

盘古大模型
3.0.0

SDK 参考

文档版本 01
发布日期 2024-12-02



版权所有 © 华为技术有限公司 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。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://www.huawei.com>

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：

<https://www.huawei.com/cn/psirt/vul-response-process>

如企业客户须获取漏洞信息，请参见如下网址：

<https://securitybulletin.huawei.com/enterprise/cn/security-advisory>

目录

1 盘古推理 SDK 简介.....	1
2 准备工作.....	2
3 使用推理 SDK.....	4
4 常见问题.....	7

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 (/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 (/text/completions)	Java、Python、Go、.NET、NodeJs、PHP	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。
- 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。



A	B	C	D	E
User Name	Access Key	Secret	Access Key	
testuser	LSKM	crIZaQ		
	AK	SK		

📖 说明

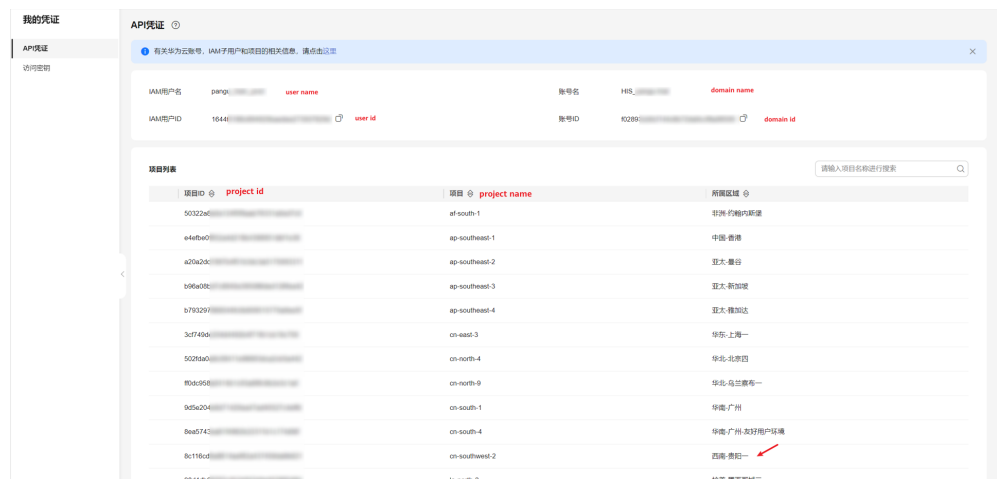
认证用的ak和sk硬编码到代码中或者明文存储都有很大的安全风险，建议在配置文件或者环境变量中密文存放，使用时解密，确保安全。

3 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 2-1 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

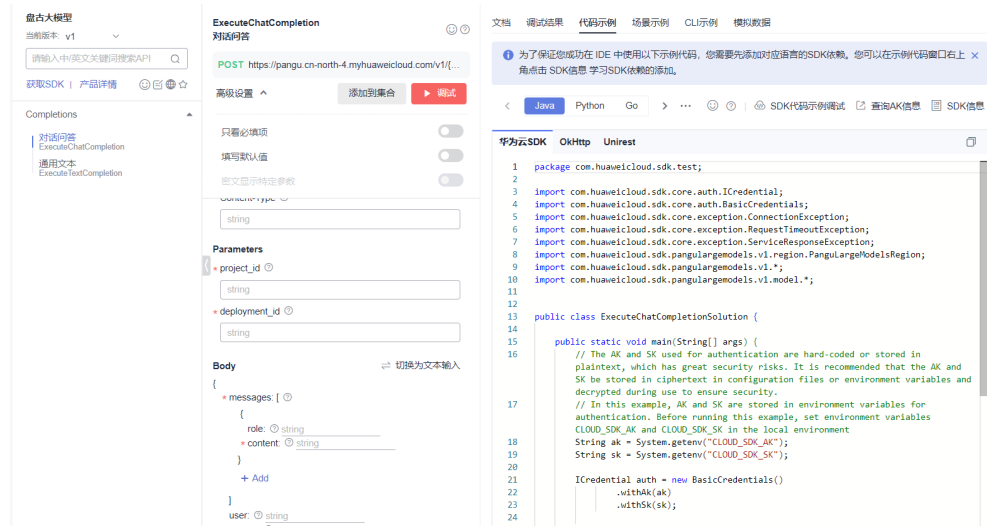
SDK语言	安装方法
Java	在您的操作系统中 下载并安装Maven ，安装完成后您只需要在Java项目的pom.xml文件中加入相应的依赖项即可。 <pre><dependency> <groupId>com.huaweicloud.sdk</groupId> <artifactId>huaweicloud-sdk-core</artifactId> <version>3.1.103</version> </dependency> <dependency> <groupId>com.huaweicloud.sdk</groupId> <artifactId>huaweicloud-sdk-pangulargemodels</artifactId> <version>3.1.103</version> </dependency></pre>
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。

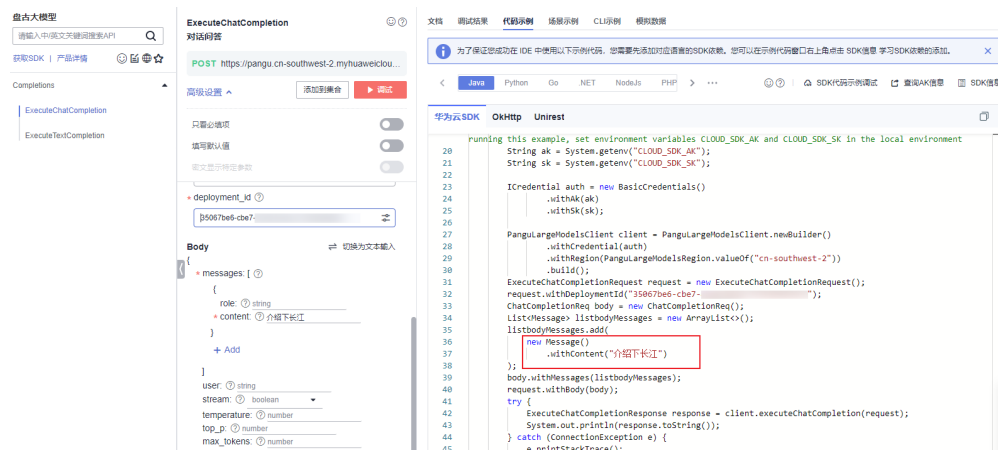
步骤1 您可以在API Explorer中具体API页面的“代码示例”页签查看对应编程语言类型的SDK代码。

图 3-1 获取 SDK 代码示例



步骤2 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。

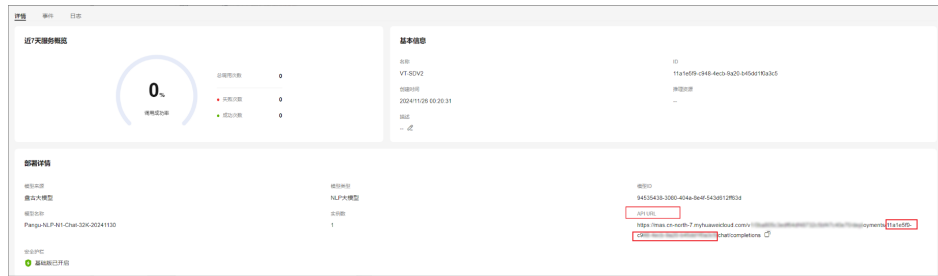
图 3-2 设置输入参数



步骤3 填写输入参数时，deployment_id为模型部署ID，获取方式如下：

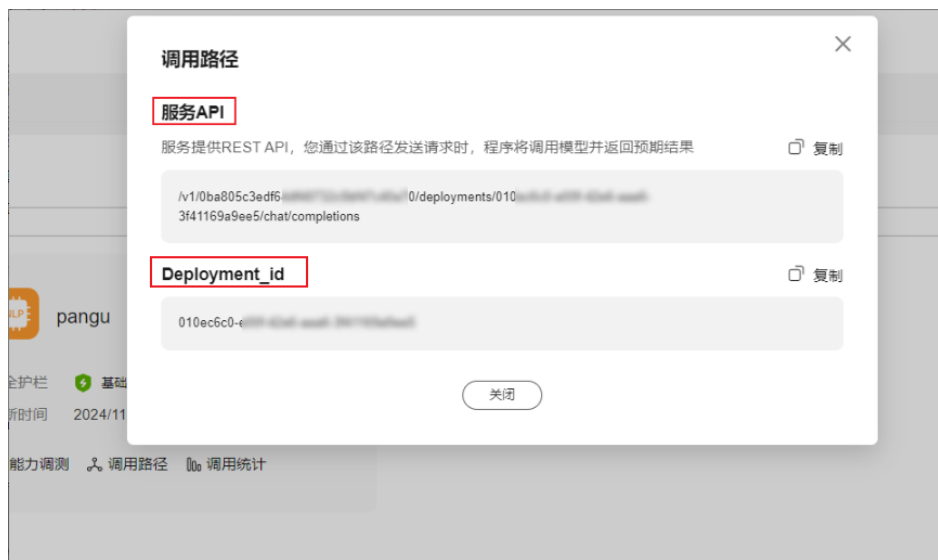
- 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的部署ID。

图 3-3 部署后的模型调用路径



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

图 3-4 预置模型的调用路径



----结束

- 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](#)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](#)。