meeting","createdTime":1735558576302}  
  
data:{"event":"message","content":":_","createdTime":1735558576302}  
  
data:{"event":"message","content":"room","createdTime":1735558576303}  
  
data:{"event":"message","content":":_status","createdTime":1735558576303}  
  
data:{"event":"message","content":"工具","createdTime":1735558576303}  
  
data:{"event":"message","content":"来","createdTime":1735558576304}  
  
data:{"event":"message","content":"查询","createdTime":1735558576304}  
  
data:{"event":"message","content":"A","createdTime":1735558576304}  
  
data:{"event":"message","content":"12","createdTime":1735558576304}  
  
data:{"event":"message","content":"会议室","createdTime":1735558576305}  
  
data:{"event":"message","content":"在","createdTime":1735558576305}  
  
data:{"event":"message","content":"9","createdTime":1735558576305}  
  
data:{"event":"message","content":":00","createdTime":1735558576305}  
  
data:{"event":"message","content":"到","createdTime":1735558576306}  
  
data:{"event":"message","content":"10","createdTime":1735558576306}  
  
data:{"event":"message","content":":00","createdTime":1735558576306}  
  
data:{"event":"message","content":"的状态","createdTime":1735558576306}
```

```
data:{"event":"message","content":"。","createdTime":1735558576306}
data:{"event":"message","content":"请","createdTime":1735558576307}
data:{"event":"message","content":"稍","createdTime":1735558576307}
data:{"event":"message","content":"等","createdTime":1735558576307}
data:{"event":"message","content":"。","createdTime":1735558576307}
data:{"event":"message","content":"_","createdTime":1735558576307}
data:{"event":"message","content":" query","createdTime":1735558576307}
data:{"event":"message","content":"_","createdTime":1735558576308}
data:{"event":"message","content":"meeting","createdTime":1735558576308}
data:{"event":"message","content":"_","createdTime":1735558576308}
data:{"event":"message","content":"room","createdTime":1735558576308}
data:{"event":"message","content":"_status","createdTime":1735558576308}
data:{"event":"message","content":"|","createdTime":1735558576308}
data:{"event":"message","content":"{}","createdTime":1735558576309}
data:{"event":"message","content":"meeting","createdTime":1735558576309}
data:{"event":"message","content":"Room","createdTime":1735558576309}
data:{"event":"message","content":"\:",createdTime":1735558576309}
data:{"event":"message","content":"{}","createdTime":1735558576309}
data:{"event":"message","content":"number","createdTime":1735558576310}
data:{"event":"message","content":"\:",createdTime":1735558576310}
data:{"event":"message","content":" 12","createdTime":1735558576310}
data:{"event":"message","content":"}","createdTime":1735558576310}
data:{"event":"message","content":"\","",createdTime":1735558576310}
data:{"event":"message","content":"start","createdTime":1735558576310}
data:{"event":"message","content":"\.\\"","createdTime":1735558576311}
data:{"event":"message","content":"9","createdTime":1735558576311}
data:{"event":"message","content":"00","createdTime":1735558576311}
data:{"event":"message","content":"\,\\"","createdTime":1735558576311}
data:{"event":"message","content":"end","createdTime":1735558576311}
data:{"event":"message","content":"\:\\"","createdTime":1735558576311}
data:{"event":"message","content":"10","createdTime":1735558576311}
data:{"event":"message","content":"00","createdTime":1735558576312}
data:{"event":"message","content":"\}","createdTime":1735558576312}
data:{"event":"message","content":"_","createdTime":1735558576312}
data:{"event":"plugin_start","type":"plugin","latency":{"overall":1.3}, "plugin":
```

```
{"name":"query_meeting_room_status","arguments":{"\"meetingRoom\": {\"number\": 12}, \"start\": \"9:00\", \"end\": \"10:00\"}","createdTime":1735558576316}

data:{"event":"plugin_end","content":{"result":"空闲"},"role":"function","latency": {"overall":1.51,"plugin":0.0}, "createdTime":1735558576521}

data:{"event":"start","createdTime":1735558576522}

data:{"event":"message","content":"A","createdTime":1735558576976}

data:{"event":"message","content":"12","createdTime":1735558576977}

data:{"event":"message","content":"会议室","createdTime":1735558576977}

data:{"event":"message","content":"在","createdTime":1735558576977}

data:{"event":"message","content":"9","createdTime":1735558576978}

data:{"event":"message","content":"00","createdTime":1735558576978}

data:{"event":"message","content":"到","createdTime":1735558576978}

data:{"event":"message","content":"10","createdTime":1735558576978}

data:{"event":"message","content":"00","createdTime":1735558576978}

data:{"event":"message","content":"的时间","createdTime":1735558576978}

data:{"event":"message","content":"段","createdTime":1735558576979}

data:{"event":"message","content":"内","createdTime":1735558576979}

data:{"event":"message","content":"是","createdTime":1735558576979}

data:{"event":"message","content":"空闲","createdTime":1735558576979}

data:{"event":"message","content":"的","createdTime":1735558576979}

data:{"event":"message","content":"。","createdTime":1735558576980}

data:{"event":"statistic_data","latency": {"overall":1.97}, "createdTime":1735558576986}

data:{"event":"summary_response","content":"A12会议室在9:00到10:00的时间段内是空闲的。","role":"assistant", "createdTime":1735558576987}

data:{"event":"done","createdTime":1735558577011}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.3.2 调用工作流

功能介绍

通过调用创建好的工作流API，输入问题，将得到工作流执行的结果。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/workflows/{workflow_id}/conversations/{conversation_id}

获取URI方式请参见[请求URI](#)。

表 3-268 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
workflow_id	是	String	参数解释: Workflow ID，获取方式如下： 在“Agent开发”页面，左侧导航栏选择“工作台 > 工作流”，在所需工作流中单击“... > 复制ID”。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
conversation_id	是	String	参数解释: 会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-269 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

参数	是否必选	参数类型	描述
stream	否	Boolean	参数解释: 是否开启流式调用。 <ul style="list-style-type: none">• true: 开启。• false: 不开启。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-270 请求 Body 参数

参数	是否必选	参数类型	描述
inputs	是	Map<String, Object>	参数解释: 用户提出的问题，作为运行工作流的输入，与工作流开始节点输入参数对应。包含默认字段query为用户输入。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
plugin_configs	否	List<PluginConfig>	参数解释: 插件配置，当工作流有配置用户自定义插件节点时，可能需要配置鉴权信息等，具体结构定义详见 表3-271 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-271 PluginConfig 参数

参数	是否必选	参数类型	描述
plugin_id	是	String	<p>参数解释: 插件Id，获取方式如下： 在“Agent开发”页面，左侧导航栏选择“工作台 > 插件”， 在所需插件中单击“... > 复制ID”。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及</p>
config	是	Map<String, String>	<p>参数解释: 插件配置信息。</p> <ul style="list-style-type: none"> 当工作流关联插件节点，并且插件是“用户级鉴权”时，需要在此配置对应的鉴权信息，例如针对如下插件，config可以配成：{"key2": "value"}。  <p>• 其他情况该参数无需传值，plugin_configs传空数组即可。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及</p>

响应参数

非流式 (Header中的stream参数为false)

状态码： 200

表 3-272 非流式输出的数据单元

参数	参数类型	描述
outputs	Map<String, Object>	<p>参数解释： 工作流最终输出内容，支持多个参数。</p> <p>说明</p> <p>outputs示例如下：</p> <ul style="list-style-type: none">• responseContent参数是默认有的，值取为工作流结束节点里的“指定回复”内容。• 支持用户在工作流结束节点的“输出参数”模块自定义配置参数，自定义配置参数将会显示在user_fields参数里。 <p>"outputs": {"user_fields": {"aaa": "1", "vvv": [{"role": "user", "content": "1"}]}, "responseContent": "你好！\ud83d\ude0a 你输入了“1”，请问有什么我可以帮助你的吗？如果有具体问题或需求，随时告诉我哦！"}</p>  <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>
messages	List<Message>	<p>参数解释： 工作流助手回复内容，如提问器节点问题消息，详见表3-273。</p> <p>约束限制： 不涉及</p> <p>取值范围： 不涉及</p> <p>默认取值： 不涉及</p>

参数	参数类型	描述
status	Map<String, Object>	参数解释: 状态，包含状态码code，状态描述desc。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
start_time	Long	参数解释: 开始时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
end_time	Long	参数解释: 结束时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-273 Message

参数	参数类型	描述
role	String	参数解释: 会话角色，支持user、assistant。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
content	String	参数解释: 会话内容。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

流式 (Header中的stream参数为true或不传)

状态码: 200

表 3-274 流式输出的数据单元

参数	参数类型	描述
data	String	参数解释: stream=true时，执行工作流的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-275 流式输出的数据单元

参数	参数类型	描述
event	String	<p>参数解释: 数据单元类型。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• workflow_started: 工作流开始事件, 表示工作流开始运行。• workflow_finished: 工作流结束事件, 表示工作流结束运行。• message: 消息事件, 表示工作流执行过程中流式返回的消息。• error: 错误事件, 表示工作流执行错误信息。• end: 结束事件, 标识请求结束。 <p>默认取值: 不涉及</p>
data	Object	<p>参数解释: 消息块内容, 不同event的消息块内容不同。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-276 workflow_started 事件的数据单元

参数	参数类型	描述
start_time	Long	<p>参数解释: 工作流开始时间。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

表 3-277 workflow_finished 事件的数据单元

参数	参数类型	描述
start_time	Long	参数解释: 工作流开始时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
end_time	Long	参数解释: 工作流结束时间。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
outputs	Map<String, Object>	参数解释: 工作流最终输出内容，支持多个参数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
status	Map<String, Object>	参数解释: 状态，包含状态码code，状态描述desc。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-278 message 事件的数据单元

参数	参数类型	描述
text	String	参数解释: 工作流输出内容消息块。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
index	Integer	参数解释: 消息块索引。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
node_id	String	参数解释: 节点id。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
node_type	String	参数解释: 支持输出message事件的节点类型。 约束限制: 不涉及 取值范围: Message: 消息节点。 End: 结束节点。 Questioner: 提问器节点。 Input: 输入节点。 默认取值: 不涉及

参数	参数类型	描述
node_name	String	参数解释: 节点名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

表 3-279 error 事件的数据单元

参数	参数类型	描述
code	String	参数解释: 工作流执行错误码。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
message	String	参数解释: 工作流执行错误消息。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
node_id	String	参数解释: 节点id。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	参数类型	描述
node_type	String	参数解释: 发送错误事件的节点类型。 约束限制: 不涉及 取值范围: Start: 开始节点。 End: 结束节点。 LLM: 大模型节点。 Workflow: 工作流节点。 Agent: Agent节点。 Branch: 判断节点。 IntentDetection: 意图识别节点。 Code: 代码节点。 Loop: 循环节点。 Plugin: 插件节点。 Mcp: MCP服务节点。 Message: 消息节点。 Questioner: 提问器节点。 Input: 输入节点。 SetVariable: 变量赋值节点。 Aggregation: 变量聚合节点。 KnowledgeRepo: 知识检索节点。 Unknown: 未知节点。 默认取值: 不涉及
node_name	String	参数解释: 节点名。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求示例

```
POST https://{{endpoint}}/v1/{{project_id}}/agent-run/workflows/{{workflow_id}}/conversations/{{conversation_id}}
```

```
Request Header:
```

```
Content-Type: application/json
```

```
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEA...  
stream: true  
Request Body:  
{  
    "inputs": {  
        "query": "你好"  
    },  
    "plugin_configs": [  
        {  
            "plugin_id": "xxxxxxxxxx",  
            "config": {  
                "key": "value"  
            }  
        }  
    ]  
}
```

响应示例

非流式 (Header 中的 stream 参数为 false)

输入节点返回:

```
{  
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",  
    "messages": [  
        {  
            "role": "assistant",  
            "content": "{\"inputs\": [{\"actualType\": \"string\", \"sourceType\": \"null\", \"description\": \"姓名\", \"name\": \"name\", \"type\": \"string\", \"required\": true}]}",  
            "nodeId": "node_1745928389632",  
            "nodeType": "Input",  
            "nodeName": "输入"  
        }  
  
,  
        "status": {  
            "code": 3,  
            "desc": "waiting"  
        },  
        "start_time": 1734336269313,  
        "end_time": 1734336270908  
    ]  
}
```

提问器节点返回:

```
{  
    "conversation_id": "f9a5540f-0c92-4f28-bd6e-f96ce04f5cc81",  
    "messages": [  
        {  
            "role": "assistant",  
            "content": "请您提供姓名, 年龄相关的信息",  
            "nodeId": "node_1745929628452",  
            "nodeType": "Questioner",  
            "nodeName": "提问器"  
        }  
  
,  
        "status": {  
            "code": 3,  
            "desc": "waiting"  
        },  
        "start_time": 1745929778250,  
        "end_time": 1745929779951  
    ]  
}
```

结束节点返回:

```
{  
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",  
    "outputs": {  
    }
```

```
        "responseContent": "你好！有什么我可以帮助你的吗？"
    },
    "messages": [],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734337068533,
    "end_time": 1734337082545
}
```

流式 (Header 中的 stream 参数为 true 或不传)

输入节点返回：

```
data:[{"event":"workflow_started","data":{"start_time":1745929087614}}]

data:[{"event":"message","data":{"text":"{\\"inputs\\": [{"actualType\\": \"string\\", \"sourceType\\": \"null\\",
\\\"description\\": \"姓名\\", \"name\\": \"name\\", \"type\\": \"string\\", \"required\\": true}]}","index":0,"node_id":"node_1745928389632","node_type":"Input","node_name":"输入"}}

data:[{"event":"message","data":
{"text":"","node_id":"node_1745928389632","node_type":"Input","node_name":"输入","is_finished":true}]

data:[{"event":"end"}]
```

提问器节点返回：

```
data:[{"event":"workflow_started","data":{"start_time":1745929709955}}]

data:[{"event":"message","data":{"text":"请您提供姓名, 年龄相关的信息
","index":0,"node_id":"node_1745929628452","node_type":"Questioner","node_name":"提问器"}}

data:[{"event":"message","data":
{"text":"","node_id":"node_1745929628452","node_type":"Questioner","node_name":"提问器
","is_finished":true}]

data:[{"event":"end"}]
```

结束节点返回：

```
data:[{"event":"workflow_started","data":{"start_time":1745929897770}}]

data:[{"event":"message","data":{"text":"","index":0,"node_id":"node_end","node_type":"End","node_name":"结束"}]

data:[{"event":"message","data":{"text":"你好
","index":1,"node_id":"node_end","node_type":"End","node_name":"结束"}]

data:[{"event":"message","data":{"text":"!
","index":2,"node_id":"node_end","node_type":"End","node_name":"结束"}]

data:[{"event":"message","data":{"text":"有什么我可以帮助你的吗？
","index":3,"node_id":"node_end","node_type":"End","node_name":"结束"}]

data:[{"event":"message","data":{"text":"","node_id":"node_end","node_type":"End","node_name":"结束
","is_finished":true}]

data:[{"event":"workflow_finished","data":{"status":{"code":1,"desc":"succeeded"},"outputs":
{"responseContent":"你好！有什么我可以帮助你的吗？
"},"start_time":1745929897770,"end_time":1745929898600}]

data:[{"event":"end"}]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.4 Token 计算器

功能介绍

为了帮助用户更好地管理和优化Token消耗，平台提供了Token计算器工具。Token计算器可以帮助用户在模型推理前评估文本的Token数量，提供费用预估，并优化数据预处理策略。

授权信息

账号具备所有API的调用权限，如果使用账号下的IAM用户调用当前API，该IAM用户需具备调用API所需的权限，具体权限要求请参见[权限和授权项](#)。

URI

POST /v1/{project_id}/deployments/{deployment_id}/caltokens

获取URI方式请参见[请求URI](#)。

表 3-280 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	参数解释: 项目ID，获取方法请参见 获取项目ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
deployment_id	是	String	参数解释: 模型的部署ID，获取方法请参见 获取模型部署ID 。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

请求参数

表 3-281 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	参数解释: 用户Token。 用于获取操作API的权限。如图2-5中响应消息头中X-Subject-Token的值即为Token。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及
Content-Type	是	String	参数解释: 发送的实体的MIME类型。 约束限制: 不涉及 取值范围: 不涉及 默认取值: application/json

表 3-282 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	List<String>	参数解释: 待统计Token数的字符串。List长度必须为奇数。 约束限制: 不涉及 取值范围: 不涉及 默认取值: 不涉及

参数	是否必选	参数类型	描述
with_prompt	否	Boolean	<p>参数解释: 是否仅统计输入字符的Token数。</p> <p>约束限制: 不涉及</p> <p>取值范围:</p> <ul style="list-style-type: none">• true: 仅统计输入字符串的Token数。• false: 统计输入字符串和推理过程产生字符的总Token数。 <p>默认取值: 不涉及</p>

响应参数

表 3-283 响应 Body 参数

参数	参数类型	描述
tokens	List<String>	<p>参数解释: 分解出的Token列表。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>
token_number	Integer	<p>参数解释: Token总数统计结果。</p> <p>约束限制: 不涉及</p> <p>取值范围: 不涉及</p> <p>默认取值: 不涉及</p>

请求示例

```
{  
    "data": [
```

```
        "你好，请介绍下西安。"
    ],
    "with_prompt": true
}
```

响应示例

```
{
    "tokens": [
        "你好",
        ",",
        "请",
        "介绍下",
        "西安",
        "。"
    ],
    "token_number": 6
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

4 附录

[状态码](#)
[错误码](#)
[获取项目ID](#)
[获取空间ID](#)
[获取模型部署ID](#)
[获取用户ID](#)
[获取负载均衡ID](#)

4.1 状态码

HTTP状态码为三位数，分成五个类别：1xx：信息响应；2xx：操作成功；3xx：重定向；4xx：客户端错误；5xx：服务器错误响应。

状态码如下所示。

状态码	编码	状态说明
100	Continue	继续请求。 这个临时响应用来通知客户端，它的部分请求已经被服务器接收，且仍未被拒绝。
101	Switching Protocols	切换协议。只能切换到更高级的协议。 例如，切换到HTTPS的新版本协议。
200	OK	服务器已成功处理了请求。
201	Created	创建类的请求完全成功。
202	Accepted	已经接受请求，但未处理完成。
203	Non-Authoritative Information	非授权信息，请求成功。

状态码	编码	状态说明
204	No Content	请求完全成功，同时HTTP响应不包含响应体。在响应OPTIONS方法的HTTP请求时返回此状态码。
205	Reset Content	重置内容，服务器处理成功。
206	Partial Content	服务器成功处理了部分GET请求。
300	Multiple Choices	多种选择。请求的资源可包括多个位置，相应可返回一个资源特征与地址的列表用于用户终端（例如：浏览器）选择。
301	Moved Permanently	永久移动，请求的资源已被永久的移动到新的URI，返回信息会包括新的URI。
302	Found	资源被临时移动。
303	See Other	查看其他地址，使用GET和POST请求查看。
304	Not Modified	所请求的资源未修改，服务器返回此状态码时，不会返回任何资源。
305	Use Proxy	所请求的资源必须通过代理访问。
306	Unused	已经被废弃的HTTP状态码。
400	Bad Request	非法请求。 建议直接修改该请求，不要重试该请求。
401	Unauthorized	在客户端提供认证信息后，返回该状态码，表明服务端指出客户端所提供的认证信息不正确或非法。
402	Payment Required	保留请求。
403	Forbidden	请求被拒绝访问。 返回该状态码，表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为该请求被设置为拒绝访问，建议直接修改该请求，不要重试该请求。
404	Not Found	所请求的资源不存在。 建议直接修改该请求，不要重试该请求。
405	Method Not Allowed	请求中带有该资源不支持的方法。 建议直接修改该请求，不要重试该请求。
406	Not Acceptable	服务器无法根据客户端请求的内容特性完成请求。
407	Proxy Authentication Required	请求要求代理的身份认证，与401类似，但请求者应当使用代理进行授权。

状态码	编码	状态说明
408	Request Timeout	服务器等待请求发生超时。 客户端可以随时再次提交该请求而无需进行任何更改。
409	Conflict	服务器在完成请求时发生冲突。 返回该状态码，表明客户端尝试创建的资源已经存在，或者由于冲突请求的更新操作不能被完成。
410	Gone	客户端请求的资源已经不存在。 返回该状态码，表明请求的资源已被永久删除。
411	Length Required	服务器无法处理客户端发送的不带Content-Length的请求信息。
412	Precondition Failed	未满足前提条件，服务器未满足请求者在请求中设置的其中一个前提条件。
413	Request Entity Too Large	由于请求的实体过大，服务器无法处理，因此拒绝请求。为防止客户端的连续请求，服务器可能会关闭连接。如果只是服务器暂时无法处理，则会包含一个Retry-After的响应信息。
414	Request URI Too Long	请求的URI过长（URI通常为网址），服务器无法处理。
415	Unsupported Media Type	服务器无法处理请求附带的媒体格式。
416	Requested Range Not Satisfiable	客户端请求的范围无效。
417	Expectation Failed	服务器无法满足Expect的请求头信息。
422	Unprocessable Entity	请求格式正确，但是由于含有语义错误，无法响应。
429	Too Many Requests	表明请求超出了客户端访问频率的限制或者服务端接收到多于它能处理的请求。建议客户端读取相应的Retry-After首部，然后等待该首部指出的时间后再重试。
500	Internal Server Error	表明服务端能被请求访问到，但是不能理解用户的请求。
501	Not Implemented	服务器不支持请求的功能，无法完成请求。
502	Bad Gateway	充当网关或代理的服务器，从远端服务器接收到了一个无效的请求。
503	Service Unavailable	被请求的服务无效。 建议直接修改该请求，不要重试该请求。

状态码	编码	状态说明
504	Gateway Timeout	请求在给定的时间内无法完成。客户端仅在为请求指定超时 (Timeout) 参数时会得到该响应。
505	HTTP Version Not Supported	服务器不支持请求的HTTPS协议的版本，无法完成处理。

4.2 错误码

当您调用API时，如果遇到“APIGW”开头的错误码，请参见[API网关错误码](#)进行处理。遇到“APIG”开头的错误码，请参考本文档进行处理。

表 4-1 错误码

模块	错误码	错误信息	说明	建议解决方法
模型推 理	PANGU.0010	parameter illegal.	请求参数错误。	请参考《API文档》输入正确的请求参数，并重新调试API。
	PANGU.0011	Authentication failed.	认证失败。	认证鉴权失败，请参考《API文档》 认证鉴权 章节重新进行认证。
	PANGU.0012	The authentication information is missing.	缺少身份验证信息。	请检查调用API时是否有传入认证鉴权信息。
	PANGU.0031	Inner service exception.	服务内部异常。	请联系服务技术支持协助解决。
	PANGU.3254	The requested inference service does not exist.	资源不存在。	请检查调用API时projectId和deploymentId是否填写正确，推理服务状态是否可用。
	PANGU.3267	The number of service invoking requests exceeds the project limit.	用户调用过于频繁。	请降低请求频率。
	PANGU.3278	required api parameter is not present.	请求参数丢失。	请检查调用API时请求参数是否填写完整、是否有拼写错误、取值是否正确。

模块	错误码	错误信息	说明	建议解决方法
	PANGU.3318	The total length of the question should be between 1 and 4096.	Content长度不合法。	请参考《API文档》检查请求参数中输入的Content参数长度是否不在范围内，并重新调试API。
	PANGU.3320	The parameter [n] can only be 1 or 2 when calling non-streaming.	非流式调用推理服务传的参数只能是1或者2。	请使用正确的取值：1或者2。
	PANGU.3321	The parameter [n] can only be 1 when calling streaming.	流式调用推理服务n只能取1。	请使用正确的取值：1。
	PANGU.3342	Failed to invoke the inference service. please check the details field.	调用推理服务失败，请查看错误详情。	调用推理服务失败，请查看错误详情。
	IIT.0201	The input param is invalid! /The input param is invalid, please check your key!	请求参数不合法。	请检查请求参数是否填写正确。
	IIT.0202	Interval Server Error!	内部错误。	请联系服务技术支持协助解决。
	IIT.0203	The input param is invalid, the input data lens is less than the train data lens!	请求参数不合法，输入参数中的数据长度小于训练所用数据长度。	请确认请求body中特征名称、特征数量是否与训练数据中的特征一致。
	PREDICT.0102	Json format is wrong! 或者其他与数据相关的特定报错信息	请求数据非JSON格式；或者其他与数据相关的特定错误。	请将请求体设置为JSON格式；或者根据数据相关的特定报错信息调整请求体。
	PREDICT.0201	The input param is invalid! /The input param is invalid, please check your key!	请求参数不合法。	请检查请求参数是否填写正确。

模块	错误码	错误信息	说明	建议解决方法
	PREDICT.0202	Interval Server Error!	内部错误。	请联系服务技术支持协助解决。
	PREDICT.0203	The input param is invalid, the input data lens is less than the train data lens!	请求参数不合法，输入参数中的数据长度小于训练所用数据长度。	请确认请求body中特征名称、特征数量是否与训练数据中的特征一致。
	APIG.0101	The API does not exist or has not been published in the environment.	访问的API不存在或尚未在环境中发布。	<ul style="list-style-type: none">请检查API的URL是否拼写正确，例如，URL中是否缺少project_id。HTTP请求方法（POST, GET等）是否正确。
	APIG.0201	Backend timeout.	请求超时。	<ul style="list-style-type: none">请检查原调用请求是否过于频繁，如果是并发过大，可以通过重试机制解决，在代码里检查返回值，碰到这个并发错误可以延时一小段时间（如2-5s）重试请求；也可以后端检查上一个请求结果，上一个请求返回之后再发送下一个请求，避免请求过于频繁。请与技术支持确认，API是否已完成部署。

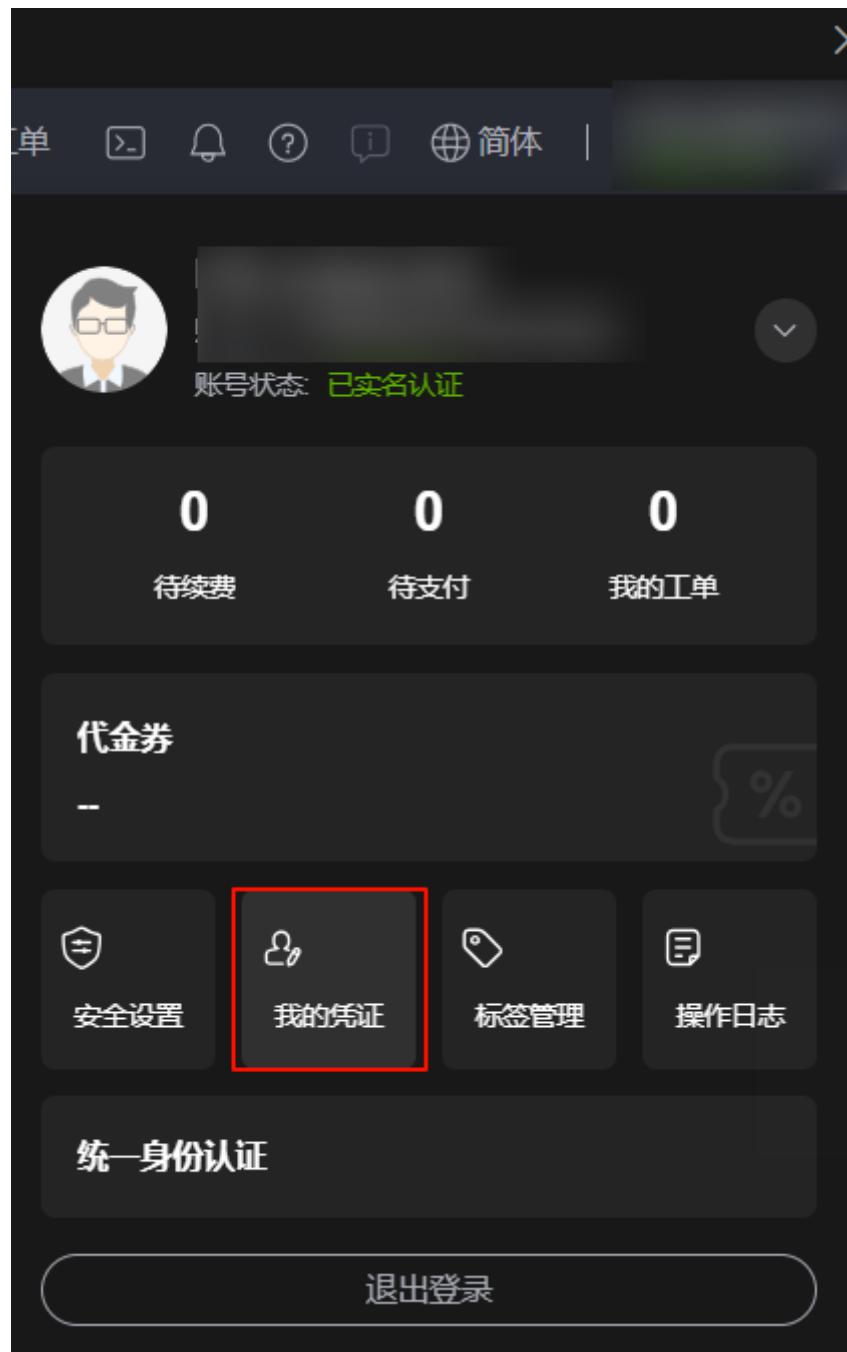
模块	错误码	错误信息	说明	建议解决方法
	APIG.0301	Incorrect IAM authentication information.	IAM身份验证信息不正确： <ul style="list-style-type: none">decrypt token fail: token解析失败。token expires: token过期。verify aksk signature fail: AK/SK认证失败。x-auth-token not found: 未找到x-auth-token参数。	<ul style="list-style-type: none">token解析失败，请检查获取token的方法，请求体信息是否填写正确，token是否正确；检查获取token的环境与调用的环境是否一致。token超时（token expires），请重新获取token，使用不过期的token。请检查AK/SK是否正确（AK对应的SK错误，不匹配；AK/SK中多填了空格）。AK/SK频繁出现鉴权出错，连续错误5次以上，被锁定5分钟（5分钟内，则一直认为其是异常的鉴权请求），5分钟后解锁重新认证。检查账号权限，是否欠费，被冻结等。检查调用API时，请求header参数X-Auth-Token是否拼写正确。
	APIG.0308	The throttling threshold has been reached: policy user over ratelimit,limit:XX ,time:1 minute.	发送请求超过了服务的默认配置限流。	<ul style="list-style-type: none">通过重试机制，在代码里检查返回值，碰到并发错误可以延时一小段时间（如2-5s）重试请求。后端检查上一个请求结果，上一个请求返回之后再发送下一个请求，避免请求过于频繁。

4.3 获取项目 ID

从控制台获取项目 ID

1. 登录[管理控制台](#)。
2. 在页面右上角的用户名的下拉列表中选择“我的凭证”。

图 4-1 我的凭证



3. 在“我的凭证”页面，获取项目ID（project_id），以及账号名、账号ID、IAM用户名和IAM用户ID。

在调用盘古API时，获取的项目id需要与盘古服务部署区域一致，例如盘古大模型当前部署在“西南-贵阳一”区域，需要获取与贵阳一区域的对应的项目id。

图 4-2 查看项目 ID



多项目时，展开“所属区域”，从“项目ID”列获取子项目ID。

调用 API 获取项目 ID

项目ID还可通过调用[查询指定条件下的项目信息](#)API获取。

获取项目ID的接口为“`GET https://[endpoint]/v3/projects`”，其中`[endpoint]`为IAM的终端节点，可以从[地区和终端节点](#)获取。接口的认证鉴权请参见[认证鉴权](#)。

响应示例如下，例如，对话机器人服务部署的区域为“`cn-north-4`”，响应消息体中查找“`name`”为“`cn-north-4`”，其中`projects`下的“`id`”即为项目ID。

```
{  
    "projects": [  
        {  
            "domain_id": "65382450e8f64ac0870cd180d14e684b",  
            "is_domain": false,  
            "parent_id": "65382450e8f64ac0870cd180d14e684b",  
            "name": "project_name",  
            "description": "",  
            "links": {  
                "next": null,  
                "previous": null,  
                "self": "https://www.example.com/v3/projects/a4a5d4098fb4474fa22cd05f897d6b99"  
            },  
            "id": "a4a5d4098fb4474fa22cd05f897d6b99",  
            "enabled": true  
        }  
    ],  
    "links": {  
        "next": null,  
        "previous": null,  
        "self": "https://www.example.com/v3/projects"  
    }  
}
```

4.4 获取空间 ID

1. 使用最终租户登录ModelArts Studio平台，进入所需空间。
2. 在左侧“空间资产”导航栏中选择资产页签，例如选择“数据”页签。
3. 在所选择的页面中按F12打开开发者工具并进入“Network”页签，单击资产列表的刷新按钮。在开发者工具的“Network”中找到assets接口，在接口的响应体（Response）中查询空间ID。

图 4-3 获取空间 ID



4.5 获取模型部署 ID

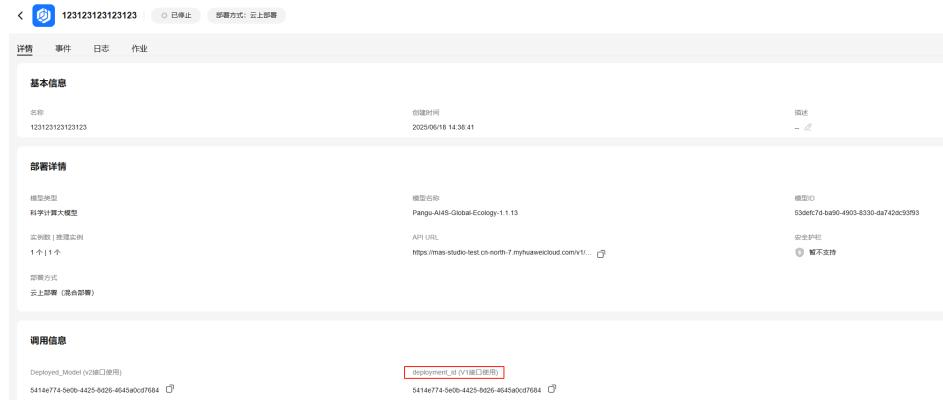
模型部署ID获取步骤如下：

步骤1 登录ModelArts Studio大模型开发平台。

步骤2 获取模型部署ID。

- 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的部署ID。

图 4-4 部署后的模型调用路径



- 若调用预置模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“预置服务”页签，模型列表单击“调用路径”，获取该模型的部署ID。

图 4-5 预置模型的部署 ID



----结束

4.6 获取用户 ID

在调用接口的时候，部分URL中需要填入用户ID，需要先在管理控制台上获取。用户ID获取步骤如下：

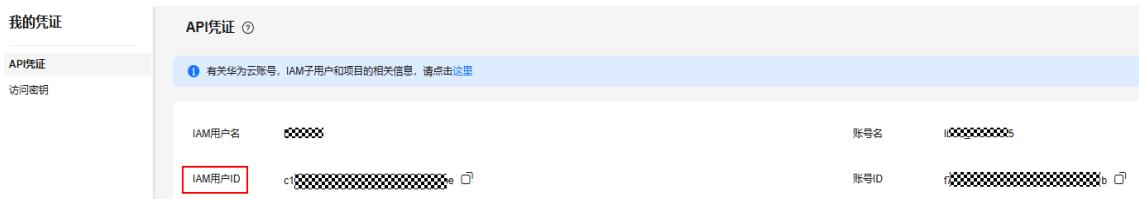
1. 登录[管理控制台](#)。
2. 在页面右上角的用户名的下拉列表中选择“我的凭证”。

图 4-6 我的凭证



3. 在“我的凭证”页面，获取IAM用户ID，即为用户ID。

图 4-7 查看项目 ID



4.7 获取负载均衡 ID

步骤1 登录ModelArts Studio大模型开发平台，进入首页界面。

步骤2 点击右上角“设置”菜单，进入资源池管理界面，选择“边缘资源池”菜单，选择对应的资源池下拉找到其负载均衡信息。

图 4-8 设置

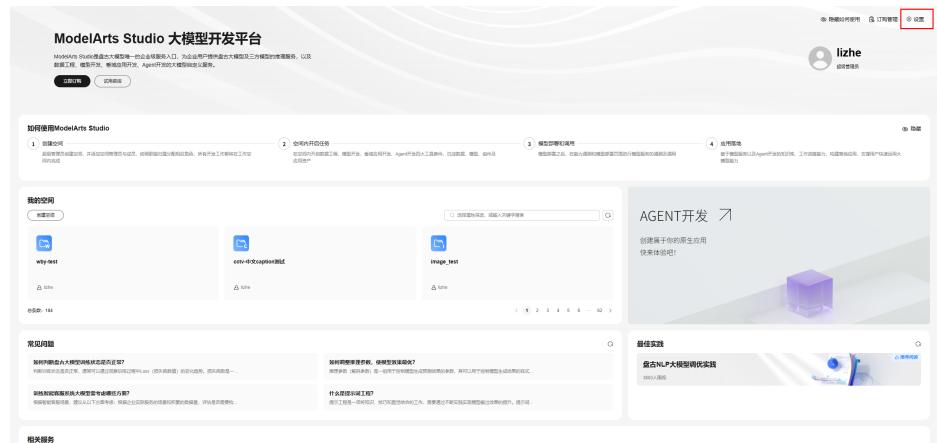
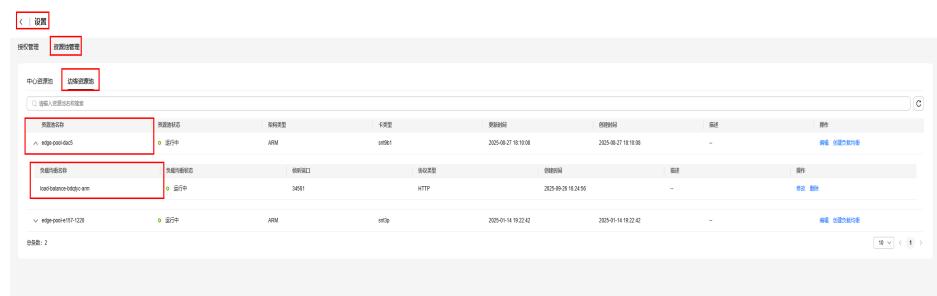


图 4-9 边缘资源池



步骤3 如果负载均衡不存在，则需要单击资源池右边的“创建负载均衡”按钮进行创建（输入名称，及监听端口，端口号必须在30000-40000之间）。

图 4-10 创建负载均衡入口

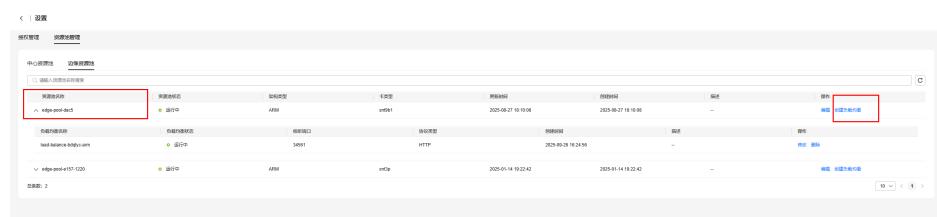


图 4-11 创建负载均衡



步骤4 创建成功后，执行[步骤2](#)查看负载均衡信息。

----结束