

盘古大模型

API 参考

文档版本 01
发布日期 2024-08-31



版权所有 © 华为云计算技术有限公司 2024。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.....	1
1.3 终端节点.....	1
1.4 基本概念.....	3
2 如何调用 REST API.....	4
2.1 开通 API.....	4
2.2 构造请求.....	5
2.3 认证鉴权.....	8
2.4 返回结果.....	12
3 API.....	14
3.1 NLP 大模型.....	14
3.1.1 文本补全.....	14
3.1.2 多轮对话.....	21
4 附录.....	34
4.1 状态码.....	34
4.2 错误码.....	36
4.3 获取项目 ID.....	39
4.4 获取模型调用 API 地址.....	41

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。

华为云盘古大模型，以下功能支持API调用。

表 1-1 API 清单

API	功能
NLP-文本补全	给定一个提示和一些参数，模型会根据这些信息生成一个或多个预测的补全。它可以用来做文本生成、自动写作、代码补全等任务。
NLP-多轮对话	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。

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 终端节点

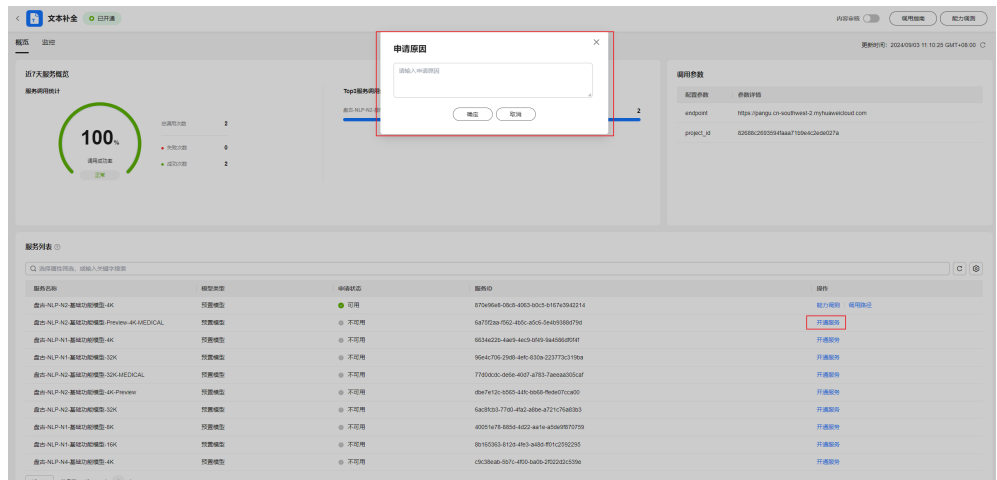
终端节点 (endpoint) 即API服务的终端地址，通过该地址与API进行通信和交互。获取步骤如下：

1. 登录[盘古大模型套件平台](#)。
2. 在左侧导航栏中选择“服务管理”，在相应服务的操作列单击“查看详情”，可在服务列表中申请需要开通的服务。

图 1-1 服务管理



图 1-2 申请开通服务



- 在“概览 > 服务列表”中选择需要调用的模型，并单击操作列的“调用路径”，在弹窗中可获取对应模型的API请求地址。

图 1-3 服务概览页面

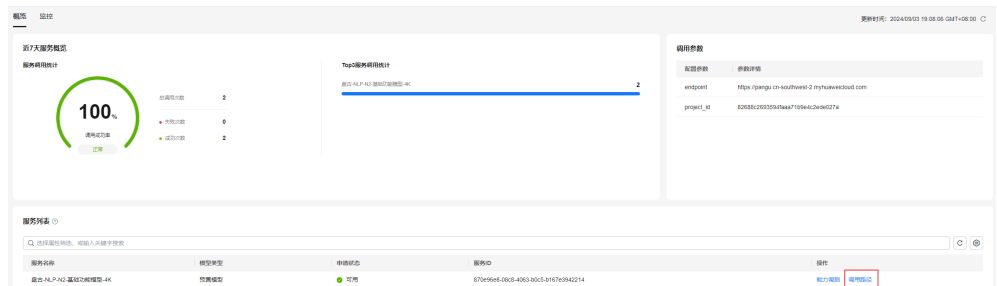
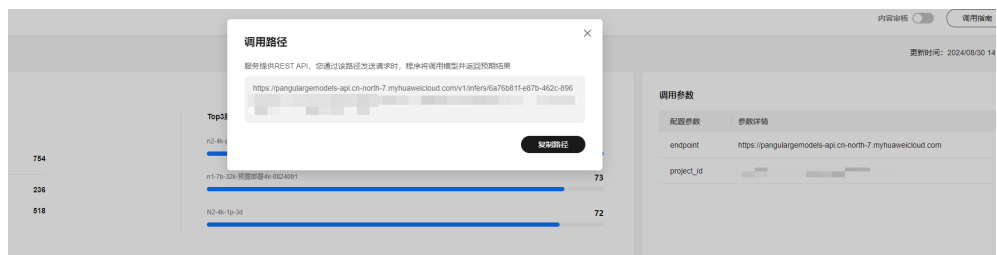


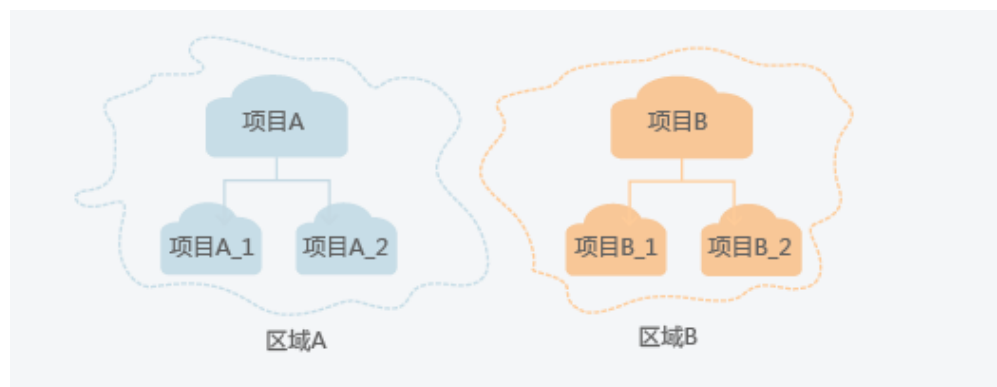
图 1-4 获取 API 请求地址



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

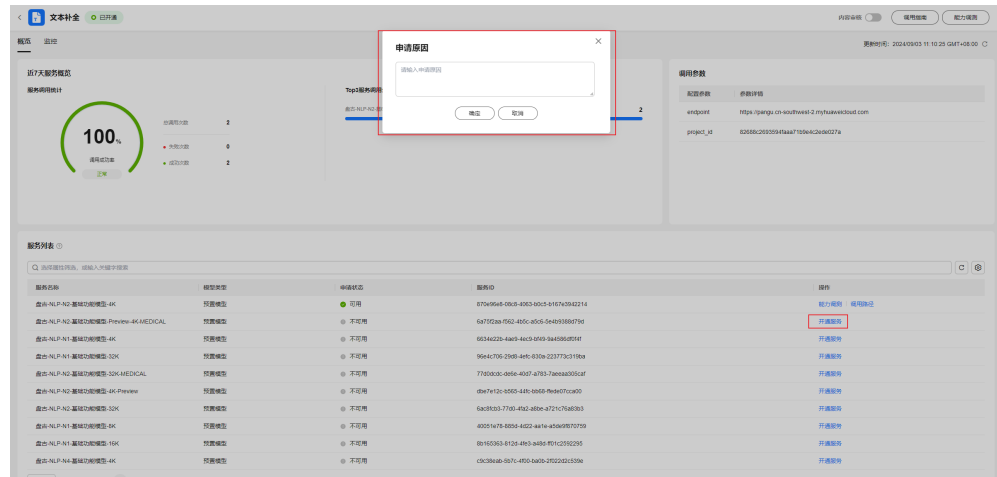
2.1 开通 API

1. 登录[盘古大模型套件平台](#)。
2. 在左侧导航栏中选择“服务管理”，在相应服务的操作列单击“查看详情”，可在服务列表中申请需要开通的服务。
 - 文本补全：给定一个提示和一些参数，模型会根据这些信息生成一个或多个预测的补全。例如让模型依据要求写邮件、做摘要总结、生成观点见解等。
 - 多轮对话：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。

图 2-1 服务管理



图 2-2 申请开通服务



3. 您可按照需要选择是否开启内容审核。
开启内容审核后，可以有效拦截大模型输入输出的有害信息，保障模型调用安全，推荐进行开启。

图 2-3 大模型内容审核



说明

盘古大模型支持通过对接内容审核，实现拦截大模型输入、输出的有害信息，保障模型调用安全。用户在开通API后，可依据需求选择是否开通、启用内容审核。

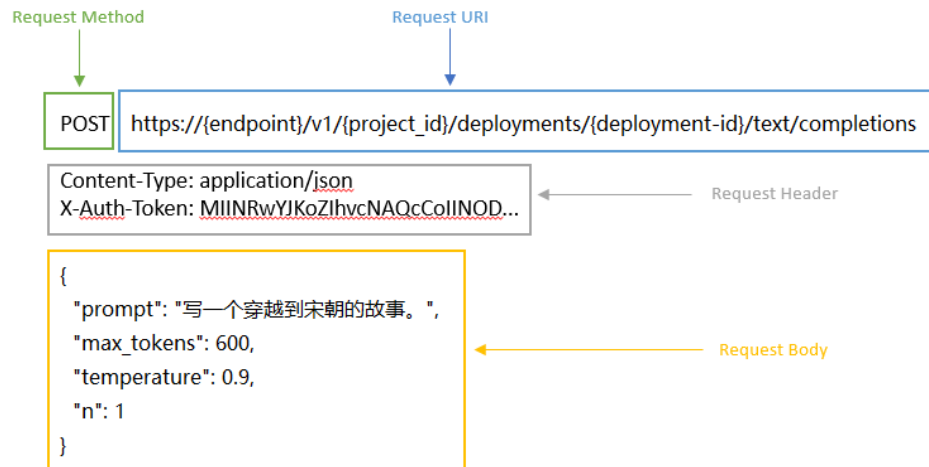
2.2 构造请求

本节介绍REST API请求的组成，并以调用IAM服务的**获取用户Token**接口说明如何调用API，该API获取用户的Token，Token可以用于调用其他API时鉴权。

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>。

请求示例如下图所示，一个请求主要有请求URI、请求方法、请求消息头和请求消息体组成，各个部分将在下文详细解释。

图 2-4 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}

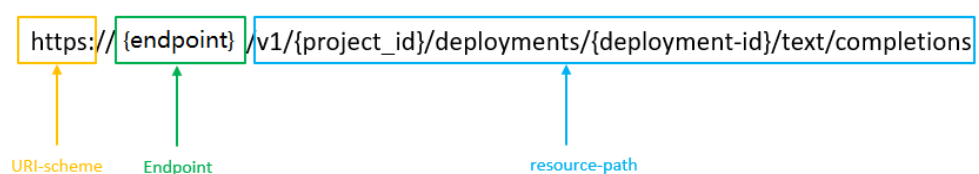
表 2-1 请求 URI

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”。

参考[终端节点](#)章节获取endpoint，并在[接口的URI](#)部分找到resource-path (/v1/{project_id}/deployments/{deployment_id}/text/completions)，其中{project_id}需要替换成用户的项目ID。拼接起来如下所示。

https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/text/completions

图 2-5 URI 示意图



📖 说明

为查看方便，每个具体API的URI，只给出resource-path部分，并将请求方法写在一起。这是因为URI-scheme都是HTTPS，而Endpoint在同一个区域也相同，所以简洁起见将这两部分省略。

请求方法

HTTP请求方法（也称为操作或动词），它告诉服务你正在请求什么类型的操作。

- **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，您可以看到其请求方法为“POST”，则其请求为：

```
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/text/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。

如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}/text/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}/text/completions
Content-Type: application/json
```

```
X-Auth-Token: MIINRwYJKoZlhvcNAQcCoIINOD...
```

```
{
  "prompt": "写一个穿越到宋朝的故事。",
  "max_tokens": 600,
  "temperature": 0.9,
  "n": 1
}
```

到这里为止这个请求需要的内容就具备齐全了，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。

2.3 认证鉴权

调用接口有如下两种认证方式，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。
- 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，可以先缓存起来，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"])
```

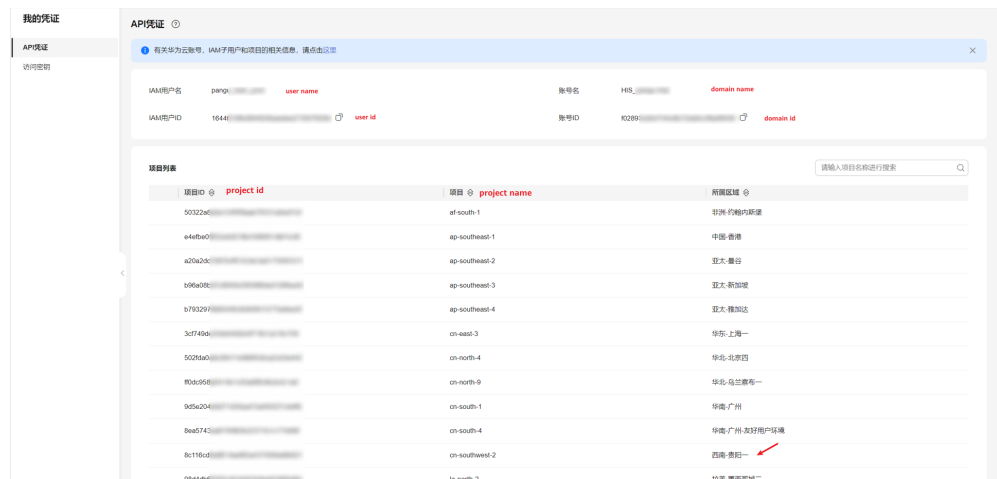
获取Token步骤:

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。

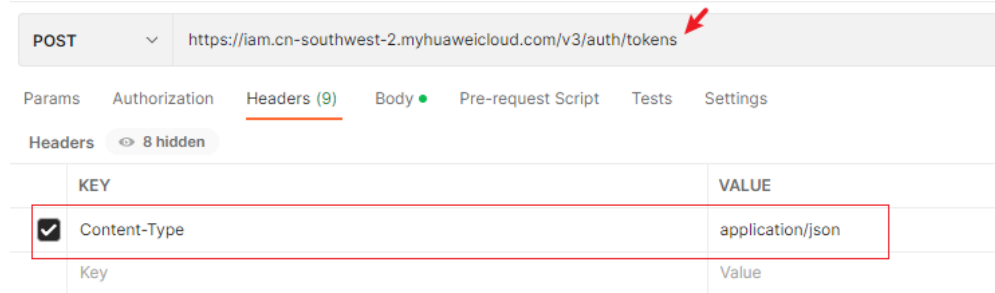
由于盘古大模型当前部署在“西南-贵阳一”区域，需要获取与“西南-贵阳一”区域对应的project id。

图 2-6 获取 user name、domain name、project id



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

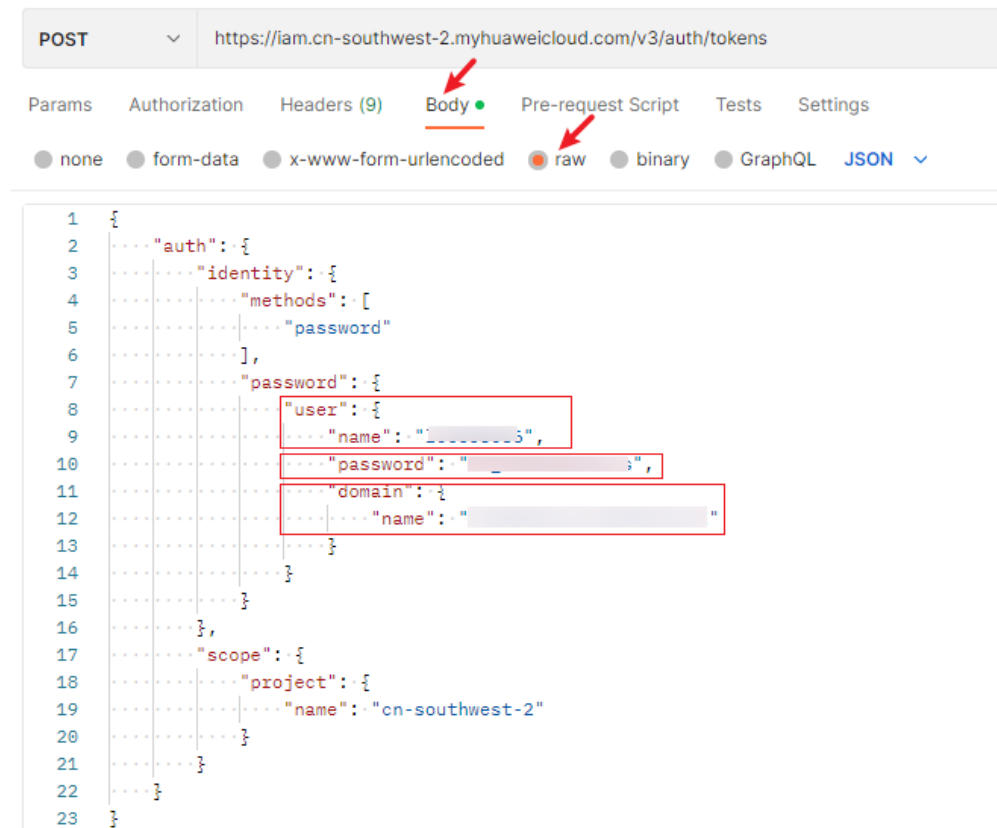
图 2-7 填写获取 Token 接口



3. 填写“获取token”接口的请求体。在Postman中选择“Body > raw”选项，参考图2-8复制并填入以下代码，并填写user name、domain name、password。

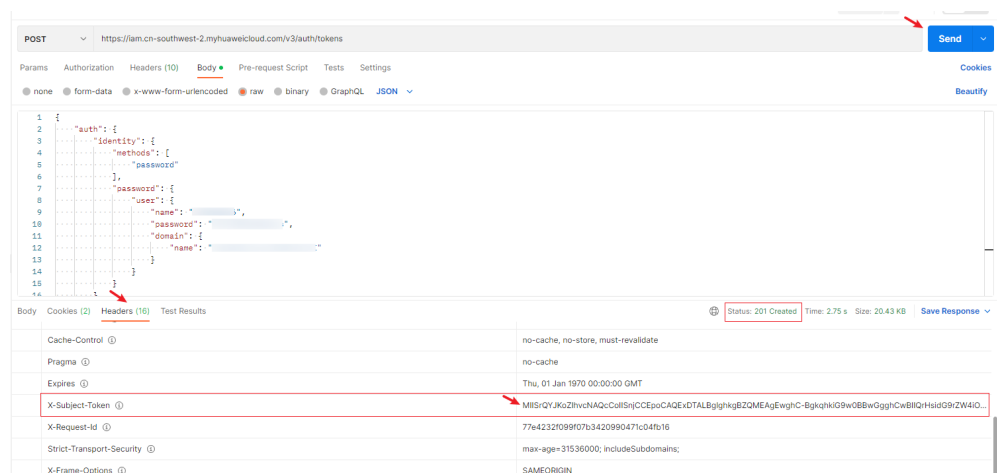
```
{
  "auth": {
    "identity": {
      "methods": [
        "password"
      ],
      "password": {
        "user": {
          "name": "username", //IAM用户名
          "password": "*****", //华为云账号密码
          "domain": {
            "name": "domainname" //账号名
          }
        }
      }
    }
  },
  "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 2-8 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 2-9 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>。

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](#)。

2.4 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 文本补全

功能介绍

给定一个提示和一些参数，模型会根据这些信息生成一个或多个预测的补全。它可以用来做文本生成、自动写作、代码补全等任务。

URI

POST /v1/{project_id}/deployments/{deployment_id}/text/completions

表 3-1 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 获取模型调用API地址 。

请求参数

表 3-2 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 3-3 请求 Body 参数

参数	是否必选	参数类型	描述
prompt	是	String	向模型输入的文本信息，单位token。 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。 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。
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。

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>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，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>最小值：0 最大值：1 缺省值：N1模型为0.4、N2模型为0.3，N4模型为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。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>取值范围：(0, 1] 缺省值：N1模型为0.7、N2模型为1，N4模型为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：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。</p> <p>缺省值：默认部署时有token长度最大值，请参见《产品介绍》“模型规格 > 模型基础信息”章节。</p>

参数	是否必选	参数类型	描述
presence_penalty	否	Float	<p>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>最小值: -2 最大值: 2 缺省值: 0 (表示该参数未生效)</p>
frequency_penalty	否	Float	<p>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>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)</p>
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>

响应参数

非流式

状态码: 200

表 3-4 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	生成的补全信息的列表，包含以下属性： <ul style="list-style-type: none">• text: 补全的文本内容。• index: 补全的索引，从0开始。
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 3-5 choices

参数	参数类型	描述
text	String	补全的文本内容。
index	Integer	补全的索引，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 3-6 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 3-7 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 3-8 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 3-9 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

- **非流式**
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/text/completions

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

Request Body:
{
 "prompt": "写一个穿越到宋朝的故事。",
 "temperature": 0.9
}
- **流式 (stream参数为true)**
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/text/completions

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

Request Body:
{
 "prompt": "介绍下长江三峡",
 "temperature": 0.9,
 "stream": "true"
}

响应示例

非流式**状态码： 200**

OK

```
{
  "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hession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
  }
}
```

流式 (stream参数为true)**状态码： 200**

OK

```
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"长江"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"三峡"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"是"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"瞿"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"塘"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"峡"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"、"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"巫"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"峡"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"和"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"西"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"陵"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"峡"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"三"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"段"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"峡谷"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"的"}]}
```

```
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"总"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"称"}]}
data:{"id":"e95727b0-fe09-4f18-96db-98354bd30e57","created":1687918751,"choices":[{"text":"。"}]}
data:[DONE]
```

流式（stream参数为true，触发内容审核）

状态码： 200

OK

event: moderation:{"suggestion":"block","reply":"作为AI语言模型，不能接受或表达任何不当内容。无论是在什么情况下，我们都应该保持对他人的尊重和礼貌，并且以积极、正向和安全的方式回答问题。"}

状态码

请参见[状态码](#)。

错误码

请参见[错误码](#)。

3.1.2 多轮对话

功能介绍

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。

URI

POST /v1/{project_id}/deployments/{deployment_id}/chat/completions

表 3-10 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 获取模型调用API地址 。

请求参数

表 3-11 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 3-12 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	多轮对话问答对，包含两个属性：role和content。 <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。 messages参数可以帮助模型根据对话的上下文生成合适的回复。 数组长度：1 - 20
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。

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>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，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>最小值：0 最大值：1 缺省值：N1模型为0.4、N2模型为0.3，N4模型为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。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>取值范围：(0, 1] 缺省值：N1模型为0.7、N2模型为1，N4模型为0.85</p>

参数	是否必选	参数类型	描述
max_tokens	否	Integer	<p>生成文本的最大token数量。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</p> <p>最大值：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。</p> <p>缺省值：默认部署时token长度最大值，请参见《产品介绍》“模型规格 > 模型基础信息”章节。</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>
presence_penalty	否	Float	<p>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>最小值：-2</p> <p>最大值：2</p> <p>缺省值：0（表示该参数未生效）</p>
frequency_penalty	否	Float	<p>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</p> <p>最小值：-2</p> <p>最大值：2</p> <p>缺省值：0（表示该参数未生效）</p>

参数	是否必选	参数类型	描述
stream	否	boolean	流式开关。 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。

表 3-13 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 3-14 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。

参数	参数类型	描述
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。
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 3-15 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 3-16 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 3-17 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 3-18 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 3-19 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 3-20 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": "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": "张家界具体在哪里" //第三轮问题
  }
],
"temperature": 0.9,
"max_tokens": 600,
"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. **鳙鱼**（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。长江好好好长，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": "。"}}]}
```

```
{"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：  
介绍：  
### 长沙岳麓山  
概述：  
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。  
自然风光：  
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。  
文化底蕴：  
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》。  
旅游活动：  
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eutic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  
Jesus Christ概述  
扫地僧身份曝光：  
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：  
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：  
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：  
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": ""}}]}
```

```
[{"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](#)。

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	服务器等候请求时发生超时。 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”。

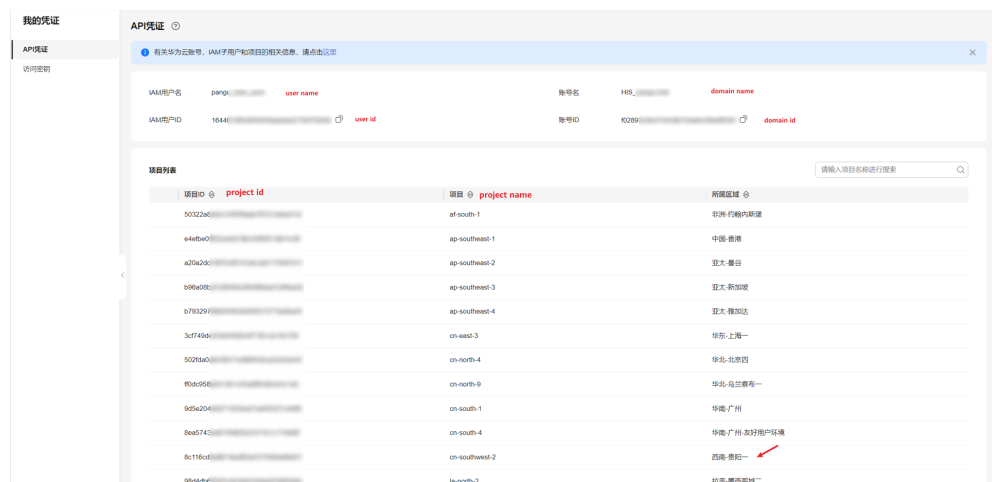
图 4-1 我的凭证



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 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 获取模型调用 API 地址

1. 登录[盘古大模型套件平台](#)。
2. 在左侧导航栏中选择“服务管理”，在相应服务的操作列单击“查看详情”，可在服务列表中申请需要开通的服务。

图 4-3 服务管理



3. 在“概览 > 服务列表”中选择需要调用的模型，并单击操作列的“调用路径”。

图 4-4 服务概览页面



4. 在弹窗中可获取对应模型的API请求地址。其中，路径选中部分即为模型的部署ID (deployment_id) 。

图 4-5 获取 API 请求地址

