2024 华为软件精英挑战赛

决赛阶段大模型服务 API 使用说明

1. 背景说明：

为确保贵重物品在运输过程中妥善处置（避免损坏并进而产生赔偿），在决赛阶段，当选手机器人拿取贵重物品时，需回答运输注意事项相关问题（考验机器人是否具备运输当前贵重物品的必备知识储备）。回答问题时，选手程序可以借助大赛平台提供的大模型服务接口来辅助进行问题回答。每个问题每个选手只有一次回答机会，回答成功后可顺利取货（回答错误后，无继续答题机会，此次选手应尽快选择其他目标搬运货物，将该货物预留给其他选手程序进行搬运），每位选手在同一货物处只能使用一次取货命令（会忽略除第一次外在该货物的取货命令）。（详情请见任务书 4.2.7 节）。

赛题组为决赛的每一支参赛队伍提供了 1 个独立部署的智能问答大模型服务（独占 2 个昇腾 NPU 资源），选手可使用 HTTP 请求的方式进行访问。选手可在赛题程序中对大模型 API 进行访问，利用大模型的智能问答能力辅助答题，从而获得贵重物品搬运权。

2. API 信息：

- 接口地址：https://example/infer/v1/xxx （邮件方式发送给各个选手）
- 请求方式：POST
- Header 信息：

字段名称	字段值	说明
X-Apig-AppCode	xxxxxxx	各个选手团队身份认证token (各个选手查收邮件方式获取)。
Content-Type	application/json	请求中指定的请求体类型

- 请求参数：

参数名称	参数类型	参数说明
prompt	字符串	对大模型提问交互的内容
temperature	浮点数，默认值	控制采样的随机性的浮点数。值越大不确定性越高。0表示贪婪采

	1.0	样。
top_k	整数，默认值-1	控制模型推理时要考虑的前几个tokens的数量的整数。设置为-1表示考虑所有tokens
top_p	浮点数，默认值1.0	控制要考虑的前几个tokens的累积概率的浮点数。必须在 (0, 1] 范围内。设置为1表示考虑所有tokens

响应状态：

状态码	说明	可能原因
200	OK	请求成功
401	Unauthorized	未在header中传入X-Apig-AppCode或无效
404	Not Found	请求地址不正确
503	Service Unavailable	服务内部错误，可能是请求参数类型错误或者缺少必要的请求参数

赛题组给参赛选手发出的邮件中，包含：

1. API 请求地址
2. 用于身份认证的 X-Apig-AppCode

以上两个关键信息专属于一支参赛队伍，请妥善保管，务必不要上传至公开代码仓，或者外发给非本队的参赛选手与赛事无关人员。**任何因为选手原因泄露 API 访问信息，导致比赛过程中的 API 性能下降甚至不可用问题，由选手自行承担。**

每个大模型 API 使用了 2 块昇腾 NPU 芯片资源独立部署，不同 API 之间互不干扰。大模型服务不支持历史对话，每一次访问在内容上都将是独立的。API 使用非流式输出，一次请求将会在模型推理完成所有内容后，一次性返回结果。

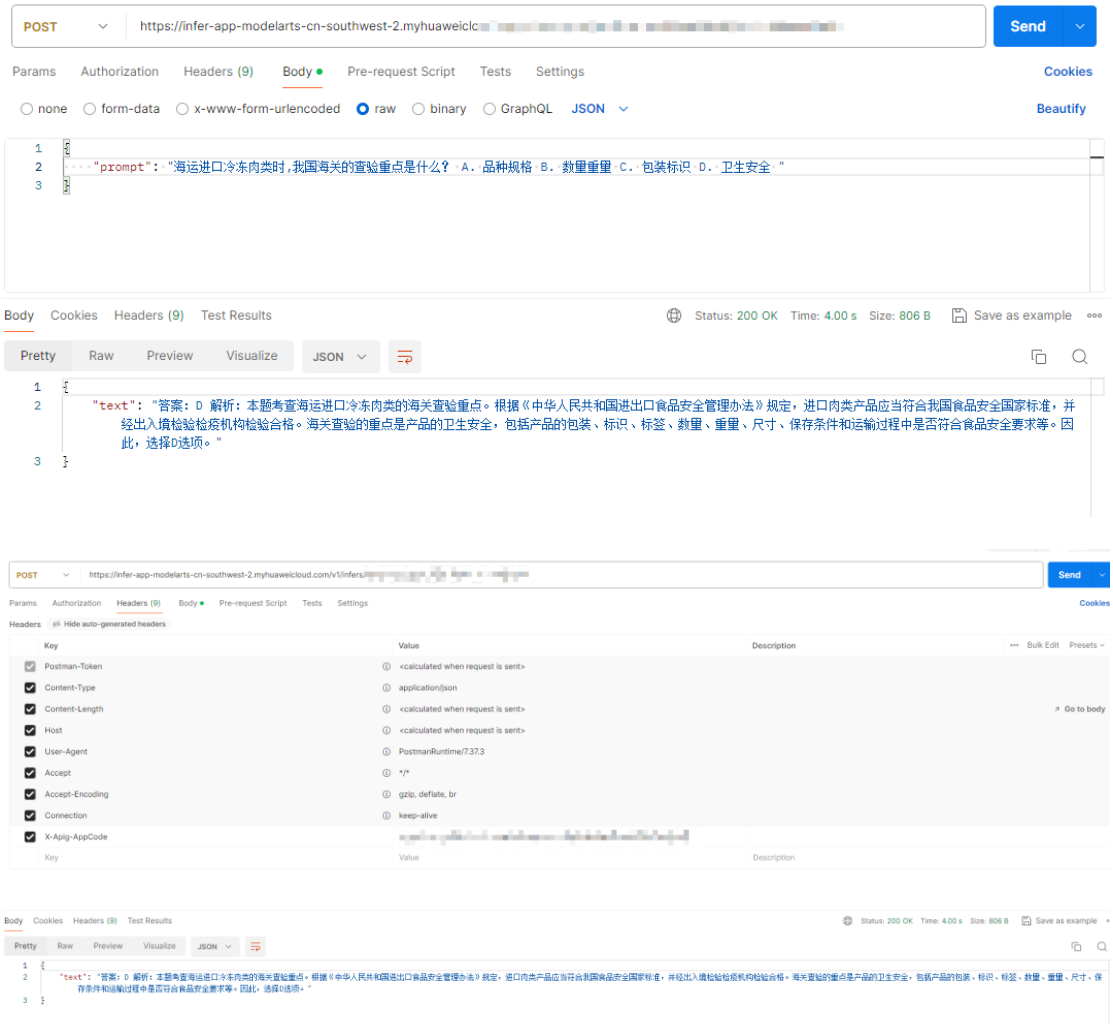
3. 注意事项：

1. 由于赛题中的问答题内容均是中文，为了避免中文乱码问题，建议对控制台输入使用 UTF-8 编码，在 HTTP 请求时也指定 UTF-8 编码，可参考本地版赛题组件中的 sdk。
2. 为防止您的专属大模型 API 受到攻击，请妥善保管好发送给您的接口请求地址与 X-Apig-AppCode（身份认证 token）。请注意：各个团队的 LLM 服务接口访问链接各不相同，请各团队务必妥善保管防止泄露。因选手自身原因保存不当所带来的接口访问异常或性能下降情况，选手自行承担相关后果。

3. 赛题组提供的决赛大模型 API 仅可用于决赛阶段赛题的智能问答，禁止用于赛事无关场景，一旦发现恶意使用或主动攻击后台服务行为，赛题组将取消参赛选手的大模型服务使用权限。
4. 解题程序中只允许使用赛题组提供的大模型 API，禁止使用外部 API。
5. 在使用程序调用大模型 API 时，请考虑可能的 HTTP 请求失败的异常情况，从而避免赛题程序中断或崩溃。
6. 如果请求时使用了错误的参数类型，可能导致返回 Internal Server Error。
7. 练习赛阶段选手在本地调试时的网络环境质量可能会小幅影响接口响应时间。

4. 请求样例：

a） 在 PC 上使用 PostMan 对大模型 API 发送请求



除此之外，赛题组还提供了 Python, Java, C++三种语言的 HTTP 请求示例代码。请关注赛题发布帖。在判题环境中，也提供了三种语言用于 HTTP 请求的必要的三方库（详情见决赛编译说明文档）。

b） 在程序中大模型 API 发送请求时可用的三方库版本与安装方式

Python:

```
pip install requests==2.31.0
```

Java:

推荐使用 Maven 进行项目构建，可将下面依赖添加到 pom.xml 中

```
<dependency>
    <groupId>org.apache.httpcomponents</groupId>
    <artifactId>httpclient</artifactId>
    <version>4.5.14</version>
</dependency>
```

Cpp:

本地环境调试时，需要安装 libcurl 库（官网：https://curl.se/download/），下载压缩包并解压到赛题目录以外的位置。

- **在 Linux/MacOS 你可以使用如下指令进行编译安装：**

```
./configure --with-ssl=/usr/local/curl
make
make install
```

- 使用 g++编译时，在命令中增加`-lcurl`
- 使用 cmake 编译时，在 CMakeLists 中增加`find_package(CURL REQUIRED)`

- **在 Windows 环境推荐直接下载源码包，配置到系统环境变量中：**

- 使用 g++编译时，在命令中增加`-lcurl -IC:/curl/include -LC:/curl/lib`
- 使用 cmake 编译时，在 CMakeLists 中增加，在 cmake 命令中增加`-DCURL_DIR=C:\curl` 指定 curl 所在路径 上述 C:\curl、C:/curl/include、C:/curl/lib 需要替换为 curl 源码所在的实际路径

```
find_path(CURL_INCLUDE_DIR NAMES curl/curl.h PATHS
${CURL_DIR}/include)
find_library(CURL_LIBRARY NAMES libcurl PATHS ${CURL_DIR}/lib)
include_directories(${CURL_INCLUDE_DIR})
target_link_libraries(main ${CURL_LIBRARY})
```

请注意：线上判题平台的 CMakeLists 中已内置上述参数，在进行线上判题时无需上传本地 CMakeLists 文件

最后，需要在主程序中增加如下头文件

```
#include <curl/curl.h>
```