闭源数据翻译模型 API 文档

概述

本文档提供了本地部署的闭源翻译模型的 API 调用指南，包括通用翻译模型。翻译模型支持的语言如下。

语系	语种&方言
印欧语系	英语、法语、葡萄牙语、德语、罗马尼亚语、瑞典语、丹麦语、保加利亚语、俄语、捷克语、希腊语、乌克兰语、西班牙语、荷兰语、斯洛伐克语、克罗地亚语、波兰语、立陶宛语、挪威语（博克马尔语）、挪威尼诺斯克语、波斯语、斯洛文尼亚语、古吉拉特语、拉脱维亚语、意大利语、奥克语、尼泊尔语、马拉地语、白俄罗斯语、塞尔维亚语、卢森堡语、威尼斯语、阿萨姆语、威尔士语、西里西亚语、阿斯图里亚语、恰蒂斯加尔语、阿瓦德语、迈蒂利语、博杰普尔语、信德语、爱尔兰语、法罗语、印地语、旁遮普语、孟加拉语、奥里雅语、塔吉克语、东意第绪语、伦巴第语、利古里亚语、西西里语、弗留利语、撒丁岛语、加利西亚语、加泰罗尼亚语、冰岛语、托斯克语、阿尔巴尼亚语、林堡语、罗马尼亚语、达里语、南非荷兰语、马其顿语僧伽罗语、乌尔都语、马加希语、波斯尼亚语、亚美尼亚语
汉藏语系	中文（简体中文、繁体中文、粤语）、缅甸语
亚非语系	阿拉伯语（标准语、内志语、黎凡特语、埃及语、摩洛哥语、美索不达米亚语、塔伊兹-阿德尼语、突尼斯语）、希伯来语、马耳他语
南岛语系	印度尼西亚语、马来语、他加禄语、宿务语、爪哇语、巽他语、米南加保语、巴厘岛语、班加语、邦阿西楠语、伊洛科语、瓦雷语（菲律宾）
德拉威语	泰米尔语、泰卢固语、卡纳达语、马拉雅拉姆语
突厥语系	土耳其语、北阿塞拜疆语、北乌兹别克语、哈萨克语、巴什基尔语、鞑靼语
壮侗语系	泰语、老挝语
乌拉尔语系	芬兰语、爱沙尼亚语、匈牙利语
南亚语系	越南语、高棉语
其他	日语、韩语、格鲁吉亚语、巴斯克语、海地语、帕皮阿门托语、卡布维尔迪亚努语、托克皮辛语、斯瓦希里语

基本信息

服务地址

- 翻译模型地址：`http://172.32.1.162:9000/v1`

认证方式

所有 API 请求需要在请求头中包含 `Authorization` 字段，格式为：

```python
Authorization: Bearer token-abc123
```

模型类型	模型名	适用场景
通用翻译模型	`model/translate`	适用于大多数语言场景的通用翻译任务

响应格式

所有 API 响应将以 JSON 格式返回。

API 端点

创建翻译请求 (Translation Request)

URL: `http://172.32.1.162:9000/v1`

Method: `POST`

请求参数:

参数名	类型	是否必需	描述
model	string	是	模型名称，`/model/translate`
messages	array	是	聊天消息数组，每个消息包含 `role` (system/user/assistant) 和 `content`
temperature	float	否	控制随机性，范围 0-2，默认 0.7
top_p	float	否	核采样参数，范围 0-1，默认 1
n	integer	否	每个提示生成的回复数量，默认 1
max_tokens	integer	否	最大生成令牌数
stream	boolean	否	是否流式返回，默认 false
enable_thinking	boolean	否	是否开启 think 模式，默认 true

参数名	类型	是否必需	描述
include_usage	boolean	否	是否在最后一个流式块返回 usage 信息

示例请求:

PYTHON

```python
import os
from openai import OpenAI

client = OpenAI(
    base_url="http://172.32.1.162:9000/v1",
    api_key="token-abc123",
)

response = client.chat.completions.create(
    model="/model/translate",
    messages=[
        {"role": "user", "content": """
        the translation content.
         """}

    ],
    stream=True,
    extra_body={"chat_template_kwargs": {"enable_thinking": False}}
)

for chunk in response:
    print(chunk.choices[0].delta.content, end="", flush=True)
```

响应参数:

参数名	类型	描述
id	string	响应唯一标识符
object	string	响应对象类型，固定为 "chat. Completion"
created	integer	创建时间戳
model	string	使用的模型名称
choices	array	生成的回复数组
index	integer	回复索引

参数名	类型	描述
message	object	消息对象
role	string	消息角色 (assistant)
content	string	消息内容
finish_reason	string	完成原因 (stop/length/content_filter)
usage	object	令牌使用统计
prompt_tokens	integer	提示令牌数
completion_tokens	integer	生成令牌数
total_tokens	integer	总令牌数

示例响应:

```
{
    "id": "chatcmpl-123456",
    "object": "chat.completion",
    "created": 1693764729,
    "model": "/model/translate",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "翻译内容"
            },
            "finish_reason": "stop"
        }
    ],
    "usage": {
        "prompt_tokens": 15,
        "completion_tokens": 20,
        "total_tokens": 35
    }
}
```

错误处理

API 可能返回以下错误码:

错误码	描述
400	错误请求，检查参数格式和必填项
401	未授权，检查 API 密钥
404	请求的资源不存在
500	服务器内部错误，请稍后重试

错误响应示例：

```
{
    "error": {
        "message": "Invalid API key",
        "type": "authentication_error",
        "param": null,
        "code": "invalid_api_key"
    }
}
```

注意事项

1. 请根据具体需求选择合适的模型，以获得最佳效果。
2. 专用模型在特定领域的术语和上下文处理上具有优势。
3. 所有模型均在本地运行，确保服务器资源充足以保证性能。
4. API 密钥需要妥善保管，避免泄露。
5. 请勿用于非法或违反保密协议的数据处理。