盘古大模型

推理服务 SDK

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 盘古推理 SDK 简介

推理 SDK 概述

盘古大模型推理SDK是对REST API进行的封装，通过该SDK可以处理用户的输入，生成模型的回复，从而实现自然流畅的对话体验。

表 1-1 推理 SDK 清单

SDK分类	SDK功能	支持语言	使用场景
推理SDK	对话问答（多轮对话）（/chat/completions）	Java、Python、Go、.NET、NodeJs	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。
	通用文本（文本补全）（/text/completions）	Java、Python、Go、.NET、NodeJs	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，还可以返回每个位置上不同词语的概率。它可以用来做文本生成、自动写作、代码补全等任务。

开发环境要求

华为云盘古大模型推理SDK要求：

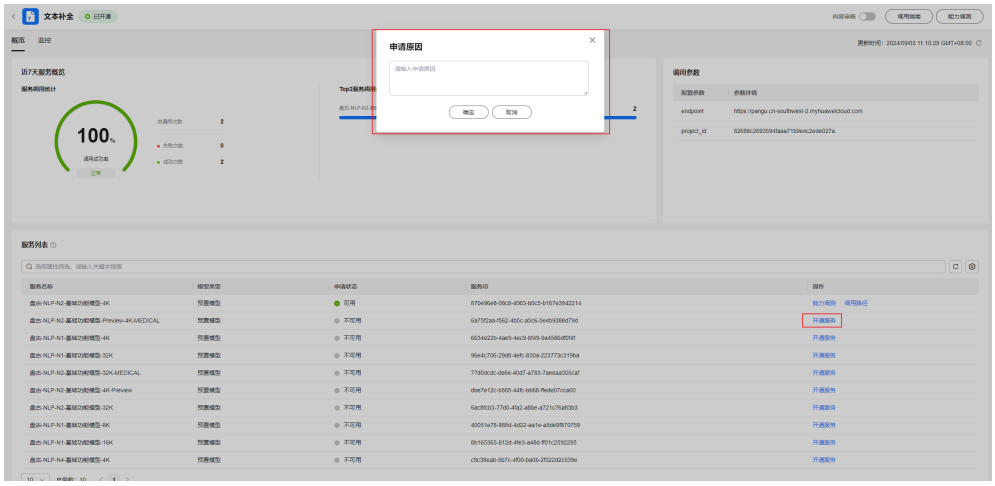
- Java SDK适用于JDK 1.8及其以上版本。
- Python SDK适用于Python3及以上版本。
- Go SDK支持go 1.14及以上版本。
- .NET SDK适用于.NET Standard 2.0及其以上版本；C# 4.0及其以上版本。
- NodeJs SDK适用于Node 10.16.1及其以上版本。

2 准备工作

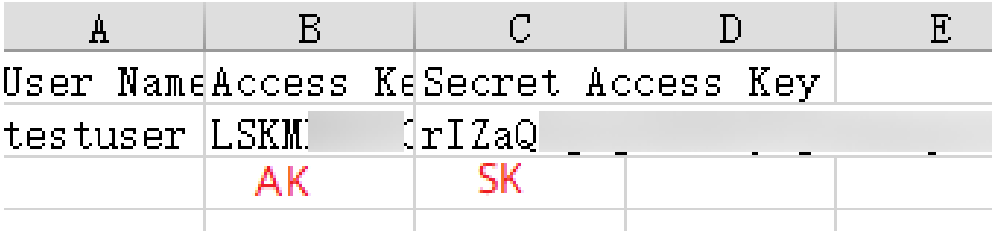
- 注册华为账号并开通华为云，并完成实名认证，账号不能处于欠费或冻结状态。
- 检查**开发环境要求**，确认本地已具备开发环境。
- 开通盘古大模型API。
  a. 登录**盘古大模型套件平台**。
  b. 在左侧导航栏中选择“服务管理”，在相应服务的操作列单击“查看详情”，可在服务列表中申请需要开通的服务。
     - 通用文本（文本补全）：文本补全接口提供单轮文本能力，常用于文本生成、文本摘要、闭卷问答等任务。
     - 对话问答（多轮对话）：多轮对话接口提供多轮文本能力，常用于多轮对话、聊天任务。

图 2-1 服务管理

图 2-2 申请开通服务



- 登录“**我的凭证** > 访问密钥”页面，依据界面操作指引获取Access Key（AK）和Secret Access Key（SK）。下载的访问密钥为credentials.csv文件，包含AK/SK信息。



说明

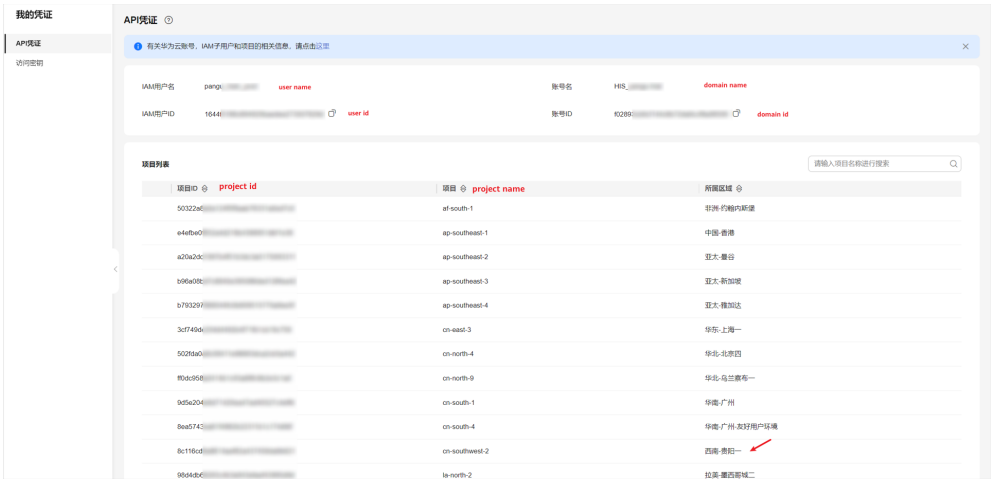
认证用的ak和sk硬编码到代码中或者明文存储都有很大的安全风险，建议在配置文件或者环境变量中密文存放，使用时解密，确保安全。

使用推理SDK章节示例代码均以ak和sk保存在环境变量中来实现身份验证。

- 登录“**我的凭证**”页面，获取“IAM用户名”、“账号名”以及待使用区域的“项目ID”。调用服务时会用到这些信息，请提前保存。

  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 2-3 获取 user name、domain name、project id



3 使用推理 SDK

安装 SDK

使用SDK前，需要安装“huaweicloud-sdk-core”和“huaweicloud-sdk-pangulargemodels”。

请在**SDK中心**获取最新的sdk包版本，替换示例中版本。

表 3-1 安装推理 SDK

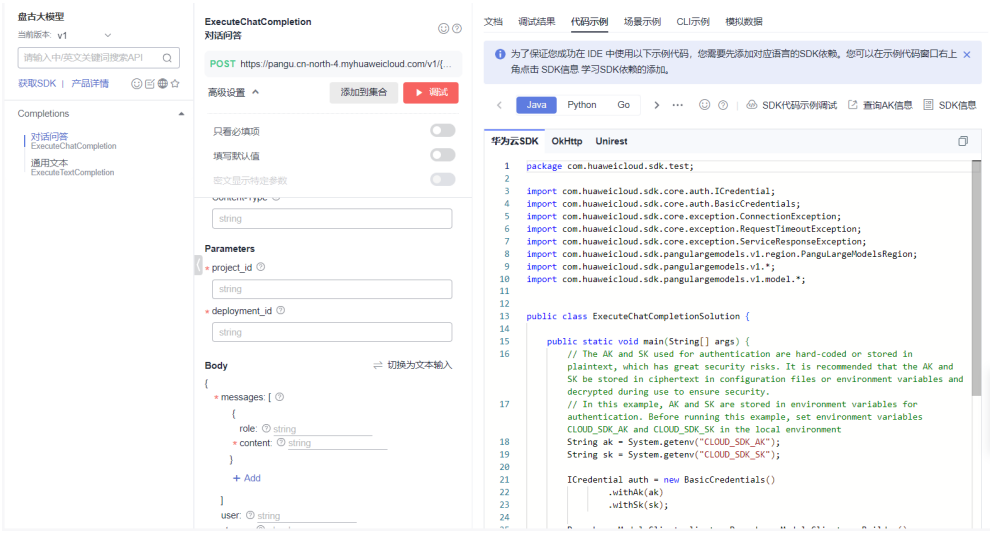
SDK语言	安装方法
Java	在您的操作系统中**下载**并**安装Maven**，安装完成后您只需要在Java项目的pom.xml文件中加入相应的依赖项即可。 <dependency> <groupId>com.huaweicloud.sdk</groupId> <artifactId>huaweicloud-sdk-core</artifactId> <version>3.1.103</version> </dependency> <dependency> <groupId>com.huaweicloud.sdk</groupId> <artifactId>huaweicloud-sdk-pangulargemodels</artifactId> <version>3.1.103</version> </dependency>
Python	使用pip安装。 #回显Successfully installed xxx表示安装成功 # 安装核心库 pip install huaweicloudsdkcore # 安装盘古服务库 pip install huaweicloudsdkpangulargemodels
Go	安装华为云Go SDK库。 // 安装华为云 Go SDK 库 go get -u github.com/huaweicloud/huaweicloud-sdk-go-v3
.NET	安装.NET SDK库 dotnet add package HuaweiCloud.SDK.Core dotnet add package HuaweiCloud.SDK.PanguLargeModels
NodeJs	安装NodeJs库 npm install @huaweicloud/huaweicloud-sdk-core npm i @huaweicloud/huaweicloud-sdk-pangulargemodels

在线生成 SDK 代码

API Explorer可根据需要动态生成SDK代码功能，降低您使用SDK的难度，推荐使用。

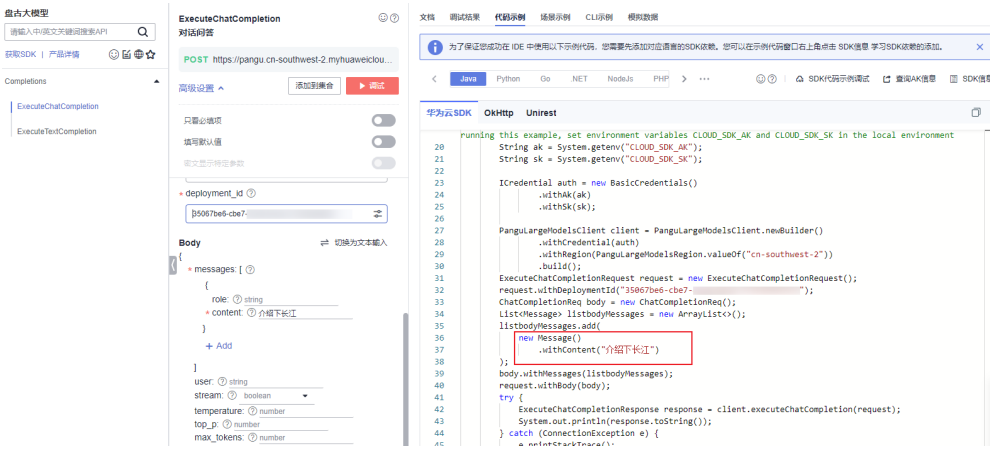
您可以在API Explorer中具体API页面的“代码示例”页签查看对应编程语言类型的SDK代码。

图 3-1 获取 SDK 代码示例



当您在中间填充栏填入对应内容时， 右侧代码示例会自动完成参数的组装。

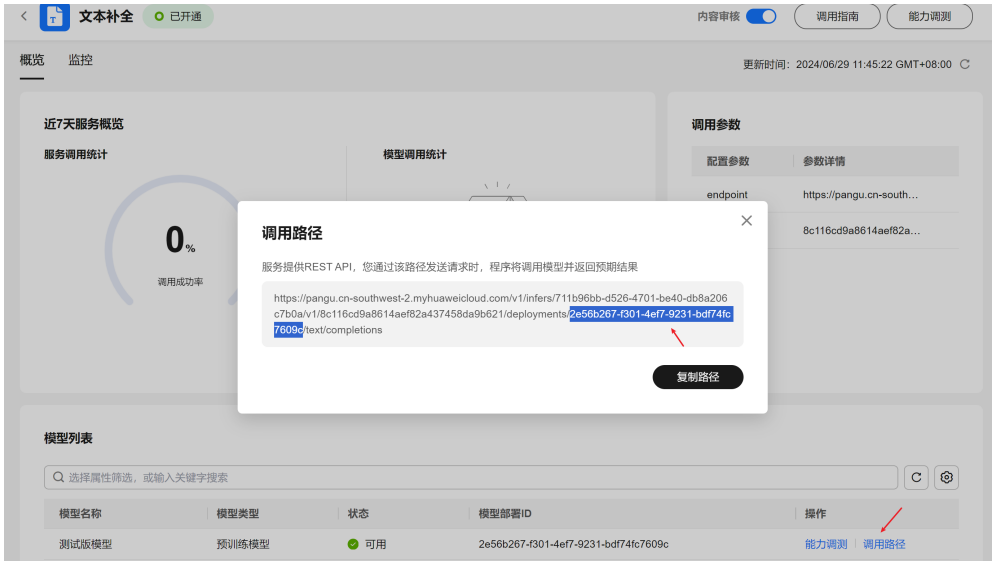
图 3-2 设置输入参数



填写输入参数时，deployment_id为模型部署ID，可以在盘古大模型套件平台“服务管理”功能中获取。

图 3-3 服务管理

图 3-4 获取 deployment_id



4 常见问题

使用 java sdk 出现第三方库冲突

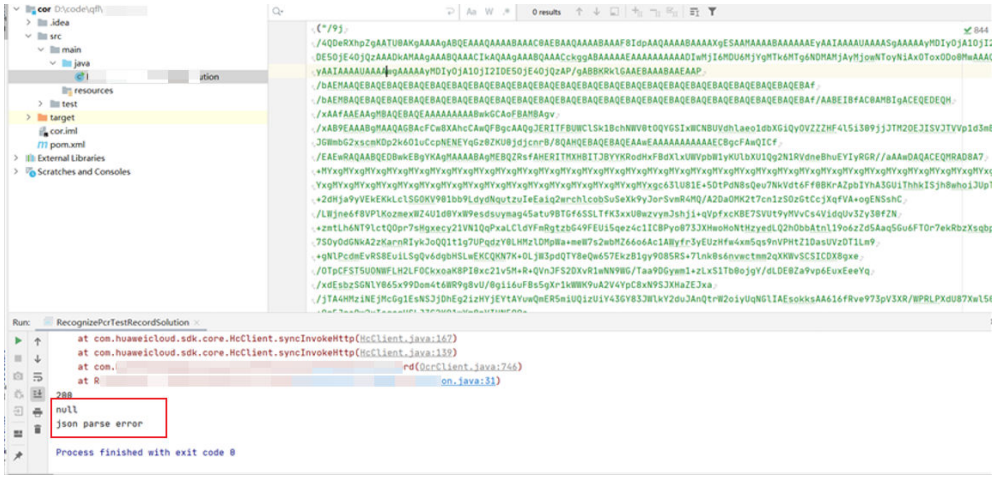
当出现第三方库冲突的时，如Jackson，okhttp3版本冲突等。可以引入如下bundle包(3.0.40-rc版本后)，该包包含所有支持的服务和重定向了SDK依赖的第三方软件，避免和业务自身依赖的库产生冲突：

```
<dependency>
    <groupId>com.huaweicloud.sdk</groupId>
    <artifactId>huaweicloud-sdk-bundle</artifactId>
    <version>[3.0.40-rc, 3.1.0)</version>
</dependency>
```

jackson版本要求请见**pom.xml**。

使用 java sdk 出现 json 解析报错

图 4-1 json 解析报错



- 服务端返回的数据格式不符合json格式，导致sdk侧解析json数据报错。
- 服务端返回的json数据不符合json反序列化的规则，和sdk定义的数据结构不一致，导致反序列化失败。
- sdk json数据解析问题。

- 建议排查服务端返回的数据是否和服务SDK设计的结构、字段一致。

SDK 运行报错

- **java.lang.NoClassDefFoundError: Could not initialize class com.huaweicloud.sdk.core.http.HttpConfig at com.huaweicloud.sdk.core.ClientBuilder.build(ClientBuilder.java:98)**

  HttpConfig这个类在sdk-core包里面找不到，造成原因为用户使用的sdk版本太老导致，建议使用**最新版本**的华为云java sdk，运行代码再具体定位。

- **java.lang.NoSuchFieldError: ALLOW_LEADING_DECIMAL_POINT_FOR_NUMBERS**

  这个字段是jackson-core里面用来标识解析json格式数据是否支持前导小数点的字段，这个报错的意思是找不到这个字段，很可能是因为用户使用的jackson版本太老导致。

  建议客户本地将jackson版本升级到和华为云java sdk一致，jackson版本要求请见**pom.xml**。

  引用华为云java sdk的**bundle包**来解决jackson版本冲突的问题。

  ```
  <dependency>
     <groupId>com.huaweicloud.sdk</groupId>
     <artifactId>huaweicloud-sdk-bundle</artifactId>
     <version>[3.0.40-rc, 3.1.0)</version>
  </dependency>
  ```

- **java.lang.ClassNotFoundException: com.fasterxml.jackson.datatype.jsr310.JavaTimeModule**

  用户本地工程引入了jackson框架，和华为云sdk引入的jackson框架冲突了，导致会报找不到某个类，建议客户在本地引入**bundle包**报来避免出现依赖冲突。

  ```
  <dependency>
     <groupId>com.huaweicloud.sdk</groupId>
     <artifactId>huaweicloud-sdk-bundle</artifactId>
     <version>[3.0.40-rc, 3.1.0)</version>
  </dependency>
  ```

- **java.lang.ClassNotFoundException: okhttp3/Interceptor**

  用户本地引入的Okhttp3版本和华为云冲突，okhttp版本要求请见**pom.xml**。