QA--RAG训练营

## 课程基本信息

### 课程大纲

[RAG 训练营课程大纲](https://shimo.im/docs/47kgMeP42pUyyB3V)

### 课程代码

1. https://github.com/huangjia2019/rag-in-action
2. https://github.com/huangjia2019/rag-project01-framework
3. https://github.com/huangjia2019/rag-project02-medical-nlp-box
4. https://github.com/huangjia2019/rag-project03-audit

#### 2.1 课程环境安装

[Cursor ：新手如何快速配置 Python 环境](https://doc.weixin.qq.com/doc/w3_ADEAXwZXAKoWA54422rSfK7ZiUqtp?scode=ACQADQdwAAoT0hgOk5ADEAXwZXAKo)

[Cursor安装教程参考](https://zhuanlan.zhihu.com/p/16508727483)

#### 2.2环境安装遇到的问题文档可共同编辑：

[遇到的问题汇总（可共同编辑）](https://doc.weixin.qq.com/doc/w3_ADEAXwZXAKoCNjD4EUAT8QtOKTeVN?scode=ACQADQdwAAohCLdhFGADEAXwZXAKo&isEnterEdit=1)

### 问题提问

小伙伴

提问请详细描述问题现象、错误消息、代码示例、环境信息，还有你已经尝试过的解决方法。

建议自己先尝试大模型（ChatGPT，Claude，DeepSeek V3）提问

————————

如果使用大模型提问，可以使用如下提示语（引用彭老师课程内容）

#### 一个让AI回答更专业、更有深度的万能Prompt

适用于任意AI,特别是放在ChatGPT自定义(ChatGPT Custom Instructions)中来使用

https://github.com/DenisSergeevitch/chatgpt-custom-instructions

1. 前往 ChatGPT
2. 导航至“设置”
3. 选择个性化
4. 在自定义说明中输入这些说明

#### 问题池

#### 黄老师历史答疑文档

[【答疑文档】RAG1期](https://doc.weixin.qq.com/doc/w3_ADEAXwZXAKoTJuQqYDMR4uXsjCGkd?scode=ACQADQdwAAoxJYWt17ADEAXwZXAKo&isEnterEdit=1)

## 课程准备

### Cursor安装配置使用

[cursor安装配置使用](http://gk.link/a/12x9H)

### Github 加速

#### github Speed

<https://ghproxy.1888866.xyz/>

#### github高级搜索

<https://github.com/search/advanced>

### Hugging Face 代理

[HF-Mirror](https://hf-mirror.com/)

* 如何快速下载huggingface模型:<https://zhuanlan.zhihu.com/p/663712983>



### 课程中使用的OpenAI key上哪里获取

#### GPT 官方方式

* 注册 ChatGPT 账号👉<http://gk.link/a/126kX>
* 外国手机号的方法: <https://sms-activate.org/cn>
* OpenAI API KEY查看和创建<https://platform.openai.com/api-keys>
* ChatGPT Plus开通攻略 <https://github.com/anyofai/anyofai.github.io>
  + ye ka 注册邀请码： https://yeka.ai/i/OG4ZRNDS

##### 有了KEY, 没有办法访问OpenAI 接口怎么办

使用国内代OPENAI\_API\_KEY理api即可. 将域名替换即可. api.openai.com -> api.openai-proxy.com

风险提示: 由于是第三方代理, 因此可以作为测试学习用. !!!不要作为生产用以防key被盗用!!!!!!

代码方法使用如下

##### 已购买代理的socks5代理

设置为全局代理可以直接使用, 如果不是设置为全局代理, 查看你代理的端口, 诸如18080

先安装库pip install PySocks

在代码开头设置socks5代理

如何验证代理是否能连接上[api.openai.com](https://api.openai.com)

打得开: <https://chat.openai.com/> 即可认为可访问[api.openai.com](https://api.openai.com)

ps: 使用ping也可以不过不够准.推荐用上述访问网页的7层协议验证

#### GPT国内代理

* <https://www.apiyi.com/>（推荐）
  + <https://www.apiyi.com/register/?aff_code=we80>(我的邀请链接)
* <https://2233.ai/api>（推荐）
* <https://www.eylink.cn/>

如果是自己购买的三方openai代理, 诸如在<https://www.apiyi.com/>上购买的，可以用

如果是自己购买的三方openai代理, 诸如在https://www.eylink.cn/上购买的，可以用

如果是自己购买的三方openai代理, 诸如在<https://2233.ai/api>上购买的，可以用

##### 注意：

**API中转站支持微调吗？**

答：由于调用逻辑关系，微调只能创建，但因为中转用了很多个Key（背后是号池：很多个的api账号），

但是非本号的 key 调用不通的

**也就是中转站不支持 fine tune 微调，另外不支持 files文件 接口（ beta.assistants.create集成了文件接口也不支持 ）。**

但支持 补全/对话，和 函数调用 functions

也就是微调 API 只能买官网API充值的直连API

### OpenAI API可否使用国内的智谱或DeepSeek

可以，这是因为他们兼容OpenAI接口规范，因此我们其实仅需要下载一个openai的库即可

自行注册智谱或DeepSeek账号后，获取API KEY，代码中进行替换，例如

DeepSeek：<https://api-docs.deepseek.com/zh-cn/>

智谱：<https://www.bigmodel.cn/dev/api/normal-model/glm-4v>

* Github社群代a码教程：[GitHub - MetaGLM/glm-cookbook: Examples and guides for using the GLM APIs](https://github.com/MetaGLM/glm-cookbook)

注意：截止2025.1.20，deepseek 还没有支持embedding model

### 开发环境

#### 怎么安装, 怎么样才算是ready

#### 如何安装miniconda 进行python环境管理

A: 见: <https://docs.anaconda.com/free/miniconda/miniconda-install/>

安装完毕之后使用: conda list

验证是否安装完毕

#### Q: 如何安装指定环境python3.10

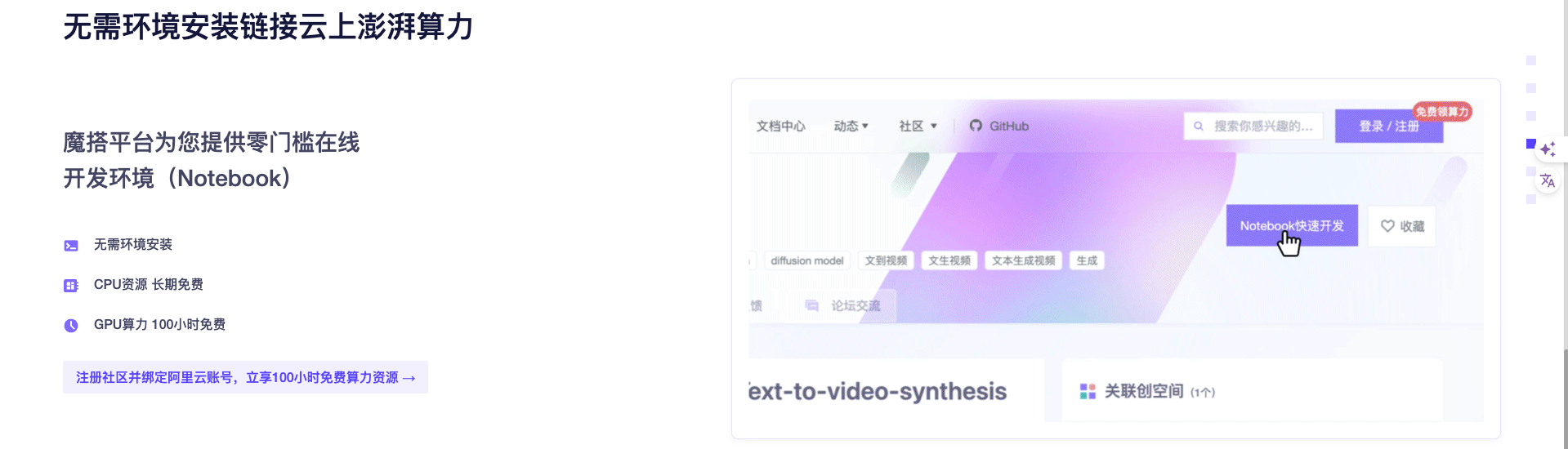
创建:conda create -n myenv python=3.10

激活: conda activate myenv

不使用时关闭环境 conda deactivate

#### Q: 有无免费可白嫖的GPU资源

* 国内: <https://modelscope.cn/home> (注册有100小时免费GPU)



* 国外: <https://colab.research.google.com/>
  + Colab使用教程（超级详细版）及Colab Pro/Pro+评测

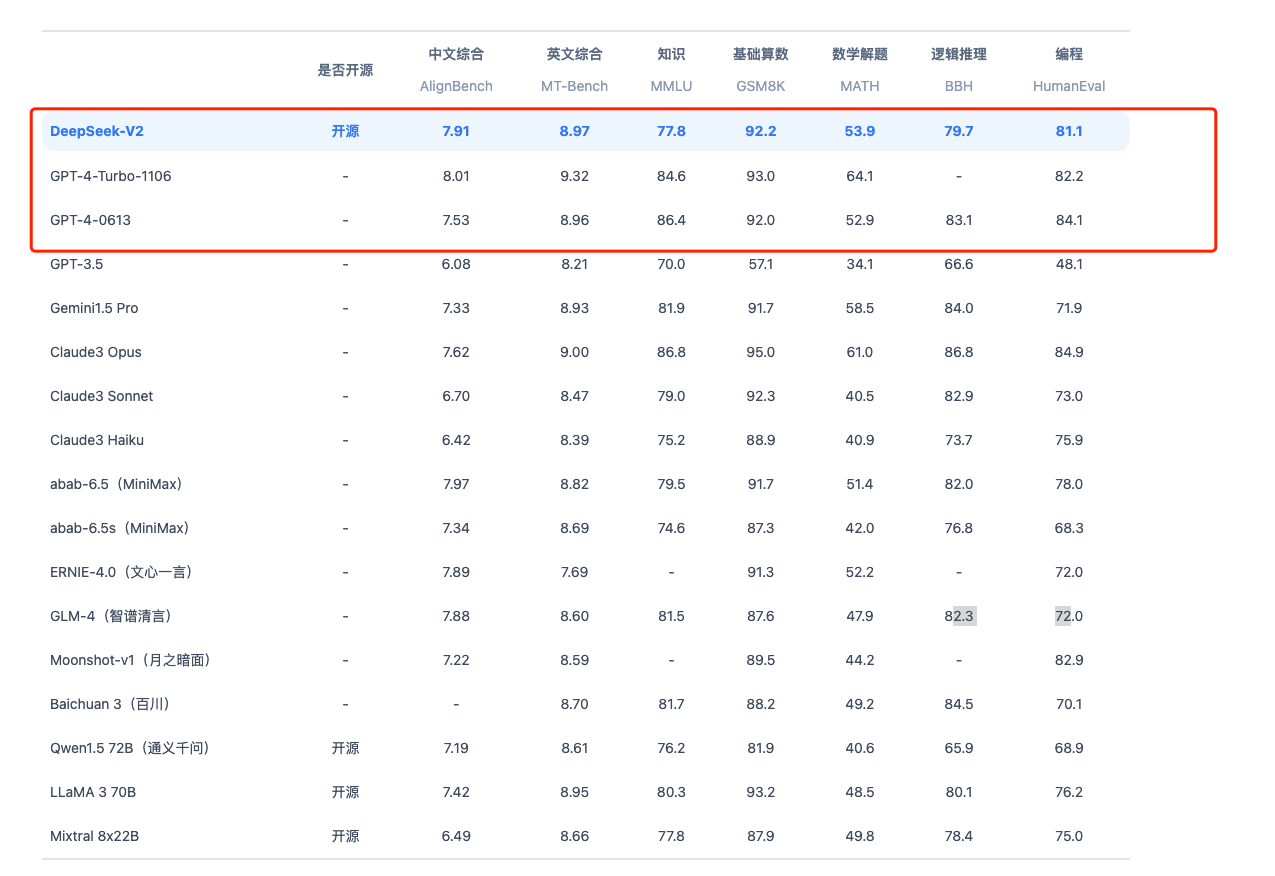
#### Q: 国内有无白嫖的类似gpt4的资源

A: 国内有不少模型注册都有. 这里总结了几个.**用的时候一个个领 ,token用不完会过期, 别一起领**.

##### DeepSeekk

<https://www.deepseek.com>

兼容openAI, 评分上靠近gpt4, 价格也便宜.



##### 月之暗面

<https://platform.moonshot.cn/console/account>

兼容openAI, 注册实名就有.

##### 智谱开放平台

* **获得的 Tokens 适用于智谱的通用大模型、超拟人大模型、向量大模型推理**
* **一共可以领取 1200W 智谱 AI 推理 tokens（价值 1200 元）**
* 从领取之日开始一个月有效，所以最好是用的时候在领取；
* 1200w 分为两次领取，其中可以先通过链接注册账号领取 800w，用差不多了之后在通过实名认证领取 400w，这样不至于同一时间到期；
* 已经注册的老用户无法获得注册的 800w tokens（换个手机号），已经实名过了也获得不了400w tokens（换个人认证）

Tokens 领取流程：

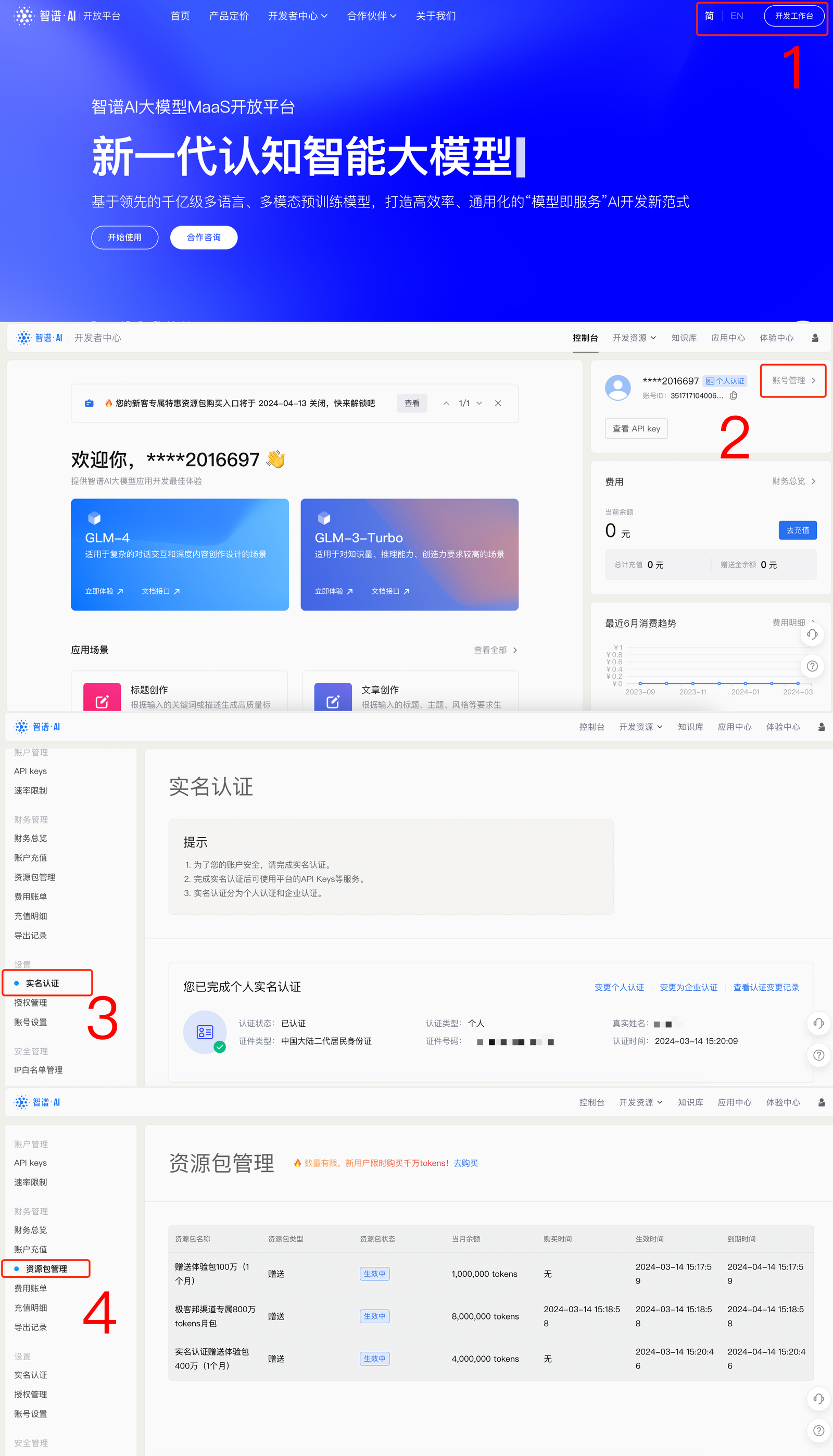
**领取链接：**[https://open.bigmodel.cn/geekbang/1004](https://open.bigmodel.cn/geekbang/1004?accessToken=eyJhbGciOiJIUzI1NiIsImtpZCI6ImRlZmF1bHQiLCJ0eXAiOiJKV1QifQ.eyJleHAiOjE3MzQ0MzM0NDUsImZpbGVHVUlEIjoidlZBWE04VjFtZ0hyWmIzbSIsImlhdCI6MTczNDQzMzE0NSwiaXNzIjoidXBsb2FkZXJfYWNjZXNzX3Jlc291cmNlIiwidXNlcklkIjoxNTYzNDc0M30.LH_iqImNNONv1uP4IJtimYS_lblYPptJsgsPXc5OCeE)

**Step1：**

通过链接注册后获取 800W tokens（一个月期限，领取完成之后可以先使用这部分）

**Step2：**

进入开放平台（控制台-账号管理），完成实名认证获取另外 400W tokens（一个月到期）



## 三、课程学习

### 作业提交

* 作业提交，建议使用github, 或者gitee提交哈
* 另外有个README文档，说明下自己的操作流程

### 基础入门

### 效率神器

### 优秀作业

### 第五章

1. [GitHub - ap198245/rag-load](https://github.com/ap198245/rag-load)
2. <https://github.com/qinglanliu/learn_rag>

## 四、常见问题

### 有没有好用的开源知识库

* <https://github.com/QuivrHQ/quivr>
* <https://doc.fastgpt.in/docs/development/intro/>

1、知识库对比：dify、fastgpt、langchatchat

2、agent构建器选择：flowise、langflow、bisheng

### 国外的huggingface源站下载太慢怎么处理

A: 试试国内镜像站: <https://hf-mirror.com/>, <https://www.modelscope.cn/models>

### 我在下载pip install包的时候出现了失败的情况怎么处理

A: 使用临时源, 诸如安装requests包: pip install requests -i<https://pypi.tuna.tsinghua.edu.cn/simple>

## 五、拓展资料

### GenAI Cookbook（大模型应用案例）

<https://github.com/FLY-Open-AI/GenAI-Showcase>

### 优秀Blog:

1. <https://mp.weixin.qq.com/mp/appmsgalbum?__biz=MzkzODUxMTY1Mg==&action=getalbum&album_id=3433186668809863174&scene=173&subscene=&sessionid=svr_3549e5aa825&enterid=1725419843&from_msgid=2247486537&from_itemidx=1&count=3&nolastread=1#wechat_redirect>
2. <https://medium.com/tag/llm>

### OpenAI API文档

<https://platform.openai.com/docs/overview>

<https://platform.openai.com/docs/api-reference/introduction>

<https://openai.xiniushu.com/docs/models>（中文）

### langgraph中文文档:

<https://github.langchain.ac.cn/langgraph/how-tos/persistence/>

### Android 安装大模型

<https://llm.mlc.ai/docs/get_started/quick_start.html>

### 试验 OpenAI 的各种模型

<https://gpt3demo.com/apps/openai-gpt-3-playground>

### 多模态 GPT-4 模型的 AI引用

<https://gpt4demo.com/all>

### 提示词使用

<https://www.promptingguide.ai/zh>

### 轻松搜索arXiv最相关论文

<https://arxivxplorer.com/>

### 大模型推理引擎

<https://github.com/DefTruth/Awesome-LLM-Inference>

### RAGFlow实战

* **RAGFlow 入门指南：解锁你的智能知识库引擎**<https://mp.weixin.qq.com/s/85sGBHXBEDk3evXd0f-UCQ>
* **RAGFlow 进阶：架构剖析与调试技巧** <https://mp.weixin.qq.com/s/dA8LZFNf8tRxNjiaFCJRYw>

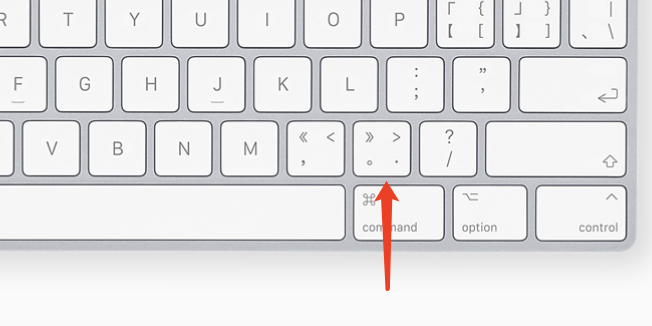
## 实用小技巧

### Github打开在线visual code

可以通过以下任一方式打开 github.dev 中的任何 GitHub 存储库：

* 要在同一浏览器标签页中打开存储库，请按 . （在 GitHub 上浏览任何存储库或拉取请求时）。

要在新的浏览器标签页中打开存储库，请按 >。



或者：**将 URL 从“github.com”更改为“github.dev”。**

例如：<https://github.dev/AlexFly666/002-openai-quickstart-jike-peng/blob/main/langchain/jupyter/model_io/model.ipynb>

改为

<https://github.dev/AlexFly666/002-openai-quickstart-jike-peng/blob/main/langchain/jupyter/model_io/model.ipynb>

或者：**查看文件时，选择“”下拉菜单，然后单击“github.dev”。**

