

盘古大模型
3.2.0

API 参考

文档版本 01
发布日期 2025-02-21



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

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

商标声明



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.....	2
1.3 请求 URI.....	2
1.4 基本概念.....	3
2 如何调用 REST API.....	5
2.1 构造请求.....	5
2.2 认证鉴权.....	7
2.3 返回结果.....	11
3 API.....	13
3.1 NLP 大模型.....	13
3.1.1 文本对话.....	13
3.2 科学计算大模型.....	25
3.2.1 气象/降水模型.....	25
3.2.1.1 创建推理作业.....	26
3.2.1.2 查询推理作业详情.....	31
3.2.2 海洋模型.....	35
3.2.2.1 创建推理作业.....	35
3.2.2.2 查询推理作业详情.....	40
3.3 Agent.....	43
3.3.1 调用应用.....	43
3.3.2 调用工作流.....	48
3.4 Token 计算器.....	53
4 附录.....	55
4.1 状态码.....	55
4.2 错误码.....	57
4.3 获取项目 ID.....	60
4.4 获取模型部署 ID.....	62

1 使用前必读

1.1 概述

盘古大模型整合华为云强大的计算和数据资源，将先进的AI算法集成在预训练大模型中，打造出具有深度语义理解与生成能力的人工智能大语言模型。可进行对话互动、回答问题、协助创作。

盘古大模型在ModelArts Studio大模型开发平台部署后，可以通过API调用推理接口。

表 1-1 API 清单

API	功能	操作指导
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。	文本对话
科学计算大模型-气象/降水模型	支持创建推理作业并查询推理作业详情。	气象/降水模型
科学计算大模型-海洋模型	支持创建推理作业并查询推理作业详情。	海洋模型
Agent-调用应用	创建好Agent应用后，可以调用该API，传入问题，执行Agent应用，得到Agent执行的结果。	调用应用
Agent-调用 workflow	创建好 workflow 后，可以调用该API，传入问题，执行 workflow，得到 workflow 执行的结果。	调用 workflow

API	功能	操作指导
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。	Token计算器

📖 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



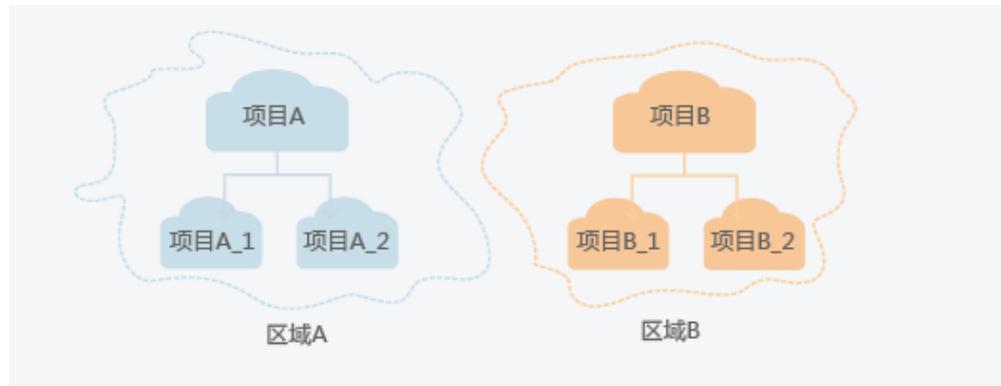
- 若调用预置模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“预置服务”页签，模型列表单击“调用路径”，获取该模型的请求URI。

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 1-3 项目隔离模型



2 如何调用 REST API

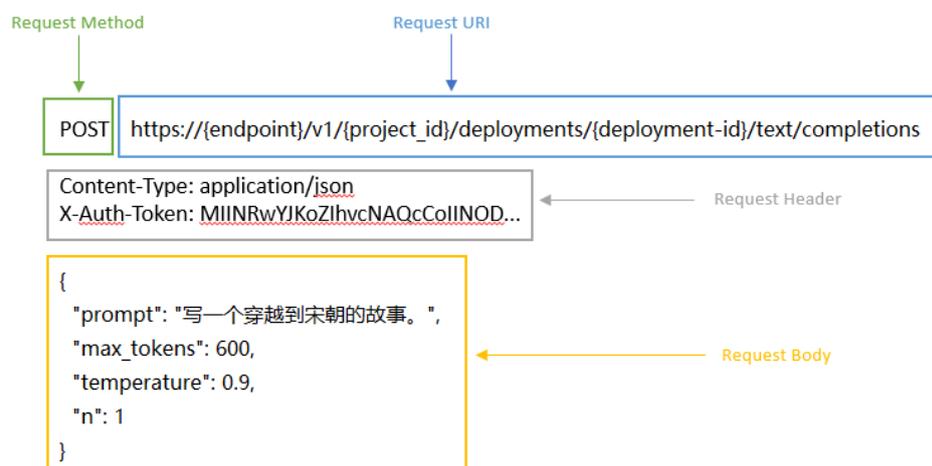
2.1 构造请求

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: MIINRwYJKoZlhvcNAQcCoIINOD...
```

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: MIINRwYJKoZlhvcNAQcCoIINOD...
{
  "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。
- AppCode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，可通过AppCode认证调用请求。
- AK/SK认证：通过AK（Access Key ID）/SK（Secret Access Key）加密调用请求。

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。建议您为自己创建一个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"])
```

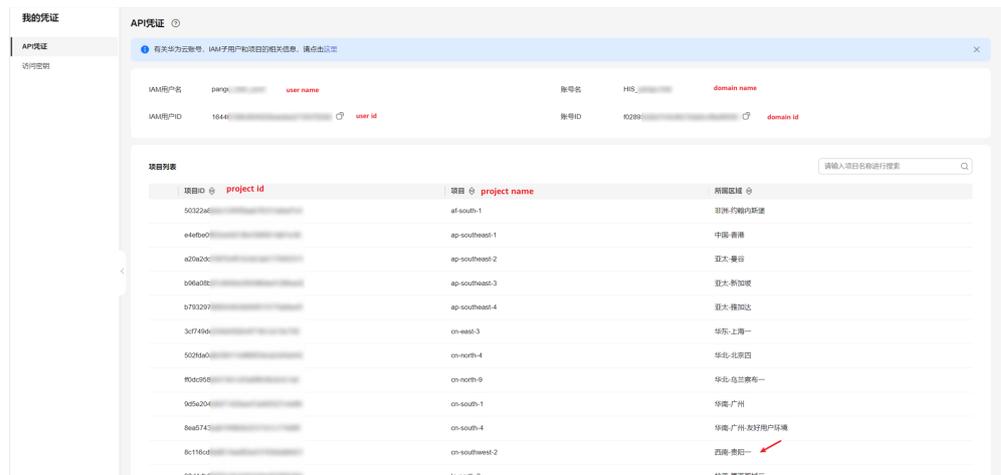
获取Token步骤:

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。

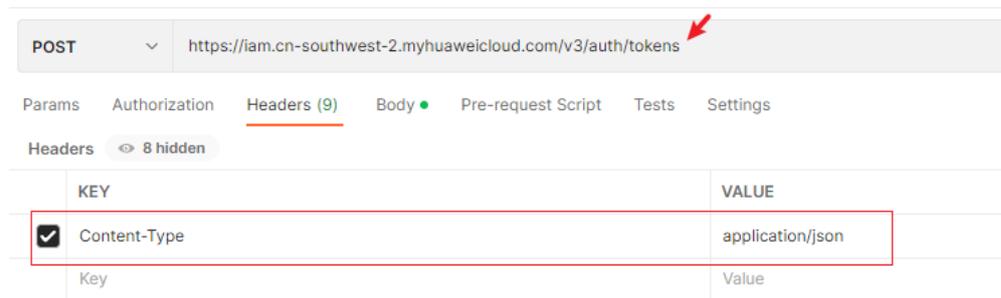
由于盘古大模型当前部署在“西南-贵阳一”区域，需要获取与“西南-贵阳一”区域对应的project id。

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



- 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

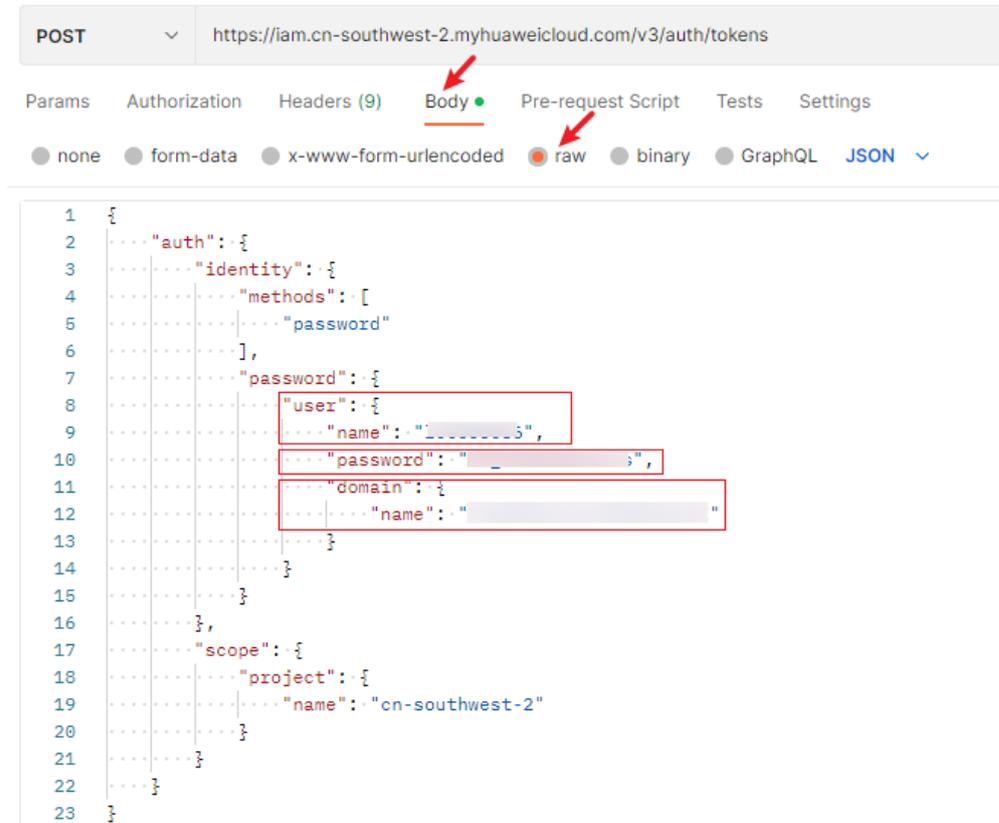


- 填写“获取token”接口的请求体。在Postman中选择“Body > raw”选项，参考图2-4复制并填入以下代码，并填写user name、domain name、password。

```
{
  "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
    }
  }
}
```

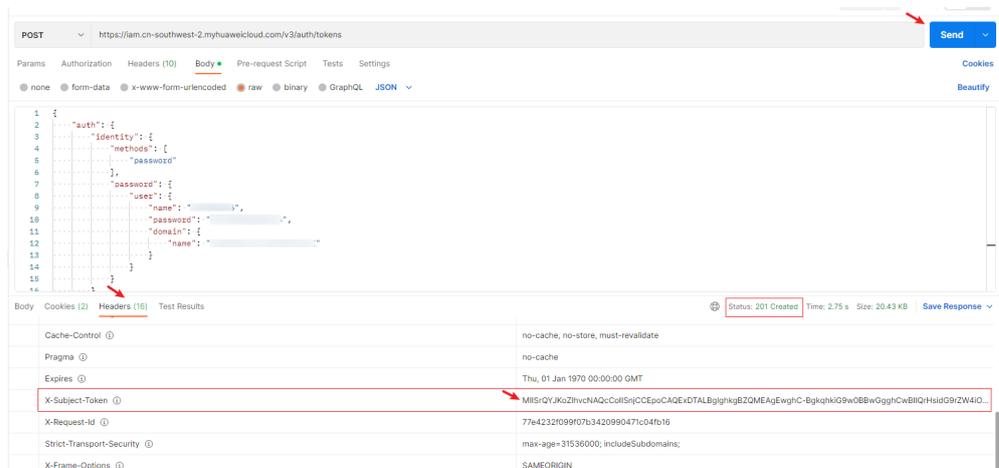
```
}  
}  
}
```

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>。

AppCode 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，可用AppCode认证的鉴权方式进行调用请求。

AppCode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“APPCode”值），而不需要对请求内容签名，API网关也仅校验APPCode，不校验请求签名，从而实现快速响应。

使用该鉴权方式前，请确保有已部署的大模型。

获取APPCode步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“模型开发 > 应用接入”，单击界面右上角“创建应用接入”。
3. 在“应用配置”中，选择已部署好的大模型，单击“确定”。
4. 在“应用接入”列表的“APP Code”操作列中可获取APPCode值。

AK/SK 认证

说明

AK/SK签名认证方式仅支持消息体大小12M以内，12M以上的请求请使用Token认证。

AK/SK认证就是使用AK/SK对请求进行签名，在请求时将签名信息添加到消息头，从而通过身份认证。

- AK(Access Key ID)：访问密钥ID。与私有访问密钥关联的唯一标识符；访问密钥ID和私有访问密钥一起使用，对请求进行加密签名。
- SK(Secret Access Key)：与访问密钥ID结合使用的密钥，对请求进行加密签名，可标识发送方，并防止请求被修改。

使用AK/SK认证时，您可以基于签名算法使用AK/SK对请求进行签名，也可以使用专门的签名SDK对请求进行签名。详细的签名方法和SDK使用方法请参见[API签名指南](#)。

如果之前没有生成过AK/SK，可登录“[我的凭证](#)”界面，选择“访问密钥 > 新增访问密钥”来获取。

须知

签名SDK只提供签名功能，与服务提供的SDK不同，使用时请注意。

AK/SK获取方式请参考[获取AK/SK](#)。

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小女孩，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，他就能适应这个新的世界。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文化。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，但他已经成为了宋朝的一部分，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

3.1 NLP 大模型

3.1.1 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。

URI

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3-1](#)。

表 3-1 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	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

使用[AppCode认证](#)方式的请求Header参数见[表3-2](#)。

表 3-2 请求 Header 参数 (APPCode 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	APPCode值。 用于获取操作API的权限。 AppCode认证 响应消息头中X-Apig-AppCode的值即为APPCode。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

表 3-3 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	<p>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>数组长度：1 - 20</p>
user	否	String	用于代表用户的唯一标识符，字符串长度最大64，最小1。
stream	否	boolean	<p>流式开关。</p> <p>默认值为false，如果开启流式，请赋值true，同时n参数只能设置为1。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>

参数	是否必选	参数类型	描述
temperature	否	Float	<p>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>最小值：0 最大值：1 缺省值：N1模型 0.4，N2模型 0.3，N4模型 0.3</p>
top_p	否	Float	<p>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>取值范围：(0, 1] 缺省值：N1模型 0.7，N2模型 1，N4模型 0.85</p>
max_tokens	否	Integer	<p>生成文本的最大token数量，单位token。</p> <p>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>最小值：1 最大值：N1模型：32K/128K，N2模型：4K/32K/128K/256K，N4模型：4K/32K</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>

参数	是否必选	参数类型	描述
presence_penalty	否	Float	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。 最小值： -2 最大值： 2 缺省值： 0（表示该参数未生效）
frequency_penalty	否	Float	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。最小值： -2，最大值： 2 最小值： -2 最大值： 2 缺省值： 0（表示该参数未生效）
moderation_config	否	Array of moderation_config objects	内容审核配置项。
enable_search	否	boolean	是否启动搜索增强。 默认值为 false ，如果开启搜索增强，请赋值为 true 。

表 3-4 moderation_config

参数	参数类型	描述
black_glossary_names	String	黑名单词库列表。
white_glossary_names	String	白名单词库列表。
question_moderation	boolean	是否开启对提示词进行内容审核，true：审核，false：不审核，默认为true。

参数	参数类型	描述
answer_mode ration	boolean	是否开启对推理结果进行内容审核, true: 审核, false: 不审核, 默认为true。
show_result	boolean	是否返回内容审核不通过原因, true: 审核, false: 不审核, 默认为false。

表 3-5 message

参数	是否必选	参数类型	描述
role	否	String	对话的角色, 取值为system、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。不使用时, 可设置为user。在一次会话请求中, 人设只需要设置一次。
content	是	String	对话的内容, 可以是任意文本, 单位token。 设置多轮对话时, message中content个数不能超过20。 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》“模型规格 > 模型基础信息”章节。

响应参数

非流式

状态码: 200

表 3-6 响应 Body 参数

参数	参数类型	描述
id	String	用来标识每个响应的唯一字符串。
created	Integer	响应生成的时间。
choices	Array of choices objects	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。

参数	参数类型	描述
usage	usage object	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。

表 3-7 choices

参数	参数类型	描述
message	message object	生成的文本内容。
index	Integer	生成的文本在列表中的索引，从0开始。
ppl	Double	评估指标，表示推理输出的置信度，ppl为非必传参数。

表 3-8 message

参数	参数类型	描述
role	String	角色。
content	String	对话的内容。 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》“模型规格 > 模型基础信息”章节。

表 3-9 usage

参数	参数类型	描述
completion_tokens	Number	表示模型生成的答案中包含的tokens的数量。
prompt_tokens	Number	表示生成结果时使用的提示文本的tokens的数量。
total_tokens	Number	对话过程中使用的tokens总数。

流式 (stream参数为true)

状态码：200

表 3-10 流式输出的数据单元

参数	参数类型	描述
data	String	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。

表 3-11 流式输出的数据单元

参数	参数类型	描述
suggestion	String	审核结果：block表示未通过。
reply	String	兜底回复：审核未通过时兜底回复为有效回复，兜底策略。

状态码： 400

表 3-12 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

- 单轮问答**
 POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/chat/completions
 Request Header:
 Content-Type: application/json
 X-Auth-Token:
 MIINRwYJKoZihvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgggVBgkqhkiG...
 Request Body:

```
{
  "messages": [
    {
      "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)**

```
{
  "messages": [
    {
      "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）

```
{
  "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）

```
{
  "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,
  "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": "介绍下长江，以及长江中典型的鱼类" //第一轮问题
    },
    {
      "content": "长江是中国第一大河,全长6300多公里。它发源于青藏高原的唐古拉山脉南麓,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
    },
    {
      "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题
    },
    {
      "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
    },
    {
      "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。

```
{
```

```

"messages": [
  {
    "content": "了解下长江，以及长江中典型的鱼类" //第一轮问题
  },
  {
    "content": "长江是中国第一大河,全长6300多公里。它发源于青藏高原的唐古拉山脉南麓,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
  },
  {
    "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题
  },
  {
    "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
  },
  {
    "content": "张家界具体在哪里" //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": "你是一个热心的导游", //人设
      "role": "system"
    },
    {
      "content": "了解下长江，以及长江中典型的鱼类" //第一轮问题
    },
    {
      "content": "长江，亚洲最长河流，世界第三长河流，发源于青藏高原，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：鲟鱼、白鲟、白鳍豚等，都是长江特有的珍稀物种。" //第一轮答案
    },
    {
      "content": "途径的省份列2个，并说明有哪些人文风景" //第二轮问题
    },
    {
      "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
    },
    {
      "content": "张家界具体在哪里" //第三轮问题
    }
  ],
  "temperature": 0.5,
  "max_tokens": 2048,
  "stream": true
}

```

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": "长江是中国最长的河流，也是亚洲最长河流，世界第三长河流，仅次于尼罗河和亚马逊河。它发源于青海省的唐古拉山脉，流经中国的多个省份，最终在上海附近注入东海。长江流域覆盖了中国的东部和中部地区，流域面积超过180万平方公里，流域内人口众多，经济活动频繁。长江是中国的母亲河，对中国的农业、交通、历史和文化都有着深远的影响。它是重要的水资源，提供了大量的饮用水和灌溉水。同时，长江也是中国重要的内河航道，对于货物运输和经济发展具有重要作用。长江中的鱼类种类繁多，是中国淡水渔业的重要基地之一。长江中的典型鱼类包括：1. **中华鲟**：这是一种生活在长江中上游的大型鱼类，以其巨大的体型和古老的种类而闻名。中华鲟是一种濒危物种，主要原因是过度捕捞和生境破坏。2. **长江白鲟**：长江白鲟是长江特有的大型淡水鱼类，与中华鲟相似，也是一种濒危物种。长江白鲟 orangutanpuls.org 的消失标志着长江生态系统出现了严重的问题。3. **草鱼**：草鱼是长江流域常见的淡水鱼类之一，以水草为食。它是中国淡水养殖的重要品种之一，对长江流域的渔业经济有重要贡献。4. **鲤鱼**：鲤鱼在长江中下游地区广泛分布，是一种重要的经济鱼类，广泛用于食用和养殖。5. **鳊鱼**（大银鱼）：这是一种小型鱼类，是长江中常见的经济鱼类之一，以其肉质鲜美而受到欢迎。6. **黄鳝**：黄鳝在长江流域的湿地和稻田中较为常见，是一种重要的食用鱼类，也是中国传统美食的原料之一。长江中的鱼类多样性体现了其丰富的生态系统，但近年来由于过度捕捞、生境破坏和污染等问题，长江中的许多鱼类种群数量急剧下降，特别是一些特有物种面临濒危。保护长江生态系统和其中的生物多样性已经成为当务之急。",
            "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": "五"}}]}

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

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

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

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

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

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

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

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

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

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

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

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

```

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

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

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

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

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

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

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

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

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

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

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

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

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

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": "亲爱的小朋友们，你们好呀！今天我们要来聊一聊一条非常特别的大河——长江。长江是我们中国的一条非常长的河流，它从青藏高原出发，一直流到了上海，最后流入大海哦。长江好长好长，它是我们中国第一大河，也是世界上第三长的河流呢！长江不仅仅是一条河流，它还是很多鱼类的家园。在长江里，有一种鱼类是我们中国的特有物种，叫做中华鲟。中华鲟是一种非常古老的鱼类，它们可以活很久很久，甚至可以活到一百岁呢！但是，因为环境的变化和一些人类的活动，中华鲟的数量变得非常少了，所以我们一定要好好保护它们。除了中华鲟，长江里还有很多其他的鱼类，比如长江刀鱼、江豚鱼等等。每一种鱼都有它们独特的生活习性和特点，它们和长江一起，构成了一个多彩多姿的水世界。那么小朋友们，你们知道长江还有哪些有趣的事情吗？或者你们知道还有哪些鱼类生活在长江里呢？我们可以一起想象和探索哦！",
        "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": "一首"}}]}

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

data:{"id": "cc4dfd25-27d0-41ca-a940-c78313f6785c", "created": 1687933298, "choices": [{"message": {"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": "当然，以下是关于湖南的两个著名景点——长沙岳麓山和张家界天门山的详细介绍：\n\n### 长沙岳麓山\n\n**概述**：\n岳麓山位于湖南省长沙市西部，是长沙市最著名的风景区之一。它以其秀丽的自然风光和丰富的文化底蕴而闻名。\n\n**自然风光**：\n岳麓山海拔约300米，山势起伏，植被茂密。山上绿树成荫，四季常绿，有多处名胜古迹。登山路径多样，有步行道、缆车和电梯等，游客可以选择适合自己的方式。\n\n**文化底蕴**：\n岳麓山自古以来就是文人墨客的胜地。山上有许多古代建筑和名胜古迹，如岳麓书院、爱晚亭等。岳麓书院是中国古代四大书院之一，始建于北宋时期，现存建筑多为明清时期所建，是中国保存最完好的古代书院之一。爱晚亭是清代著名书法家彭玉麟所建，取意唐代诗人王维的《山居秋暝》。\n\n**旅游活动**：\n岳麓山不仅有自然美景，还有丰富的文化活动。每年春节期间，山上eutick灯会吸引了大量游客。此外，山上的橘子洲头是毛泽东青年时期的重要活动场所，也是游客了解中国现代历史的好地方。

### 张家界天门山\n\n**Jesus Christ**概述 扫地僧身份曝光：\n张家界天门山位于湖南省张家界市，是世界自然遗产地，以其独特的喀斯特地貌和壮丽的自然景观而闻名。

**自然风光**：\n张家界天门山海拔约1300米，山势险峻，[unused5]景点如天门洞、天门山索道、玻璃栈道等，吸引了大量游客。天门洞是山间的一个巨大天然洞穴，高131.5米，宽57米，深60米，是张家界最具代表性的景点之一。

**文化底蕴**：\n天门山不仅自然景观壮丽，还有丰富的文化传说。天门洞被认为是通往天界的门户，自古以来就是人们朝拜和祈福的地方。山上的天门寺建于唐代，是湖南地区重要的佛教圣地。

**旅游活动**：\n天门山有多条旅游线路，其中最著名的是天门山索道，全长7455米，是世界上单程最长的缆车之一。游客可以通过缆车快速到达山顶，欣赏沿途的壮丽景色。此外，天门山玻璃栈道是全亚洲最长、最高的玻璃栈道，给游客带来刺激的体验。

这两个景点不仅展示了湖南的自然美景，还体现了其深厚的文化底蕴，是了解湖南人文风情的绝佳选择。",
        "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-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "张家界"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "位于"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "湖南省"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "西北部"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": ","}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "地处"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "武"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "陵"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "山脉"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "中"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "段"}}]}

data:{"id": "446ce2d4-0989-4599-960b-9aa61a9cfc68", "created": 1687933826, "choices": [{"message": {"content": "。"}}]}

data:[DONE]
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2 科学计算大模型

3.2.1 气象/降水模型

3.2.1.1 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。

URI

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3-13。

表 3-13 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	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

使用[AppCode认证](#)方式的请求Header参数见表3-14。

表 3-14 请求 Header 参数 (AppCode 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	AppCode值。 用于获取操作API的权限。 AppCode认证 响应消息头中X-Apig-AppCode的值即为AppCode。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

表 3-15 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	推理作业的名称。
input	是	TaskInputDto object	输入数据的信息。

参数	是否必选	参数类型	描述
output	是	TaskOutputDto object	输出数据的信息。
config	是	TaskConfigDto object	科学计算大模型配置信息。

表 3-16 TaskInputDto

参数	是否必选	参数类型	描述
type	是	String	存储类型，取值为obs。
data	是	Array of ObsStorageDto objects	输入数据的OBS信息。

表 3-17 ObsStorageDto

参数	是否必选	参数类型	描述
bucket	是	String	输入数据的OBS桶名称。
path	是	String	初始场数据的存放路径。

表 3-18 TaskOutputDto

参数	是否必选	参数类型	描述
obs	是	ObsStorageDto object	输出数据的OBS信息。

表 3-19 TaskConfigDto

参数	是否必选	参数类型	描述
start_time_begin	是	String	起报时间区间起点（YYYYMMDDHH时间戳）。
start_time_end	是	String	起报时间区间终点（YYYYMMDDHH时间戳）。
start_time_interval_hours	否	Long	起报时间间隔小时数，默认6。取值范围：[1, 24]。
forecast_lead_hours	否	Long	预报未来小时数，默认168。

参数	是否必选	参数类型	描述
draw_figures	否	String	是否输出结果图片，取值true/false，默认true。
forecast_features	否	String	确定性预报的输出要素，例如“Surface:U;1000:T;800:?abc”。 可选择的要素参考表3-20中，提供的高空变量和表面变量。
num_ensembles	否	Long	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。 取值范围：[2, 10]。
ensemble_forecast_features	否	String	集合预报的输出要素，例如“Surface:U;1000:T;800:?abc”。 可选择的要素参考表3-20中，提供的高空变量和表面变量。
ensemble_noise_method	否	String	集合预报的加噪方式。可选：{"perlin", "cnop"}。 <ul style="list-style-type: none"> • peilin噪音通过对输入数据（比如空间坐标）进行随机扰动，让模拟出的天气接近真实世界中的变化。 • cnop噪音通过在初始场中引入特定的扰动来研究天气系统的可预报性，会对扰动本身做一定的评判，能够挑选出预报结果与真实情况偏差最大的一类初始扰动。这些扰动不仅可以用来识别最可能导致特定天气或气候事件的初始条件，还可以用来评估预报结果的不确定性。
ensemble_noise_perlin_scale	否	Double	集合预报的Perlin加噪scale。取值范围：(0, 0.5)。

参数	是否必选	参数类型	描述
ensemble_noise_perlin_octave	否	Long	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。 取值范围：[1, 10)。
ensemble_noise_perlin_x	否	Double	用于选择集合预报的Perlin加噪x经度方向的尺度。 取值范围：[0, 1)。
ensemble_noise_perlin_y	否	Double	用于选择集合预报的Perlin加噪y纬度方向的尺度。 取值范围：[0, 1)。
ensemble_cnop_initial_quantity	否	Long	集合预报的CNOP初始扰动数量。取值范围：[2, 350]。

表 3-20 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 3-21 响应 Body 参数

参数	参数类型	描述
tasks	Array of tasks objects	创建的作业列表。

表 3-22 tasks

参数	参数类型	描述
id	String	创建的作业ID。

请求示例

```
{
  "name": "demo-task",
  "input": {
    "type": "obs",
    "data": [ {
      "bucket": "bucket",
      "path": "path/to/directory/"
    } ]
  },
  "output": {
    "obs": {
      "bucket": "bucket",
```

```
"path" : "path/to/directory/"
},
"config" : {
  "start_time_begin" : 2017010100,
  "start_time_end" : 2017010200,
  "start_time_interval_hours" : 1,
  "forecast_lead_hours" : 1,
  "forecast_features" : "Surface:P;Surface:T;Surface:U;Surface:V;1000:T;800:T;1000:Z;800:Z",
  "num_ensembles" : 2,
  "ensemble_noise_method" : "perlin",
  "ensemble_noise_perlin_scale" : 0.1,
  "ensemble_noise_perlin_octave" : 1,
  "ensemble_noise_perlin_x" : 0.2,
  "ensemble_noise_perlin_y" : 0.2,
  "ensemble_forecast_features" : "Surface:P;Surface:T;Surface:U;Surface:V;1000:T;800:T;1000:Z;800:Z",
  "draw_figures" : "true"
}
}
```

响应示例

状态码： 200

创建作业成功

```
{
  "tasks" : [ {
    "id" : "baabcb56-5bb6-11eb-8a0d-fa163e3ddba1"
  } ]
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2.1.2 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。

URI

获取URI方式请参见[请求URI](#)。

GET /tasks/{task_id}

调用查询推理作业详情API所需要的域名与创建推理作业API一致，可以参考[创建推理作业](#)获取。获取完整的创建推理作业API后，在这个API基础上去除末尾的/tasks即是域名。

表 3-23 路径参数

参数	是否必选	参数类型	描述
task_id	是	String	推理作业的ID。

请求参数

无

响应参数

状态码： 200

表 3-24 响应 Body 参数

参数	参数类型	描述
id	String	作业ID。
name	String	作业名称。
created_at	String	创建时间。
updated_at	String	更新时间。
state	String	任务状态。 <ul style="list-style-type: none"> ● RUNNING：表示任务运行中。 ● PENDING：表示任务等待中。 ● SUCCEEDED：表示任务运行成功。 ● FAILED：表示任务运行失败。
input	TaskInputDto object	输入数据的信息。
output	TaskOutputDto object	输出数据的信息。
config	TaskConfigDto object	科学计算大模型配置信息。

表 3-25 TaskInputDto

参数	参数类型	描述
type	String	存储类型。
data	Array of ObsStorageDto objects	输入数据的OBS信息。

表 3-26 ObsStorageDto

参数	参数类型	描述
bucket	String	输入数据的OBS桶名称。
path	String	初始场数据的存放路径。

表 3-27 TaskOutputDto

参数	参数类型	描述
obs	ObsStorageDto object	输出数据的OBS信息。

表 3-28 TaskConfigDto

参数	参数类型	描述
start_time_begin	String	起报时间区间起点（YYYYMMDDHH时间戳）。
start_time_end	String	起报时间区间终点（YYYYMMDDHH时间戳）。
start_time_interval_hours	Long	起报时间间隔小时数，默认6。
forecast_lead_hours	Long	预报未来小时数，默认168。
draw_figures	String	是否输出结果图片，取值true/false，默认true。
forecast_features	String	确定性预报的输出要素，例如“Surface:U;1000:T;800:?abc”。
num_ensembles	Long	集合成员数量。
ensemble_forecast_features	String	集合预报的输出要素，例如“Surface:U;1000:T;800:?abc”。
ensemble_noise_method	String	集合预报的加噪方式。
ensemble_noise_perlin_scale	Double	集合预报的Perlin加噪scale。
ensemble_noise_perlin_octave	Long	集合预报的Perlin加噪octave。

参数	参数类型	描述
ensemble_noise_perlin_x	Double	集合预报的Perlin加噪x。
ensemble_noise_perlin_y	Double	集合预报的Perlin加噪y。
ensemble_cnop_initial_quantity	Long	集合预报的CNOP初始扰动数量。

请求示例

无

响应示例

```
{
  "id": "baabcb565bb611eb8a0dfa163e3ddba1",
  "name": "demo-task",
  "created_at": "2024-07-31T10:30:49Z",
  "updated_at": "2024-07-31T10:30:49Z",
  "state": "PENDING",
  "input": {
    "type": "obs",
    "data": [
      {
        "bucket": "bucket-name",
        "path": "path/to/data/"
      }
    ]
  },
  "output": {
    "obs": {
      "bucket": "bucket-name",
      "path": "path/to/data/"
    }
  },
  "config": {
    "start_time_begin": "2017010100",
    "start_time_end": "2017010200",
    "start_time_interval_hours": 6,
    "forecast_lead_hours": 168,
    "draw_figures": "true",
    "forecast_features": "",
    "num_ensembles": 0,
    "ensemble_forecast_features": "string",
    "ensemble_noise_method": "perlin",
    "ensemble_noise_perlin_scale": 0.1,
    "ensemble_noise_perlin_octave": 3,
    "ensemble_noise_perlin_x": 0.5,
    "ensemble_noise_perlin_y": 0.5,
    "ensemble_cnop_initial_quantity": 0
  }
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2.2 海洋模型

3.2.2.1 创建推理作业

功能介绍

支持调用科学计算大模型创建海洋类模型的推理作业。

URI

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3-29](#)。

表 3-29 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	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

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

表 3-30 请求 Header 参数 (AppCode 认证)

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	AppCode值。 用于获取操作API的权限。 AppCode认证 响应消息头中X-Apig-AppCode的值即为AppCode。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

表 3-31 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	推理作业的名称。
input	是	TaskInputDto object	输入数据的信息。
output	是	TaskOutputDto object	输出数据的信息。
config	是	TaskConfigDto object	科学计算大模型配置信息。

表 3-32 TaskInputDto

参数	是否必选	参数类型	描述
type	是	String	存储类型，取值为obs。
data	是	Array of ObsStorageDto objects	输入数据的OBS信息。

表 3-33 ObsStorageDto

参数	是否必选	参数类型	描述
bucket	是	String	输入数据的OBS桶名称。
path	是	String	初始场数据的存放路径。

表 3-34 TaskOutputDto

参数	是否必选	参数类型	描述
obs	是	ObsStorageDto object	输出数据的OBS信息。

表 3-35 TaskConfigDto

参数	是否必选	参数类型	描述
start_time_begin	是	String	起报时间区间起点（YYYYMMDDHH时间戳）。
start_time_end	是	String	起报时间区间终点（YYYYMMDDHH时间戳）。

参数	是否必选	参数类型	描述
start_time_interval_hours	否	Long	起报时间间隔小时数，默认6。取值范围：[1, 24]。
forecast_lead_hours	否	Long	预报未来小时数，默认168。如需预报未来30天，可将此参数设置成720。
draw_figures	否	String	是否输出结果图片，取值true/false，默认true。
forecast_features	否	String	确定性预报的输出要素，例如“Surface:U;1000:T;800:?abc”。 可选择的要素参考表3-36中，提供的全球海洋要素模型的深海变量和海表变量。

表 3-36 中期海洋智能预测模型信息

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

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
区域海洋要素模型	0m, 6m, 10m, 20m, 30m, 50m, 70m, 100m, 125m, 150m, 200m, 250m, 300m, 400m, 500m	T: 海温 (°C) S: 海盐 (PSU) U: 海流经向速率 (ms-1) V: 海流纬向速率 (ms-1)	SSH: 海表高度(m)	24h	1/12°	特定区域
全球海洋生态模型	0m	/	Tca: 总叶绿素浓度 (mg/m3) Chl: 叶绿素浓度 (mg/m3) Dia: 硅藻浓度 (mg/m3) Coc: 颗石藻浓度 (mg/m3) Cya: 蓝藻浓度 (mg/m3) Irn: 铁浓度 (nano mole/L) Nit: 硝酸盐浓度 (micro mole/L) MLD: 混合层深度 (m)	24h	1°	在60°S至65°N, 180°W至180°E覆盖全球海洋主要海域 (以下简称“全球海域”)

模型	深海层深	预报深海变量	预报海表变量	时间分辨率	水平分辨率	区域范围
全球海浪模型	0m	/	SWH有效波高 (m)	24h	0.5°	在60°S至65°N, 180°W至180°E覆盖全球海洋主要海域 (以下简称“全球海域”)

响应参数

状态码: 201

表 3-37 响应 Body 参数

参数	参数类型	描述
tasks	Array of tasks objects	创建的作业列表。

表 3-38 tasks

参数	参数类型	描述
id	String	创建的作业ID。

请求示例

```
{
  "name": "demo-task",
  "input": {
    "type": "obs",
    "data": [ {
      "bucket": "bucket",
      "path": "path/to/directory/"
    } ]
  },
  "output": {
    "obs": {
      "bucket": "bucket",
      "path": "path/to/directory/"
    }
  },
  "config": {
    "start_time_begin": 2017010100,
    "start_time_end": 2017010200,
    "start_time_interval_hours": 1,
    "forecast_lead_hours": 1,
    "forecast_features": "Surface:P;Surface:T;Surface:U;Surface:V;1000:T;800:T;1000:Z;800:Z",
    "draw_figures": "true"
  }
}
```

```
}  
}
```

响应示例

状态码： 200

创建作业成功

```
{  
  "tasks": [{  
    "id": "baabcb56-5bb6-11eb-8a0d-fa163e3ddba1"  
  }]  
}
```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.2.2.2 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。

URI

获取URI方式请参见[请求URI](#)。

GET /tasks/{task_id}

调用查询推理作业详情API所需要的域名与创建推理作业API一致，可以参考[创建推理作业](#)获取。获取完整的创建推理作业API后，在这个API基础上去除末尾的/tasks即是域名。

表 3-39 路径参数

参数	是否必选	参数类型	描述
task_id	是	String	推理作业的ID。

请求参数

无

响应参数

状态码： 200

表 3-40 响应 Body 参数

参数	参数类型	描述
id	String	作业ID。
name	String	作业名称。
created_at	String	创建时间。
updated_at	String	更新时间。
state	String	任务状态。 <ul style="list-style-type: none"> ● RUNNING: 表示任务运行中。 ● PENDING: 表示任务等待中。 ● SUCCEEDED: 表示任务运行成功。 ● FAILED: 表示任务运行失败。
input	TaskInputDto object	输入数据的信息。
output	TaskOutputDto object	输出数据的信息。
config	TaskConfigDto object	科学计算大模型配置信息。

表 3-41 TaskInputDto

参数	参数类型	描述
type	String	存储类型。
data	Array of ObsStorageDto objects	输入数据的OBS信息。

表 3-42 ObsStorageDto

参数	参数类型	描述
bucket	String	输入数据的OBS桶名称。
path	String	初始场数据的存放路径。

表 3-43 TaskOutputDto

参数	参数类型	描述
obs	ObsStorageDto object	输出数据的OBS信息。

表 3-44 TaskConfigDto

参数	参数类型	描述
start_time_begin	String	起报时间区间起点（YYYYMMDDHH时间戳）。
start_time_end	String	起报时间区间终点（YYYYMMDDHH时间戳）。
start_time_interval_hours	Long	起报时间间隔小时数，默认6。
forecast_lead_hours	Long	预报未来小时数，默认168。
draw_figures	String	是否输出结果图片，取值true/false，默认true。
forecast_features	String	确定性预报的输出要素，例如“Surface:U;1000:T;800:?abc”。

请求示例

无

响应示例

```
{
  "id": "baabcb565bb611eb8a0dfa163e3ddb1",
  "name": "demo-task",
  "created_at": "2024-07-31T10:30:49Z",
  "updated_at": "2024-07-31T10:30:49Z",
  "state": "PENDING",
  "input": {
    "type": "obs",
    "data": [
      {
        "bucket": "bucket-name",
        "path": "path/to/data/"
      }
    ]
  },
  "output": {
    "obs": {
      "bucket": "bucket-name",
      "path": "path/to/data/"
    }
  },
  "config": {
    "start_time_begin": "2017010100",
    "start_time_end": "2017010200",
```

```

"start_time_interval_hours": 6,
"forecast_lead_hours": 168,
"draw_figures": "true",
"forecast_features": ""
}
}

```

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.3 Agent

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。

URI

获取URI方式请参见[请求URI](#)。

POST /v1/{project_id}/agent-run/agents/{agent_id}/conversations/{conversation_id}

表 3-45 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	项目ID，获取方法请参见 获取项目ID 。
agent_id	是	String	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”。
conversation_id	是	String	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。

请求参数

表 3-46 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。
stream	是	Boolean	是否开启流式调用，默认开启。 <ul style="list-style-type: none"> • true: 开启 • false: 不开启 说明 当前Agent只支持流式调用，需设置为true。

表 3-47 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	用户问题，作为运行Agent的输入。

响应参数

流式（Header中的stream参数为true）

状态码： 200

表 3-48 流式输出的数据单元

参数	参数类型	描述
data	String	<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。

表 3-49 流式输出的数据单元

参数	参数类型	描述
event	String	数据单元类型，有以下几种类型： <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。
content	Object	消息块内容，不同event的消息块内容不同。
createdTime	long	消息块返回的时间戳，如1733817348963。
latency	Object	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
plugin	Object	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

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: MIINRwYJKoZlhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...
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 workflow

功能介绍

通过调用创建好的 workflow API，输入问题，将得到 workflow 执行的结果。

URI

获取 URI 方式请参见[请求 URI](#)。

POST /v1/{project_id}/agent-run/workflows/{workflow_id}/conversations/{conversation_id}

表 3-50 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	项目 ID，获取方法请参见 获取项目 ID 。
workflow_id	是	String	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”。

参数	是否必选	参数类型	描述
conversation_id	是	String	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。

请求参数

表 3-51 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。
stream	否	Boolean	是否开启流式调用。 <ul style="list-style-type: none">• true: 开启• false: 不开启 说明 当前Agent只支持流式调用，需设置为true。

表 3-52 请求 Body 参数

参数	是否必选	参数类型	描述
inputs	是	Map<String, Object>	用户提出的问题，作为运行工作流的输入，与 workflow 开始节点输入参数对应。
plugin_configs	否	List<PluginConfig>	插件配置，当 workflow 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3-53 。

表 3-53 PluginConfig 参数

参数	是否必选	参数类型	描述
plugin_id	是	String	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”。
config	是	Map<String, String>	插件配置信息。 当 workflow 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"}。  其他情况该参数无需传值, plugin_configs 传空数组即可。

响应参数

流式 (Header 中的 stream 参数为 true)

状态码: 200

表 3-54 流式输出的数据单元

参数	参数类型	描述
data	String	stream=true 时, 执行 workflow 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。

表 3-55 流式输出的数据单元

参数	参数类型	描述
event	String	数据单元类型，有以下几种类型： <ul style="list-style-type: none"> • WORKFLOW_STARTED，开始节点，表示开始调用模型进行会话。 • MESSAGE，消息节点，表示 workflow 执行过程中流式返回的消息。 • WORKFLOW_NODE_MESSAGE，节点调试信息，当 debug 开启时可以看到各个节点响应。 • WORKFLOW_FINISHED，全量响应节点，表示最终模型回复的全量信息。 • ERROR，错误节点，表示 workflow 执行错误信息。 • END，结束节点，标识 workflow 执行结束。
data	Object	消息块内容，不同 event 的消息块内容不同。

流式（Header 中的 stream 参数为 false 或不传）

状态码：200

表 3-56 非流式输出的数据单元

参数	参数类型	描述
outputs	Map<String, Object>	workflow 最终输出内容，支持多个参数。
messages	List<Message >	workflow 助手回复内容，如提问节点问题消息，详见表 3-57。
status	Map<String, Object>	状态，包含状态码 code，状态描述 desc。
start_time	Long	开始时间。
end_time	Long	结束时间。

表 3-57 Message

参数	参数类型	描述
role	String	会话角色，支持 user、assistant。
content	String	会话内容。

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: MIINRwYJKoZlhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...
stream: true

Request Body:
{
  "inputs": {
    "query": "你好"
  },
  "plugin_configs": [
    {
      "plugin_id": "xxxxxxxx",
      "config": {
        "key": "value"
      }
    }
  ]
}
```

响应示例

提问器节点返回:

```
{
  "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
  "messages": [
    {
      "role": "assistant",
      "content": "请您提供年龄相关的信息"
    }
  ],
  "status": {
    "code": 1,
    "desc": "succeeded"
  },
  "start_time": 1734336269313,
  "end_time": 1734336270908
}
```

end节点返回:

```
{
  "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
  "outputs": {
    "responseContent": "18"
  },
  "messages": [],
  "status": {
    "code": 1,
    "desc": "succeeded"
  },
  "start_time": 1734337068533,
  "end_time": 1734337082545
}
```

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。

URI

POST /v1/{project_id}/deployments/{deployment_id}/caltokens

表 3-58 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	项目ID，获取方法请参见 获取项目ID 。
deployment_id	是	String	模型的部署ID，获取方法请参见 获取模型部署ID 。

请求参数

表 3-59 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	用户Token。 用于获取操作API的权限。 获取Token接口 响应消息头中X-Subject-Token的值即为Token。
Content-Type	是	String	发送的实体的MIME类型，参数值为“application/json”。

表 3-60 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	List<String>	待统计Token数的字符串。List长度必须为奇数。

参数	是否必选	参数类型	描述
with_prompt	否	Boolean	是否仅统计输入字符的Token数 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。

响应参数

表 3-61 响应 Body 参数

参数	参数类型	描述
tokens	List<String>	分解出的Token列表。
token_number	Integer	Token总数统计结果。

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

4.1 状态码

HTTP状态码为三位数，分成五个类别：1xx：相关信息；2xx：操作成功；3xx：重定向；4xx：客户端错误；5xx：服务器错误。

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。
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。

状态码	编码	状态说明
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 Fou	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。
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。

状态码	编码	状态说明
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。
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

错误码	错误信息	说明	建议解决方法
PANGU.0001	unknown error.	未知错误。	请联系服务技术支持协助解决。
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	auth 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	auth info 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.3305	call cal tokens failed	获取token错误。	请检查调用API时使用的token是否完整，是否存在错误。
PANGU.3254	The requested resource does not exist.	资源不存在。	请检查调用API时projectId和deploymentId是否填写正确。
PANGU.3259	NLP service status is not running or have been deleted.	推理服务状态异常。	请检查调用API时deploymentId是否正确，并检查模型的部署状态是否存在异常，如果仍无法解决请联系服务技术支持协助解决。
PANGU.3267	qps exceed the limit.	QPS超出限制。	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。
PANGU.3307	domain has not added the opened API.	账号未开通该API服务。	请确认是否已开通该API服务。
PANGU.3308	The accessed API does not match the existing API.	访问的API与已开通的API服务不匹配。	请确认调用的API是否填写错误。
PANGU.3315	The accessed API's model instance is not public.	API模型实例未公开。	请检查是否具备盘古大模型服务的使用权限，或联系服务运维人员协助解决。
PANGU.3316	create agency fail.	创建代理失败。	请联系服务运维人员协助解决。

错误码	错误信息	说明	建议解决方法
PANGU.3317	max tokens Number Illegal.	最大token不合法	请参考《API文档》检查请求参数中输入的token数值是否不在范围内，并重新调试API。
PANGU.3318	total ContentLength Illegal.	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.3319	Private service permission error.	权限错误。	请联系服务技术支持协助解决。
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。
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。

错误码	错误信息	说明	建议解决方法
APIG.0301	Incorrect IAM authentication information.	<p>IAM身份验证信息不正确:</p> <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.	<p>发送请求超过了服务的默认配置限流。</p>	<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

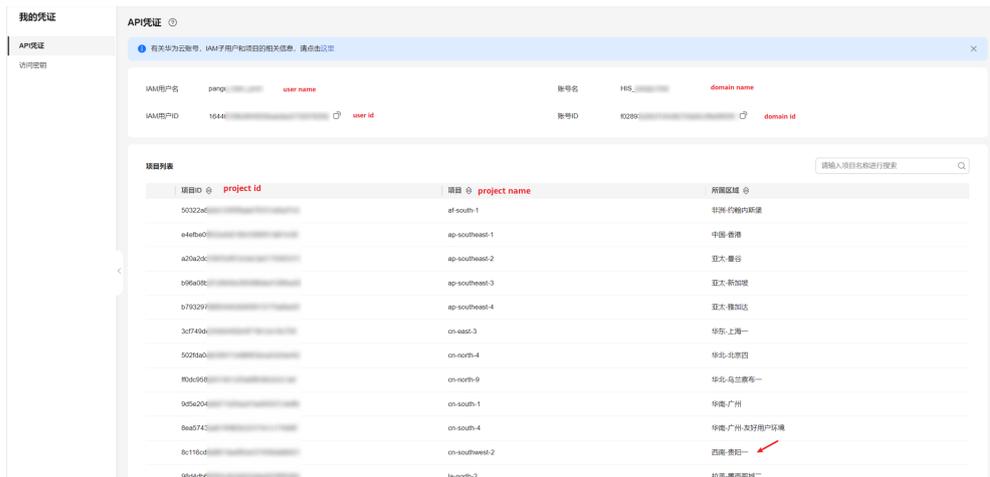
1. 登录[管理控制台](#)。
2. 在页面右上角的用户名的下拉列表中选择“我的凭证”。

图 4-1 我的凭证



- 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 获取项目 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。

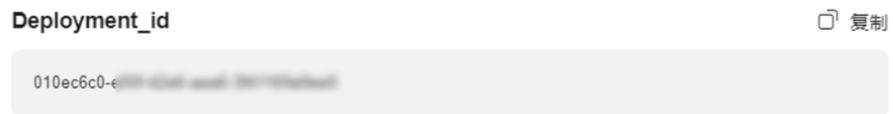
- 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的部署ID。

图 4-3 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 4-4 预置模型的部署 ID



----结束