

盘古大模型

API 参考

文档版本 01
发布日期 2023-09-30



版权所有 © 华为云计算技术有限公司 2023。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..... 1
- 2 如何调用 REST API..... 3
 - 2.1 构造请求..... 3
 - 2.2 认证鉴权..... 6
 - 2.3 返回结果..... 9
- 3 API..... 11
 - 3.1 通用文本..... 11
 - 3.2 对话问答..... 18
- 4 附录..... 30
 - 4.1 状态码..... 30
 - 4.2 错误码..... 32
 - 4.3 获取项目 ID..... 35
- 5 修订记录..... 37

1 使用前必读

1.1 概述

盘古大模型整合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，当前提供以下能力。

表 1-1 API 清单

API	功能
通用文本	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。
对话问答	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](#)。

1.3 终端节点

终端节点（endpoint）即调用API的请求地址，认证鉴权所需要的终端节点请参考[认证鉴权](#)获取，API接口的endpoint信息请参考[构造请求](#)获取。

1.4 基本概念

- 租户

租户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。为了确保租户安全，建议您不要直接使用租户进行日常管理工作，而是创建用户并使用他们进行日常管理工作。

- 用户

由租户在ManageOne运营面创建的用户，是云服务的使用人员，具有身份凭证（密码和访问密钥）。

在个人设置下，您可以查看租户ID和用户ID。通常在调用API的鉴权过程中，您需要用到租户名、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。

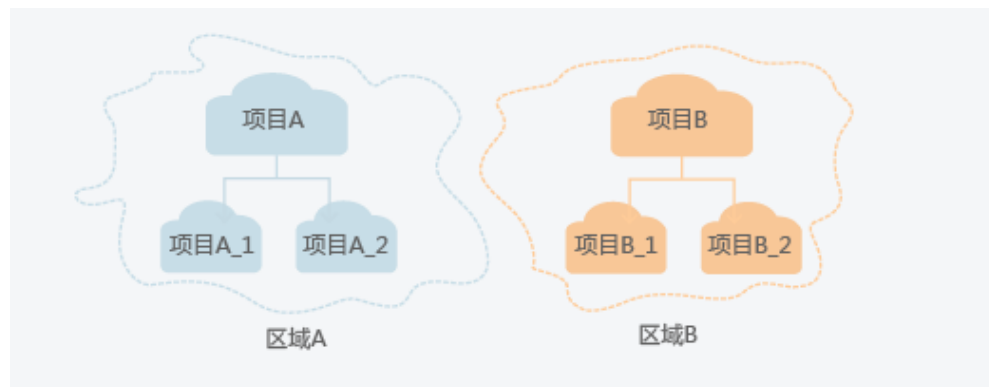
- 可用区（AZ，Availability Zone）

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

云服务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，用户可以访问您租户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，并在子项目中创建资源，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 1-1 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。

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 2-1 请求示例图

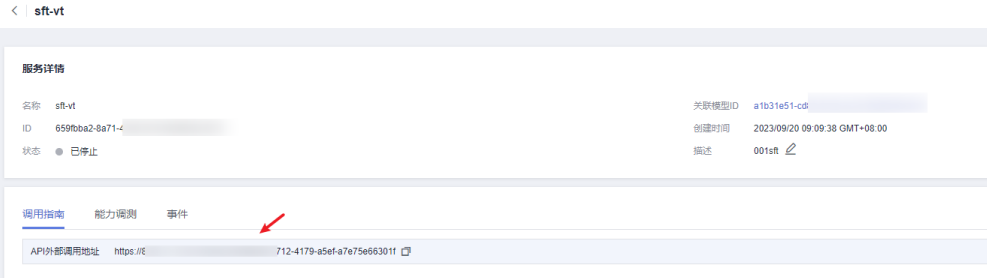


获取 https://{endpoint}

模型部署成功后，将部署为在线服务，用户可以通过API调用模型能力。

在模型部署列表中，单击模型名称，进入模型调用指南页面，可查看模型的API接口信息。

图 2-2 获取 endpoint



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 2-3 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

云服务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。

2.2 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。

1. 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。
2. 用户所属的帐户名，通过下面的步骤获取：
 - a. 登录ManageOne运营面。
 - b. 在页面右上角单击用户图标，在下拉列表中选择“个人设置”。
 - c. 在“个人设置”页面，显示的帐户即为该用户所属的帐户名。
3. 调用IAM接口获取token之前，需要将IAM地址和域名的映射，写入系统的hosts文件，通过下面的步骤实现。
 - a. 示例中的GLOBAL_DOMAIN_NAME请从HCCI Turnkey工具安装智能中枢服务的参数导出表>1.1基本参数章节中参数为“global_domain_name”的规划值，如下图所示：

参数分类	参数名称	参数key	规划值	规划值限制	是否
	Yum源软件仓库root密码	yum_password	*****		CON
	内部Global域名	global_domain_name	internal3.com	demo.com	TRUE
	外部Global Level Domain	external_global_domain_name	external3.com	demo.com	TRUE

- b. 在安装底座时导出的汇总文件《xxx_export_all_v2_CN.xlsx》的“2.1工程生成的IP参数”页签，查找参数名称为“ManageOne-Tenant-Float-IP”的IP地址。

参数名称	规划值	是否
external on ManageOne-Managed-Tenant-Float-IP	external on	TRUE

4. 打开机器的“C:\Windows\System32\drivers\etc\hosts”文件，在文件的最后一行写入步骤3获取到的IP和步骤2获取到的域名iam-cache-proxy. {GLOBAL_DOMAIN_NAME}的映射。
{IP} {iam-cache-proxy.{GLOBAL_DOMAIN_NAME}}

Token可通过调用“获取用户Token”接口获取，调用本服务API需要project级别的Token，即调用获取用户Token接口时，请求body中“auth.scope”的取值需要选择“project”，请求示例如下所示。“POST https://{endpoint}/v3/auth/tokens”接口中的{endpoint}可联系服务技术支持获取。

user name、domain name、project id获取方法如下。

1. 登录盘古大模型平台，在界面右上角，单击租户名称后，选择“个人设置”，进入个人信息界面。

图 2-4 个人设置



2. 在个人信息页，参考下图获取user name、domain name、project id。

图 2-5 获取 user name、domain name、project id



- 伪码
POST https://{endpoint}/v3/auth/tokens
Content-Type: application/json
{
 "auth": {
 "identity": {
 "methods": [
 "password"
],
 "password": {
 "user": {
 "name": "username", //用户名
 "password": "*****", //密码
 "domain": {
 "name": "domainname" //租户名
 }
 }
 }
 },
 "scope": {
 "project": {
 "name": "project name" //替换为实际的project name
 }
 }
 }
}

}

- Python

```
import requests
import json

url = "https://{endpoint}/v3/auth/tokens"
payload = json.dumps({
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",
                    "password": "*****",
                    "domain": {
                        "name": "domainname"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "projectname"//替换为实际的项目名称
            }
        }
    }
})
headers = {
    'Content-Type': 'application/json'
}

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

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

如下图所示，返回的响应消息头中“x-subject-token”就是需要获取的用户Token。获取Token之后，您就可以使用Token认证调用OCR服务API。

图 2-6 获取用户 Token 响应消息头

[illegible]

AK/SK 认证

说明

AK/SK签名认证方式仅支持消息体大小12MB以内，12MB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。

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](#)。

对于盘古大模型服务接口，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”。SIS服务响应消息头无特殊用途，可用于定位问题使用。

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": "dd5b73dd5775d53366b6a61aac6080d5",
  "created": 20230512025050,
  "choices": [
    {
      "index": 0,
      "text": "好的,我来为您写一个穿越到宋朝的故事。故事的主人公是一个名叫张云的年轻人,他因为一次意外被送到了北宋时期。在这个陌生的时代里,张云经历了许多奇妙的事情,结交了许多有趣的朋友,也遇到了不少危险和挑战。最终,他在这个时代中找到了自己的归属感和意义所在,成为了一名伟大的历史学家。"
    }
  ],
  "usage": {
    "completion_tokens": 72,
    "prompt_tokens": 7,
    "total_tokens": 79
  }
}
```

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 通用文本

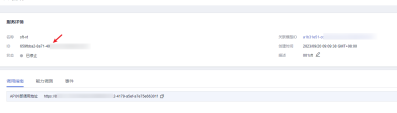
功能介绍

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。 模型在使用 workflows 训练部署完成后，单击模型部署列表中的模型，进入详情页获取。 

请求参数

说明

盘古大模型支持百亿/千亿两类模型，不同模型在参数取值上有所差异，存在差异点的参数请参见[表3-2](#)。

表 3-2 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：1408 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。
temperature	否	Float	用于控制生成文本的多样性和创造力。 参数的取值范围是 (0, 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。 temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两个参数。 取值范围：(0, 1] 缺省值：百亿模型 0.4，千亿模型 0.5

参数	是否必选	参数类型	描述
top_p	否	Float	<p>用于控制生成文本的多样性。它的含义是在每个时间步，只考虑概率累积和小于top_p的词作为候选词。例如，如果top_p=0.9，那么只有概率最高的10%的词会被选中，其他的词会被忽略。这样可以避免生成一些过于常见或者无意义的词，提高生成文本的质量和创造性。top_p的取值范围是0到1，越接近1表示越多样，越接近0表示越保守。</p> <p>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]</p> <p>缺省值：0.7</p> <p>说明</p> <p>top_p参数背后的技术方法称为nucleus sampling（核采样），它是一种文本生成中的解码策略，它的目的是让生成的文本更多样化和自然。在nucleus sampling中，时间步是指生成文本的每个位置，也就是每次从词汇表中采样一个词的过程。在每个时间步，nucleus sampling会给定一个概率阈值p，然后从词汇表中选择一个最小集合，使得它们的概率和大于等于p，这个集合就叫做nucleus。然后再从nucleus中按照概率分布进行采样，得到下一个词。这样的好处是，在不同的时间步，根据词汇表的概率分布的不同，nucleus的大小也会动态变化，从而避免了生成低概率或者重复的词。</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</p> <p>最大值：640</p> <p>缺省值：16</p>

参数	是否必选	参数类型	描述
n	否	Integer	<p>表示对每个问题生成多少条答案。</p> <p>n参数的默认值是1，表示只生成一个答案。如果想要生成多条答案，可以设置n参数为一个大于1的整数，例如n=2。这样，API会返回一个包含2个答案的数组。</p> <p>最小值：1 最大值：2 缺省值：1</p> <p>说明</p> <ul style="list-style-type: none">流式只能传1，非流式时可传1和2。因为这个参数会产生很多条答案，n设置较大时，会很快消耗完tokens的配额。请谨慎使用，并确保对 max_tokens 进行了合理的设置。
presence_penalty	否	Float	<p>用于控制生成文本中的重复程度。正值会根据它们到目前为止在文本中的现有频率来惩罚新tokens，从而降低模型逐字重复同一行的可能性。</p> <p>presence_penalty 参数可以用来提高生成文本的多样性和创造性，避免生成单调或重复的内容。</p> <p>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 3-3 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 3-4 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。

表 3-5 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 3-6 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 3-7 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： default

表 3-8 响应 Body 参数

参数	参数类型	描述
error	error object	调用API失败时的错误信息。

表 3-9 error

参数	参数类型	描述
code	String	错误码。
message	String	错误信息。
param	String	错误参数。
type	String	错误类型。

请求示例

说明

调用API的请求地址{endpoint}获取方法请参见[获取https://{endpoint}](#)。

- 非流式
POST https://{endpoint}/v1/{project-id}/deployments/{deployment-id}/text/completions

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

Request Body:
{
 "prompt": "写一个穿越到宋朝的故事。",
 "max_tokens": 600,
 "temperature": 0.9,
 "n": 1
}
- 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:
MIINRwYJKoZlHvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgggVBGkqhkiG...

Request Body:
{
  "prompt": "介绍下长江三峡",
  "max_tokens": 600,
  "temperature": 0.9,
  "n": 1,
  "stream": "true"
}
```

响应示例

非流式

状态码： 200

OK

```
{
  "id": "dd5b73dd5775d53366b6a61aac6080d5",
  "created": 20230512025050,
  "choices": [
    {
      "index": 0,
      "text": "好的,我来为您写一个穿越到宋朝的故事。故事的主人公是一个名叫张云的年轻人,他因为一次意外被送到了北宋时期。在这个陌生的时代里,张云经历了许多奇妙的事情,结交了许多有趣的朋友,也遇到了不少危险和挑战。最终,他在这个时代中找到了自己的归属感和意义所在,成为了一名伟大的历史学家。"
    }
  ],
  "usage": {
    "completion_tokens": 72,
    "prompt_tokens": 7,
    "total_tokens": 79
  }
}
```

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:[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

状态码	描述
200	OK
default	Service unavailable

错误码

请参见[错误码](#)。

3.2 对话问答

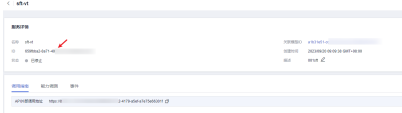
功能介绍

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。 模型在使用 workflows 训练部署完成后，单击模型部署列表中的模型，进入详情页获取。 

请求参数

说明

盘古大模型支持百亿/千亿两类模型，不同模型在参数取值上有所差异，存在差异点的参数请参见表3-11。

表 3-11 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of message objects	messages表示用户与模型进行对话时，输入的信息。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。 messages参数可以帮助模型根据对话的上下文生成合适的回复。
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。
temperature	否	Float	用于控制生成文本的多样性和创造力。 参数的取值范围是 (0, 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。 temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两个参数。 取值范围：(0, 1] 缺省值：百亿模型 0.4，千亿模型 0.5

参数	是否必选	参数类型	描述
top_p	否	Float	<p>用于控制生成文本的多样性。它的含义是在每个时间步，只考虑概率累积和小于top_p的词作为候选词。例如，如果top_p=0.9，那么只有概率最高的10%的词会被选中，其他的词会被忽略。这样可以避免生成一些过于常见或者无意义的词，提高生成文本的质量和创造性。top_p的取值范围是0到1，越接近1表示越多样，越接近0表示越保守。</p> <p>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]</p> <p>缺省值：0.7</p> <p>说明</p> <p>top_p参数背后的技术方法称为nucleus sampling（核采样），它是一种文本生成中的解码策略，它的目的是让生成的文本更多样化和自然。在nucleus sampling中，时间步是指生成文本的每个位置，也就是每次从词汇表中采样一个词的过程。在每个时间步，nucleus sampling会给定一个概率阈值p，然后从词汇表中选择一个最小集合，使得它们的概率和大于等于p，这个集合就叫做nucleus。然后再从nucleus中按照概率分布进行采样，得到下一个词。这样的好处是，在不同的时间步，根据词汇表的概率分布的不同，nucleus的大小也会动态变化，从而避免了生成低概率或者重复的词。</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，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</p> <p>最大值：640</p> <p>缺省值：16</p> <p>说明</p> <p>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>
n	否	Integer	<p>表示对每个问题生成多少条答案。</p> <p>n参数的默认值是1，表示只生成一个答案。如果想要生成多条答案，可以设置n参数为一个大于1的整数，例如n=2。这样，API会返回一个包含2个答案的数组。</p> <p>最小值：1</p> <p>最大值：2</p> <p>缺省值：1</p> <p>说明</p> <ul style="list-style-type: none">流式只能传1，非流式时可传1和2。因为这个参数会产生很多条答案，n设置较大时，会很快消耗完tokens的配额。请谨慎使用，并确保对 max_tokens 进行了合理的设置。

参数	是否必选	参数类型	描述
presence_penalty	否	Float	用于控制生成文本中的重复程度。正值会根据它们到目前为止在文本中的现有频率来惩罚新tokens，从而降低模型逐字重复同一行的可能性。 presence_penalty 参数可以用来提高生成文本的多样性和创造性，避免生成单调或重复的内容。 最小值：-2 最大值：2 缺省值：0（表示该参数未生效）
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 3-12 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	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：2048 注意：在输入content值时，最大输入长度为2048token，但在进行推理计算时，会从输入内容的后面往前截取，截断成1408token后传入模型。

响应参数

非流式
状态码： 200

表 3-13 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。
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 3-14 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。

表 3-15 message

参数	参数类型	描述
role	String	对话的角色。
content	String	对话的内容，单位token。 最小长度： 1 最大长度： 640 缺省值： 640

表 3-16 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。

参数	参数类型	描述
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 3-17 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 3-18 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： default

表 3-19 响应 Body 参数

参数	参数类型	描述
error	error object	调用API失败时的错误信息。

表 3-20 error

参数	参数类型	描述
code	String	错误码。
message	String	错误信息。
param	String	错误参数。

参数	参数类型	描述
type	String	错误类型。

请求示例

说明

调用API的请求地址{endpoint}获取方法请参见[获取https://{endpoint}](https://{endpoint})。

- 单轮问答
POST https://{endpoint}/v1/{project-id}/deployments/{deployment-id}/chat/completions

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

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

```
{
  "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": "2f8e891225d486190c8bea91207e9aa1",
  "created": 20230512084843,
  "choices": [
    {
      "index": 0,
      "message": {
        "role": null,
        "content": "长江是中国第一大河,发源于青藏高原的唐古拉山脉。全长6300千米,流域面积达180万平方千米。\\n典型的鱼类包括:草鱼、鲤鱼、鲢鱼等淡水鱼和鳙鱼、鲶鱼等咸水鱼。"
      }
    }
  ],
  "usage": {
    "completion_tokens": 61,
    "prompt_tokens": 11,
    "total_tokens": 72
  }
}
```

- 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:[DONE]
```

- 带有人设的单轮问答（role参数值为system）

```
{
  "id": "df01600b0da594d596a98902fae795dd",
  "created": 20230719020941,
  "choices": [
    {
      "index": 0,
      "message": {
        "role": null,
        "content": "长江是中国第一大河,全长6300多公里。它发源于青藏高原上的唐古拉山脉,经过青海、西藏、四川、云南等省份,最终注入东海。\\n在长江中生活着许多著名的鱼类,比如:草鱼、鲢鱼、鳙鱼、鲫鱼、鲤鱼等等。其中最著名的当属中华鲟了!它是中国特有的淡水鱼类之一,也是世界上现存最古老的脊椎动物之一。此外还有白鳍豚、扬子鳄等多种珍稀物种生活在长江之中。"
      }
    }
  ],
  "usage": {
    "completion_tokens": 110,
    "prompt_tokens": 126,
    "total_tokens": 236
  }
}
```

- ```
}
}
```
- 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": "c238908bd80aef4e4fbfc9fad24460d",  
 "created": 20230512104144,  
 "choices": [  
 {  
 "index": 0,  
 "message": {  
 "role": null,  
 "content": "1、张家界:是湖南省辖地级市,原名大庸市。辖2个市辖区(永定区、武陵源区)、2个县(慈利县、桑植县)。位于湖南西北部,澧水中上游,属武陵山区腹地。\\n2、凤凰古城:建于清康熙四十三年(1704年),距铜仁凤凰机场27公里;这里与吉首的德夯苗寨、永顺的猛洞河、贵州的梵净山相毗邻,是怀化、吉首、贵州铜仁三地之间的必经之路。湖南十大文化遗产之一。曾被新西兰作家路易·艾黎称赞为中国最美丽的小城,与云南丽江古城、山西平遥古城媲美,享有“北平遥、南凤凰”之名。是国家历史文化名城,国家AAAA级景区。"  
 }  
 },  
 {  
 "index": 1,  
 "message": {  
 "role": null,  
 "content": "2、凤凰古城:建于清康熙四十三年(1704年),距铜仁凤凰机场27公里;这里与吉首的德夯苗寨、永顺的猛洞河、贵州的梵净山相毗邻,是怀化、吉首、贵州铜仁三地之间的必经之路。湖南十大文化遗产之一。曾被新西兰作家路易·艾黎称赞为中国最美丽的小城,与云南丽江古城、山西平遥古城媲美,享有“北平遥、南凤凰”之名。是国家历史文化名城,国家AAAA级景区。"  
 }  
 }  
 ],  
 "usage": {  
 "completion\_tokens": 180,  
 "prompt\_tokens": 9,  
 "total\_tokens": 189  
 }  
}
  - 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": "张家界"}}]}





# 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。<br>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。<br>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。<br>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                     | 查看其它地址，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。<br>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。<br>返回该状态码，表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为该请求被设置为拒绝访问，建议直接修改该请求，不要重试该请求。 |
| 404 | Not Found                     | 所请求的资源不存在。<br>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。<br>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。<br>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。<br>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。<br>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，从远端服务器接收到了一个无效的请求。                                                        |
| 503 | Service Unavailable             | 被请求的服务无效。<br>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网关服务的错误码进行处理。

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                                               | 权限校验失败。              | 请联系服务技术支持协助解决。                     |
| PANGU.0012 | auth info missing                                         | 权限校验信息丢失。            | 请联系服务技术支持协助解决。                     |
| 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                                    | 待查询的资源不存在。           | 请联系服务技术支持协助解决。                     |
| PANGU.3254 | The requested resource does not exist.                    | 资源不存在。               | 请确认projectId和deploymentId是否正确。     |
| PANGU.3259 | model instance status is not running or have been deleted | 模型实例状态异常。            | 请确认deploymentId是否正确，或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。              | 请检查请求参数是否正确。                       |
| PANGU.3306 | The accessed API's model does not match instance's model  | 访问的API模型与实例模型不匹配。    | 请检查depolymntId是否正确，或联系服务技术支持协助解决。  |
| 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 | domain has not added the opened API                       | 访问的API与已开通的API服务不匹配。 | 请确认调用的API是否正确，或需要新开通该API服务。        |
| PANGU.3315 | The accessed API's model instance is not public.          | API模型实例未公开。          | 请检查depolymntId是否正确，或联系服务运维人员协助解决。  |
| PANGU.3317 | maxtokensNumbe rllleagl                                   | 最大token不合法           | 请参考API文档输入正确的请求参数，并重新调试API。        |

| 错误码        | 错误信息                                                                          | 说明                                                                                                                                                                                                                                    | 建议解决方法                                                                                                                                                                                                                                                                                                                                       |
|------------|-------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| PANGU.3318 | totalContentLengt<br>hilleagl                                                 | Content长度<br>不合法                                                                                                                                                                                                                      | 请参考API文档输入正确的请求参数，并重新调试API。                                                                                                                                                                                                                                                                                                                  |
| PANGU.3318 | Private service<br>permission error.                                          | 权限错误。                                                                                                                                                                                                                                 | 请联系服务技术支持协助解决。                                                                                                                                                                                                                                                                                                                               |
| APIG.0101  | The API does not<br>exist or has not<br>been published in<br>the environment. | 访问的API不存在或尚未在环境中发布。                                                                                                                                                                                                                   | <ul style="list-style-type: none"><li>请检查API的URL是否拼写正确，例如，URL中是否缺少project_id。</li><li>HTTP请求方法（POST，GET等）是否正确。</li></ul>                                                                                                                                                                                                                     |
| APIG.0201  | Backend timeout.                                                              | 请求超时。                                                                                                                                                                                                                                 | <ul style="list-style-type: none"><li>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。</li><li>请与技术支持确认，API是否已完成部署。</li></ul>                                                                                                                                             |
| APIG.0301  | Incorrect IAM<br>authentication<br>information.                               | <p>IAM身份验证信息不正确：</p> <ul style="list-style-type: none"><li>decrypt token fail: token解析失败。</li><li>token expires: token过期。</li><li>verify aksk signature fail: AK/SK认证失败。</li><li>x-auth-token not found: 未找到x-auth-token参数。</li></ul> | <ul style="list-style-type: none"><li>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。</li><li>token超时（token expires），请重新获取token，使用不过期的token。</li><li>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）。</li><li>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。</li><li>检查请求header参数X-Auth-Token是否拼写正确。</li></ul> |

| 错误码       | 错误信息                                                                                          | 说明                | 建议解决方法                                                                                                                                       |
|-----------|-----------------------------------------------------------------------------------------------|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| APIG.0308 | The throttling threshold has been reached: policy user over ratelimit,limit:XX,time:1 minute. | 发送请求超过了服务的默认配置限流。 | <ul style="list-style-type: none"><li>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。</li><li>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。</li></ul> |

## 4.3 获取项目 ID

### 从控制台获取项目 ID

1. 登录ManageOne运营面。
2. 鼠标移动到右上角的用户名上，在下拉列表中选择“个人设置”。
3. 在“个人设置”页面，选择“资源集列表”页签，查看项目及项目ID。

图 4-1 查看项目 ID



### 调用 API 获取项目 ID

获取项目ID的接口为“GET https://{endpoint}/v3/projects”，其中“{endpoint}”为IAM的终端节点。接口的认证鉴权请参见[认证鉴权](#)。

响应示例如下，依据服务的部署区域在响应消息体中查找“name”参数“projects”下的“id”即为项目ID。

```
{
 "projects": [{
 "domain_id": "65382450e8f64ac0870cd180d14e684b",
 "is_domain": false,
 "parent_id": "65382450e8f64ac0870cd180d14e684b",
 "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"
 }
}
```

# 5 修订记录

| 发布日期       | 修改说明     |
|------------|----------|
| 2023-09-30 | 第一次正式发布。 |