盘古大模型

快速入门

文档版本 01

发布日期 2024-08-31

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 体验盘古大模型功能 1
1.1 体验盘古预置模型能力 1
1.2 体验盘古驱动的应用百宝箱 5
2 获取 API 认证鉴权信息（获取 Token） 7
3 调用盘古大模型 API 10
4 启用盘古大模型搜索增强能力 16

1 体验盘古大模型功能

1.1 体验盘古预置模型能力

登录盘古大模型套件平台，在左侧导航栏中单击“能力调测”。

如图所示，能力调测页面提供了文本补全和多轮对话功能，且每种功能都提供了预置的盘古大模型供用户体验。用户可以在页面右侧进行参数设置，然后在输入框中输入问题，模型就会返回对应的答案内容，具体参数信息如下表。

图 1-1 体验预置模型功能

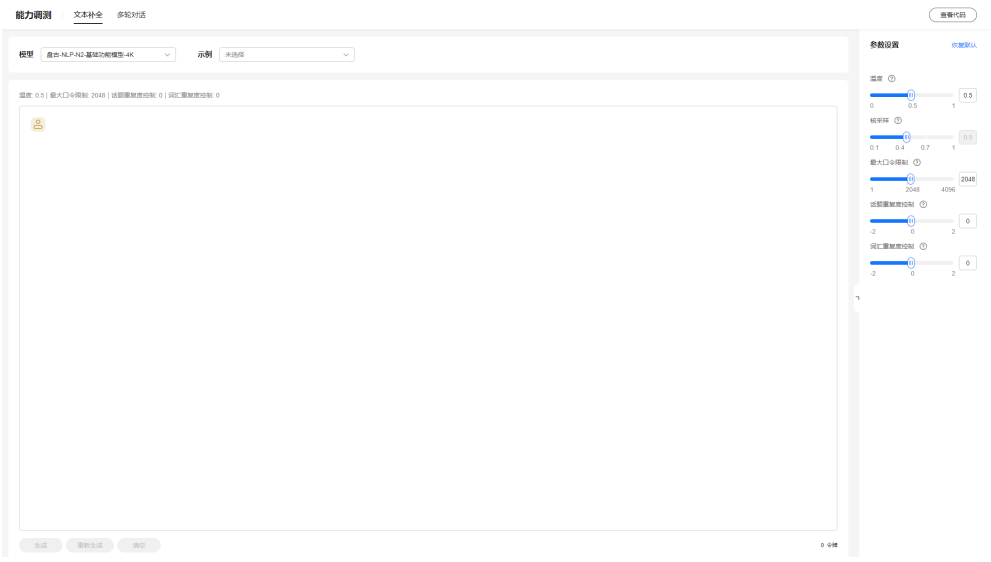


表 1-1 能力调测参数信息表

参数名称	描述
温度	控制语言模型输出的随机性与创造性。温度设置越低，输出更可预测；温度设置越高，输出种类更多，更不可预测。
核采样	控制生成文本多样性和质量。

参数名称	描述
最大口令限制	用于控制聊天回复的长度和质量。一般来说，设置较大的参数值可以生成较长和较完整的回复，但也可能增加生成无关或重复内容的风险。较小的参数值可以生成较短和较简洁的回复，但也可能导致生成不完整或不连贯的内容，请避免该值小于10，否则可能生成空值或极差的效果。因此，需要根据不同的场景和需求来选择合适的参数值。
话题重复度控制	用于调整模型对新令牌（Token）的处理方式。即如果一个Token已经在之前的文本出现过，那么模型在生成这个Token时会受到一定的惩罚。当值为正数时，模型会更倾向于生成新的Token，即更倾向于谈论新的话题。
词汇重复度控制	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。
历史对话保留轮数	选择要包含在每个新API请求中的过去消息数。这有助于为新用户查询提供模型上下文。参数设置为10，表示包括5个用户查询和5个系统响应。该参数只涉及多轮对话功能。

- 体验预置模型文本补全能力
  a. 进入“文本补全”页签，选择模型与示例，参数设置为默认参数，在输入框输入问题，单击“生成”，模型将基于问题进行回答。

  图 1-2 体验预置模型文本补全能力

  b. 修改参数以查看模型效果，示例如下：
     i. **将“核采样”参数调小，如改为0.1**，保持其他参数不变，单击“重新生成”，再单击“重新生成”，可以看到模型前后两次回复内容的多样性降低。

图 1-3 “核采样”参数调小后生成结果 1

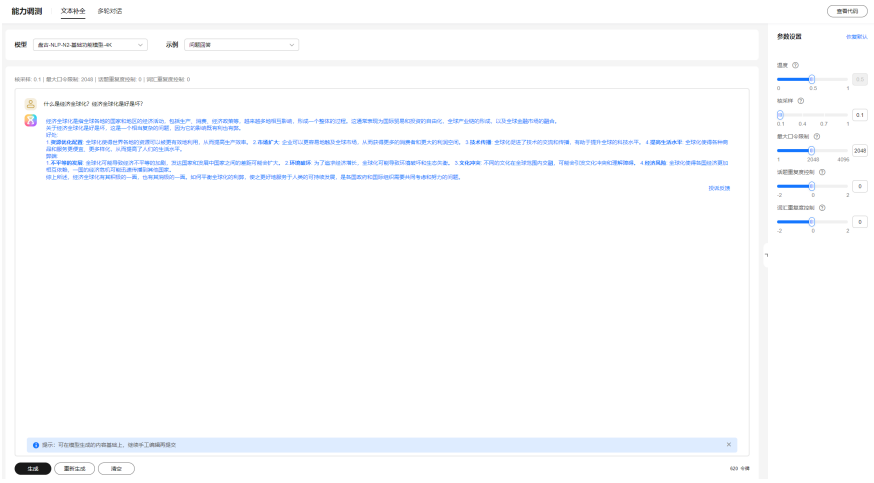
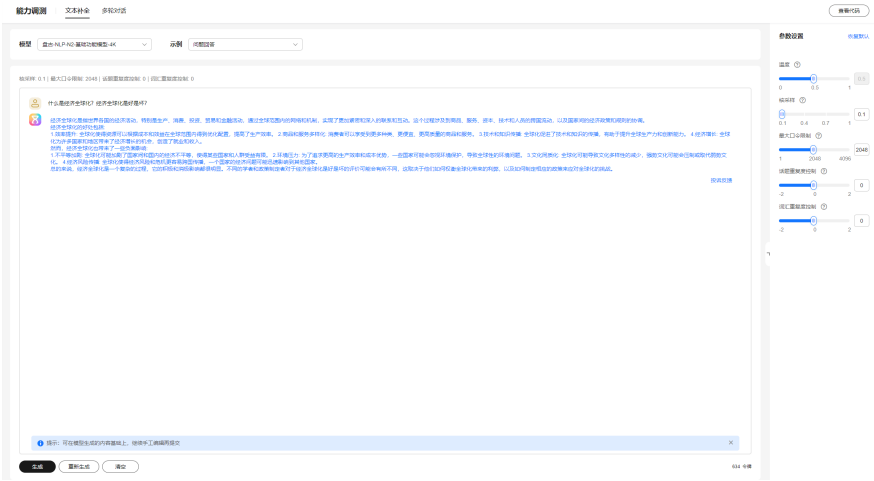


图 1-4 “核采样”参数调小后生成结果 2



ii. **将“核采样”参数调大，如改为1**，保持其他参数不变，单击“重新生成”，再单击“重新生成”，可以看到模型前后两次回复内容的多样性提高。

图 1-5 “核采样”参数调大后生成结果 1

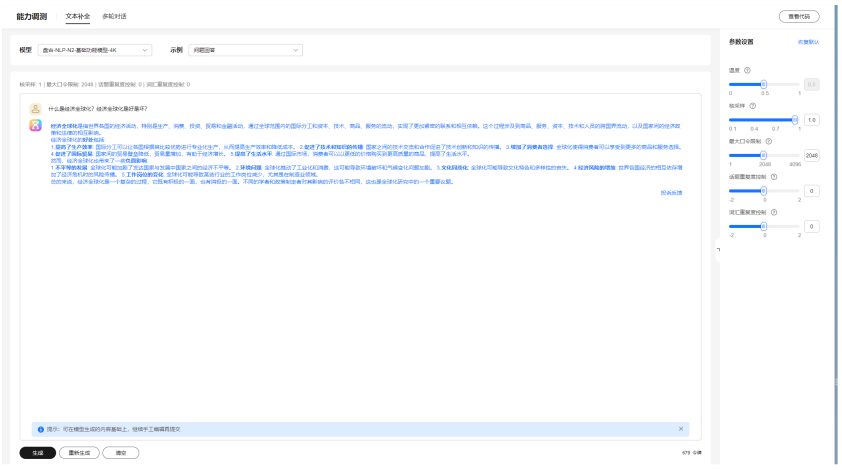
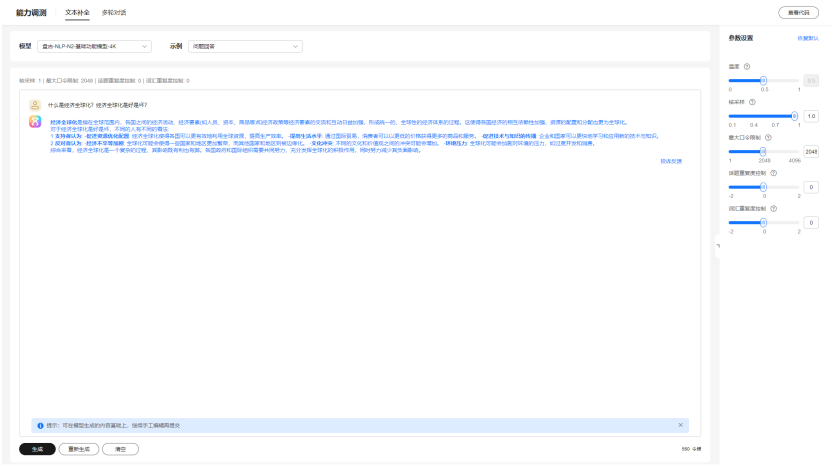


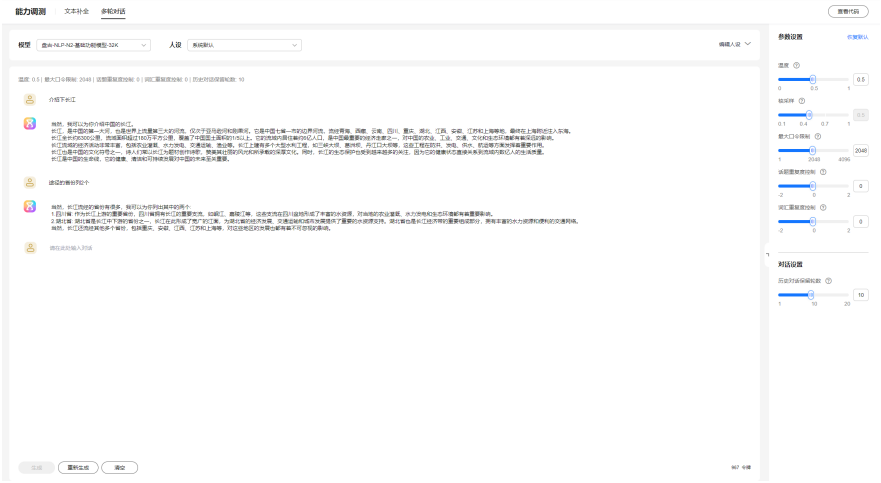
图 1-6 “核采样”参数调大后生成结果 2



- 体验预置模型的多轮对话能力

  a. 进入“多轮对话”页签，选择模型与人设，参数设置为默认参数，在输入框输入问题，单击“生成”，模型将基于问题进行回答。

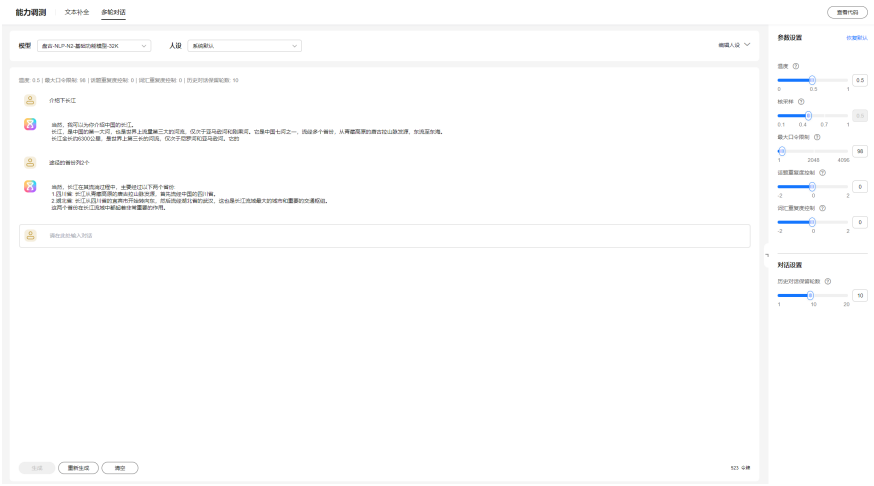
  图 1-7 体验预置模型多轮对话能力

  b. 修改参数以查看模型效果，示例如下：

  将“最大口令限制”参数调小，如改为98，保持其他参数不变，单击“重新生成”，可以看到模型回复内容长度减小。

图 1-8 修改“最大口令限制”参数



1.2 体验盘古驱动的应用百宝箱

应用百宝箱是盘古大模型为用户提供的便捷AI应用集，用户可在其中使用盘古大模型预置的场景应用和外部应用，轻松体验大模型开箱即用的强大能力。

1. 登录**盘古大模型套件平台**，在左侧导航栏中选择“应用百宝箱”，进入“应用百宝箱”页面。
2. 在“应用市场”页签中，选择场景应用，立即体验应用能力。

图 1-9 应用市场

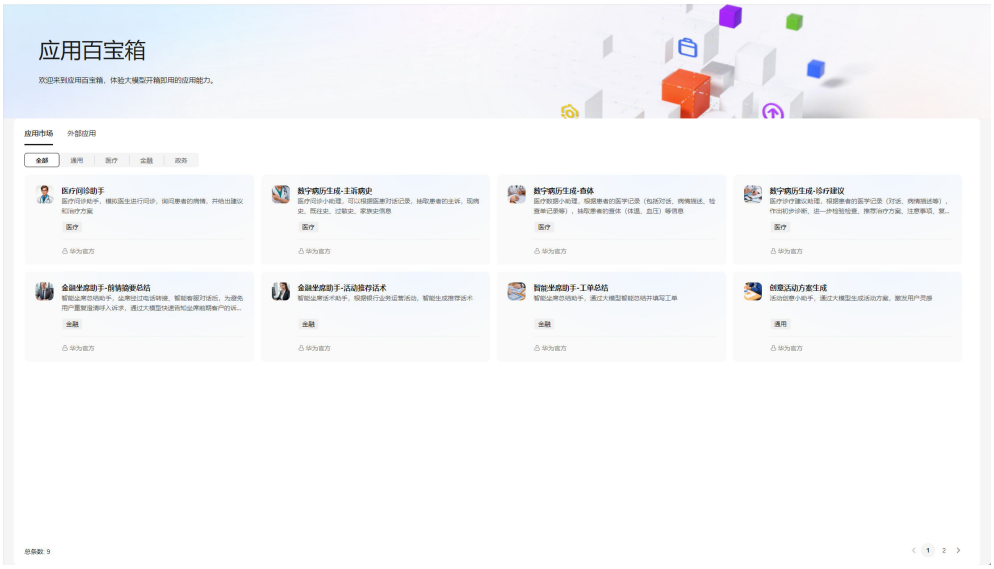
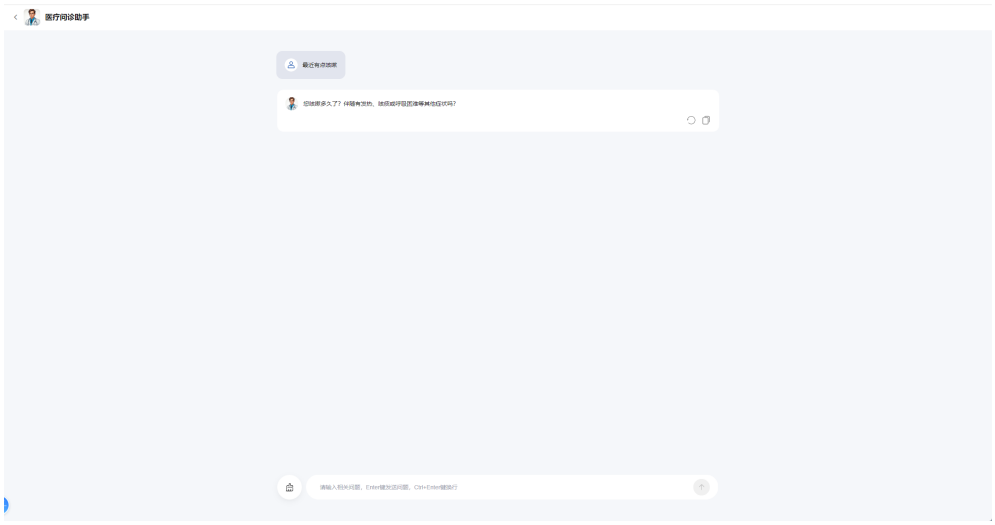


图 1-10 应用试用



3. 在“外部应用”页签中，选择外部应用，单击“继续前往”，页面将跳转至外部应用页面供用户体验。

图 1-11 外部应用

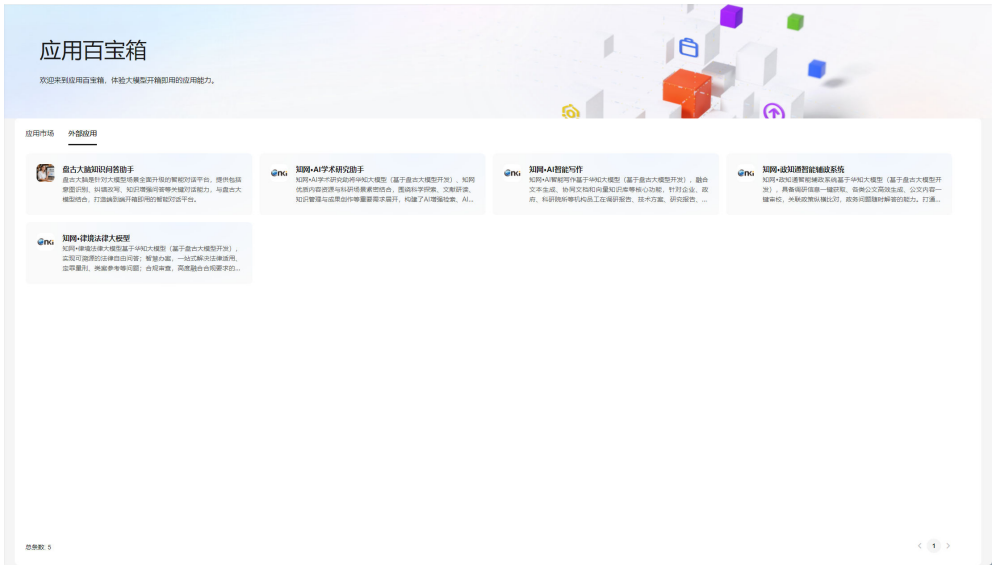
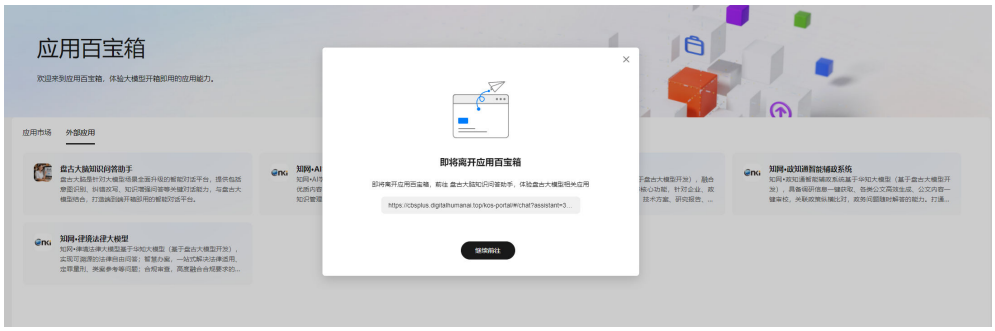


图 1-12 外部应用试用



2 获取 API 认证鉴权信息（获取 Token）

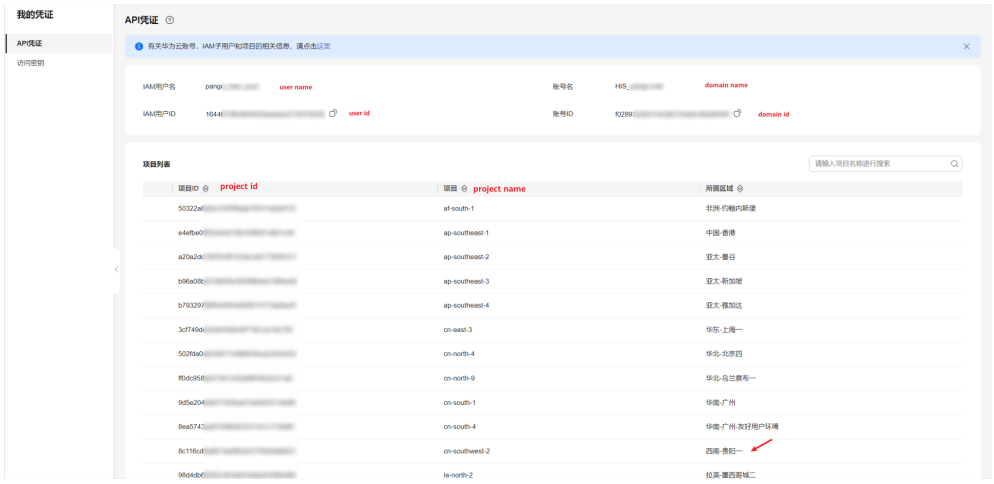
1. 登录“**我的凭证 > API凭证**”页面，获取user name、domain name、project id。

   project id参数需要与盘古服务部署区域一致。例如，盘古大模型部署在“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 2-1 查看盘古服务区域

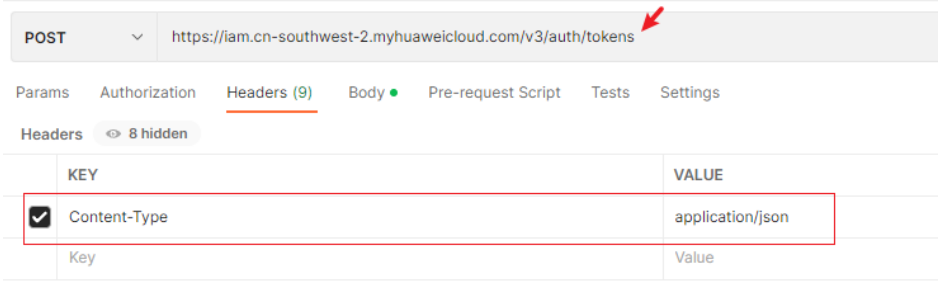
   图 2-2 获取 user name、domain name、project id

2. 下载并安装**Postman**调测工具。
3. 打开Postman，新建一个POST请求，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

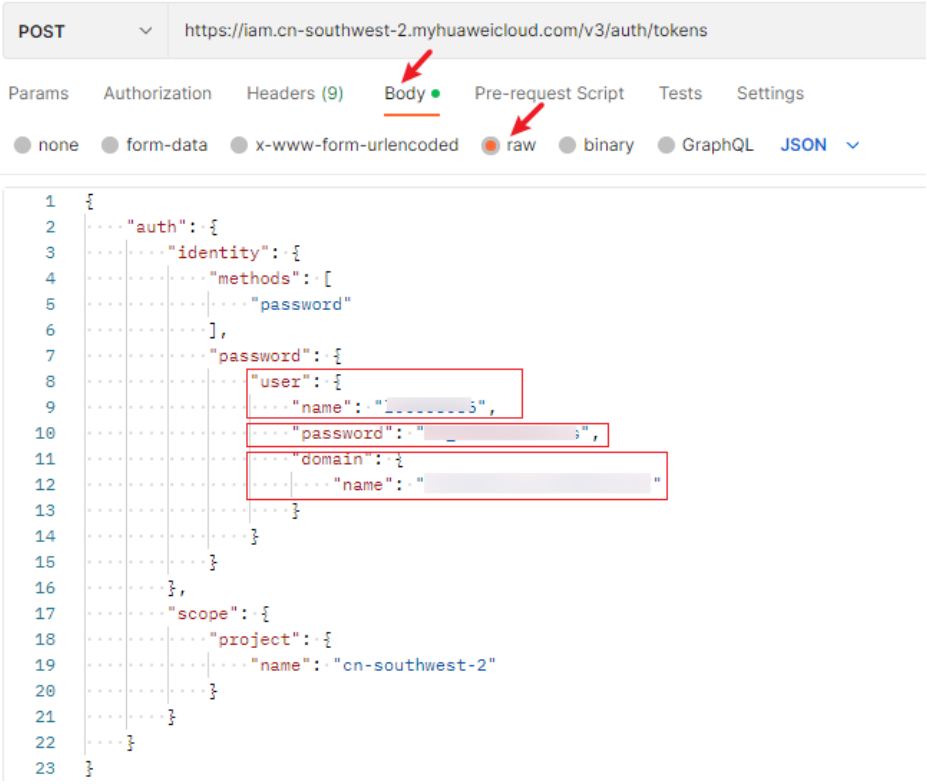
图 2-3 填写获取 Token 接口



4. 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。

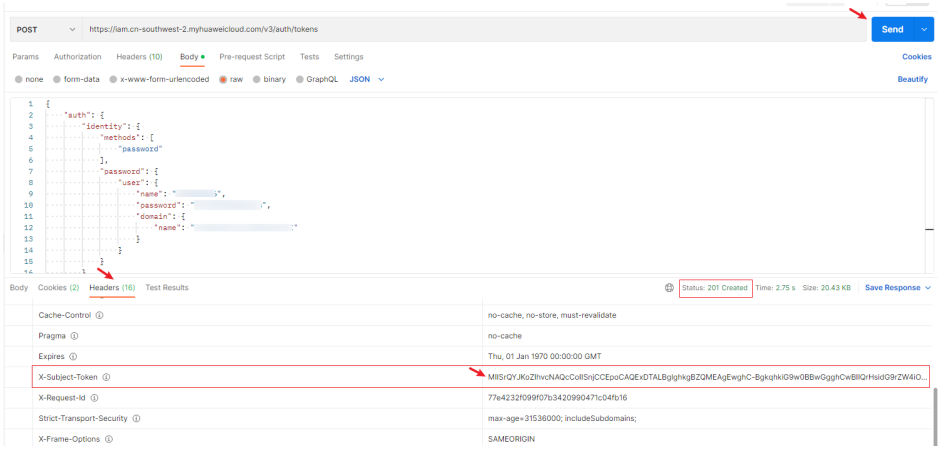
```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username", //IAM用户名
                    "password": "********", //华为云账号密码
                    "domain": {
    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



5. 单击Postman界面的“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。单击“Headers”选项，复制“X-Subject-Token”参数对应的值，该值即为获取的Token。

图 2-5 获取 Token



3 调用盘古大模型 API

用户可以通过API调用盘古大模型服务的基模型以及用户训练后的模型。训练后的模型只有在使用“在线部署”功能时，才可以使用本章节提供的方法进行调用。本章节将介绍如何使用Postman调用API，仅供测试使用。

前提条件

使用API调用模型前，请先完成盘古大模型服务订购和开通操作。

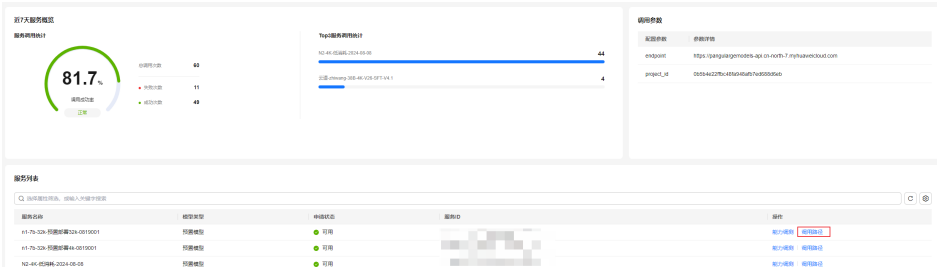
使用 Postman 调用 API

1. 获取API请求地址。
   a. 在“服务管理”页面，单击所需API的“查看详情”按钮。

图 3-1 服务管理

   b. 在“服务列表”中选择需要调用的模型，单击操作栏中的“调用路径”，复制对应模型的API请求地址。

图 3-2 获取 API 请求地址



2. 获取Token。

   在调用盘古API过程中，Token起到了身份验证和权限管理的作用。

   在调用盘古API前，需要先使用“获取Token”接口，获取Token值，再将Token值传入盘古API的请求header参数中，实现盘古服务在接收到用户的API请求时进行身份验证。

   说明

   关于Token有效期的详细说明请参见**获取IAM用户Token（使用密码）**。

   获取token步骤如下：

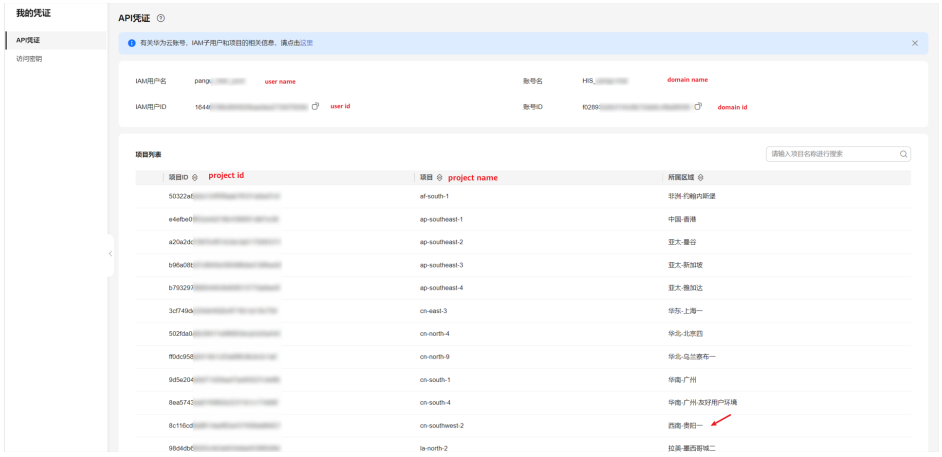
   a. 登录**“我的凭证 > API凭证”**页面，获取user name、domain name、project id。

      project id参数需要与盘古服务部署区域一致。例如，盘古大模型部署在“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 3-3 查看盘古服务区域

      图 3-4 获取 user name、domain name、project id

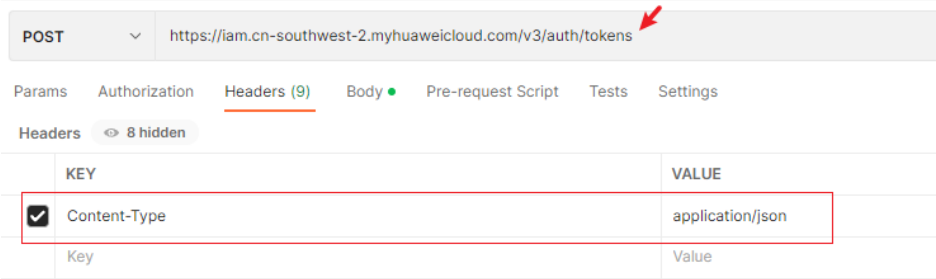
      

   b. 下载并安装**Postman**调测工具。

   c. 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

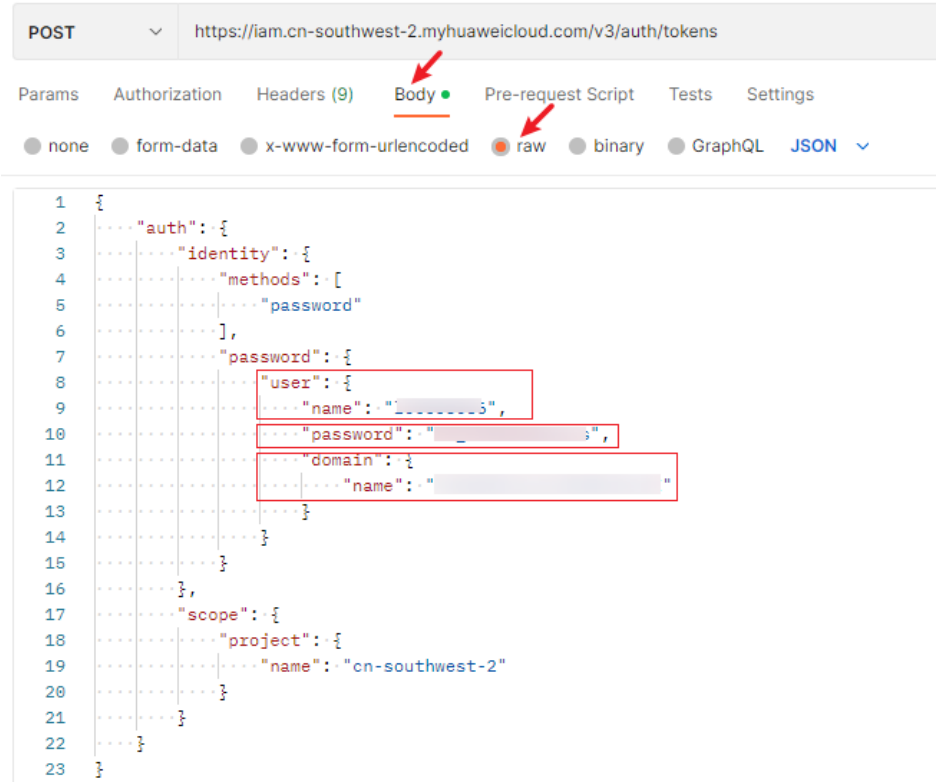
图 3-5 填写获取 Token 接口



d. 填写“获取token”接口的请求体。在Postman中选择“Body > raw”选项，参考图3-6复制并填入以下代码，并填写user name、domain name、password。

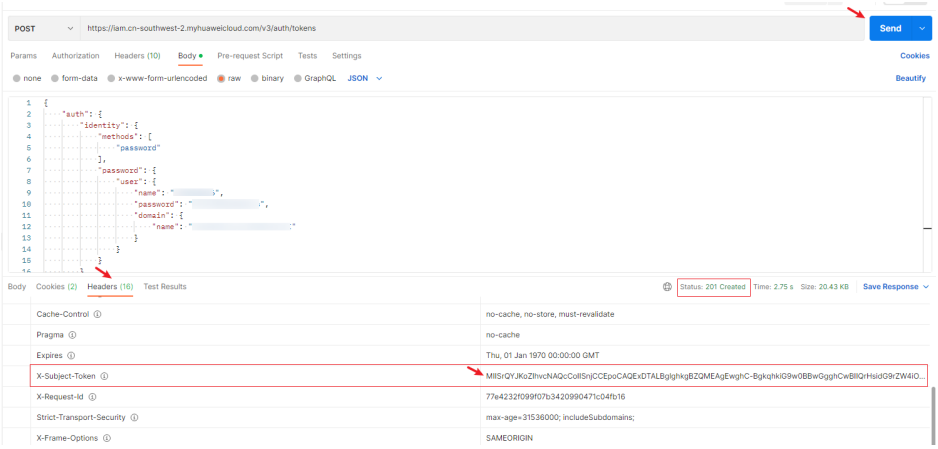
```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username", //IAM用户名
                    "password": "********", //华为云账号密码
                    "domain": {
                        "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 3-6 填写请求 Body



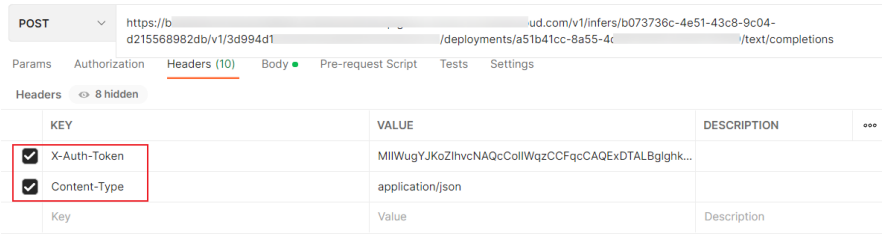
e. 单击Postman界面“Send”按钮，发送请求。当接口返回状态为201时，表示Token接口调用成功，此时单击“Headers”选项，找到并复制“X-Subject-Token”参数对应的值，该值即为需要获取的Token。

图 3-7 获取 Token



3. 调用盘古API。
   a. 在Postman中新建POST请求，并填入盘古API请求地址。
   b. 参考图3-8填写2个请求Header参数。
      - 参数名为Content-Type，参数值为application/json。
      - 参数名为X-Auth-Token，参数值为**获取Token**中获取的Token值。

图 3-8 填写盘古 API



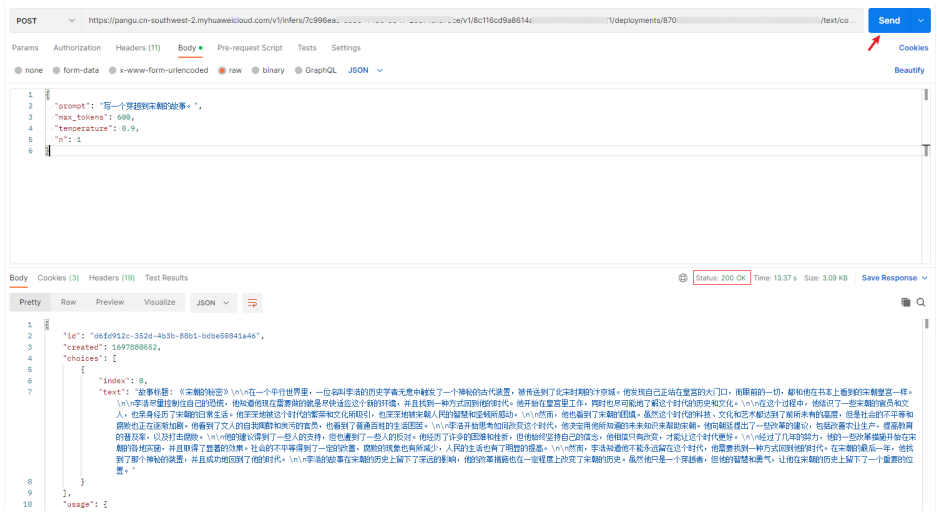
c. 在Postman中选择“Body > raw”选项，参考图3-9复制并填入以下代码，填写请求Body。

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

图 3-9 填写盘古请求 Body

d. 单击Postman界面的“Send”按钮，发送请求。当接口返回状态为200时，表示盘古API调用成功，并可在Postman中看到接口的返回信息。

图 3-10 获取盘古 API 调用结果



说明

- 使用Postman调用API时，如果出现SSL证书无效相关的报错，如“self signed certificate”（自签名证书）、“certificate has expired”（证书已过期）或“unable to verify the first certificate”（无法验证第一个证书）等。可以在Postman的设置中关闭“SSL certificate verification”选项。
- 关于盘古大模型API的详细请求参数、响应参数介绍请参见《API参考》文档。

4 启用盘古大模型搜索增强能力

大模型在训练时使用的是静态的文本数据集，这些数据集通常是包含了截止到某一时间点的所有数据。因此，对于该时间点之后的信息，大模型可能无法提供。

通过将大模型与盘古搜索结合，可以有效解决数据的时效性问题。当用户提出问题时，模型先通过搜索引擎获取最新的信息，并将这些信息整合到大模型生成的答案中，从而提供既准确又及时的答案。

1. 登录**盘古大模型套件平台**，在左侧导航栏中选择“能力调测”。
2. 单击“多轮对话”页签，选择使用N2系列模型，在页面右侧“参数设置”中可以开启搜索增强功能。

图 4-1 体验搜索增强能力