AI 大模型开发工程师
之GPT大模型开发基础

讲师：李希沅

目录

1 OpenAI 账户注册

2 OpenAI 官网介绍

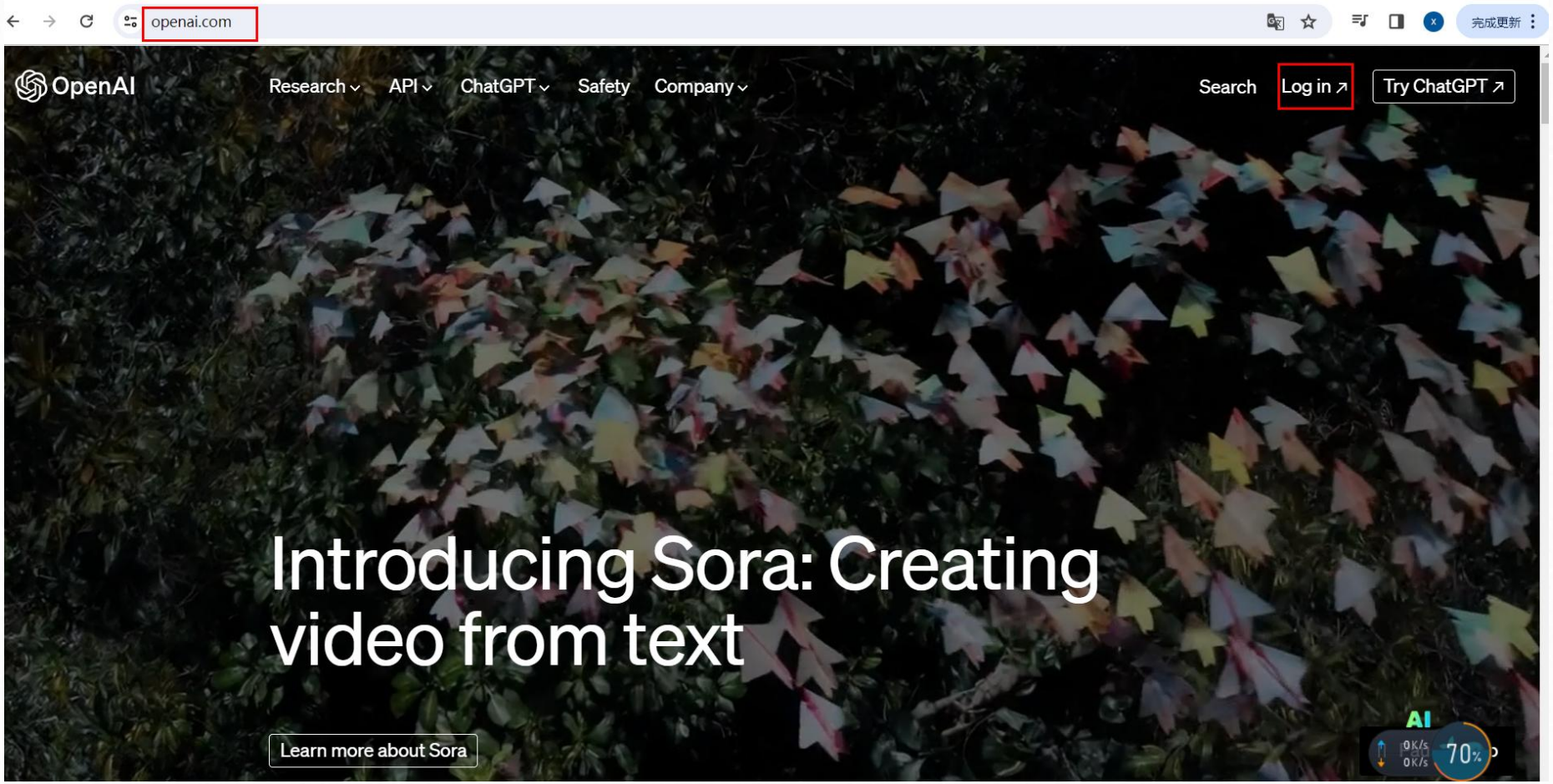
3 OpenAI GPT费用计算

4 OpenAI key获取与配置

5 OpenAI 大模型总览

1 OpenAI 账户注册

01、OpenAI账户注册

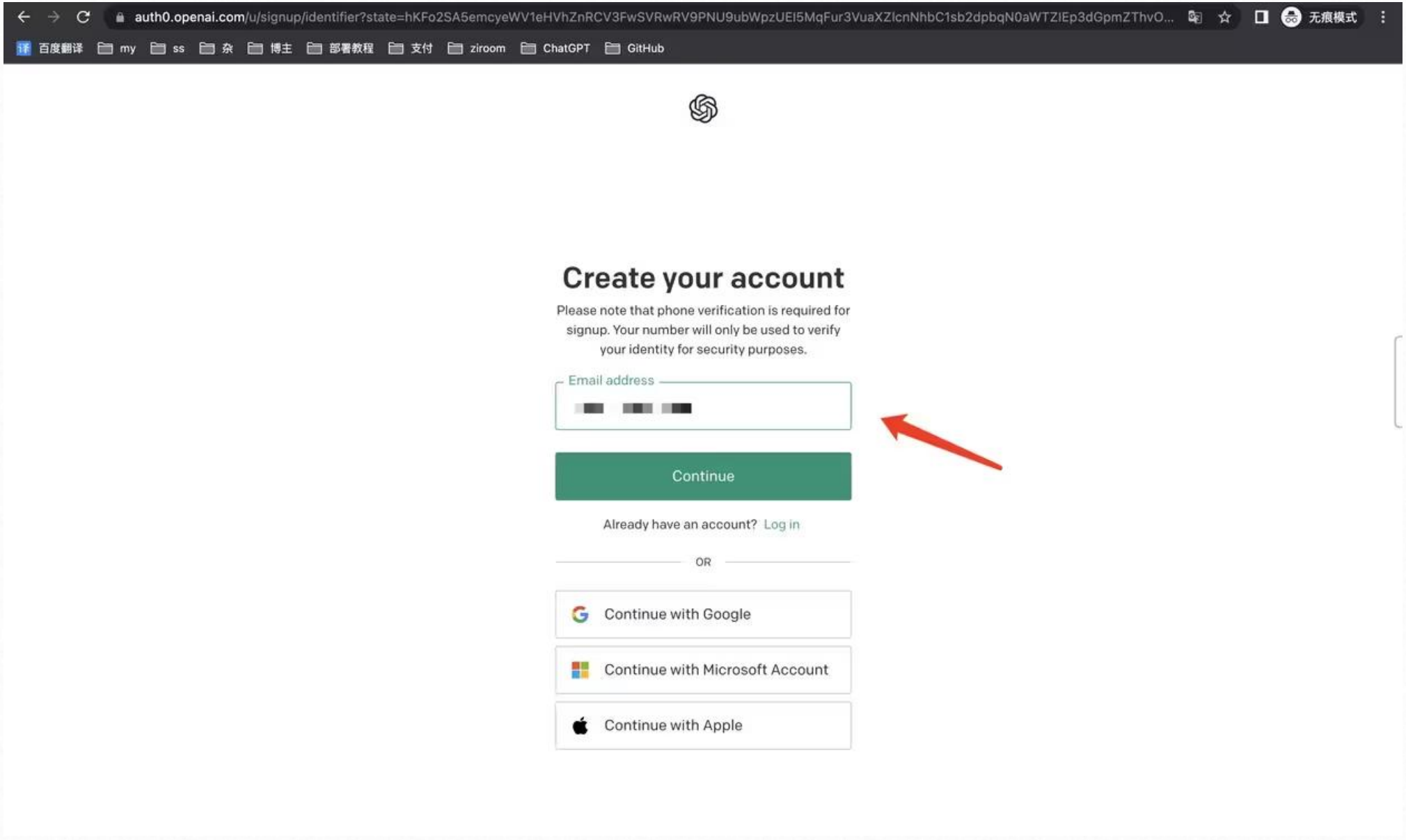


科学上网

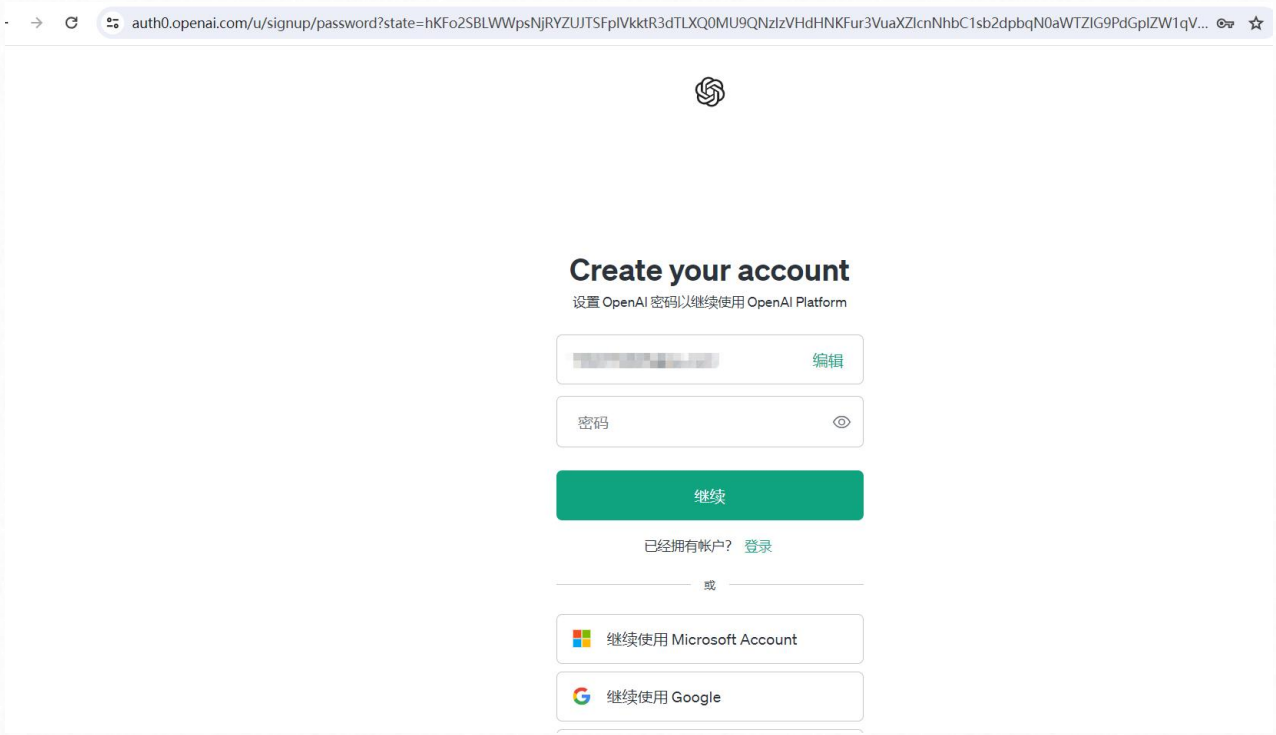
国外邮箱

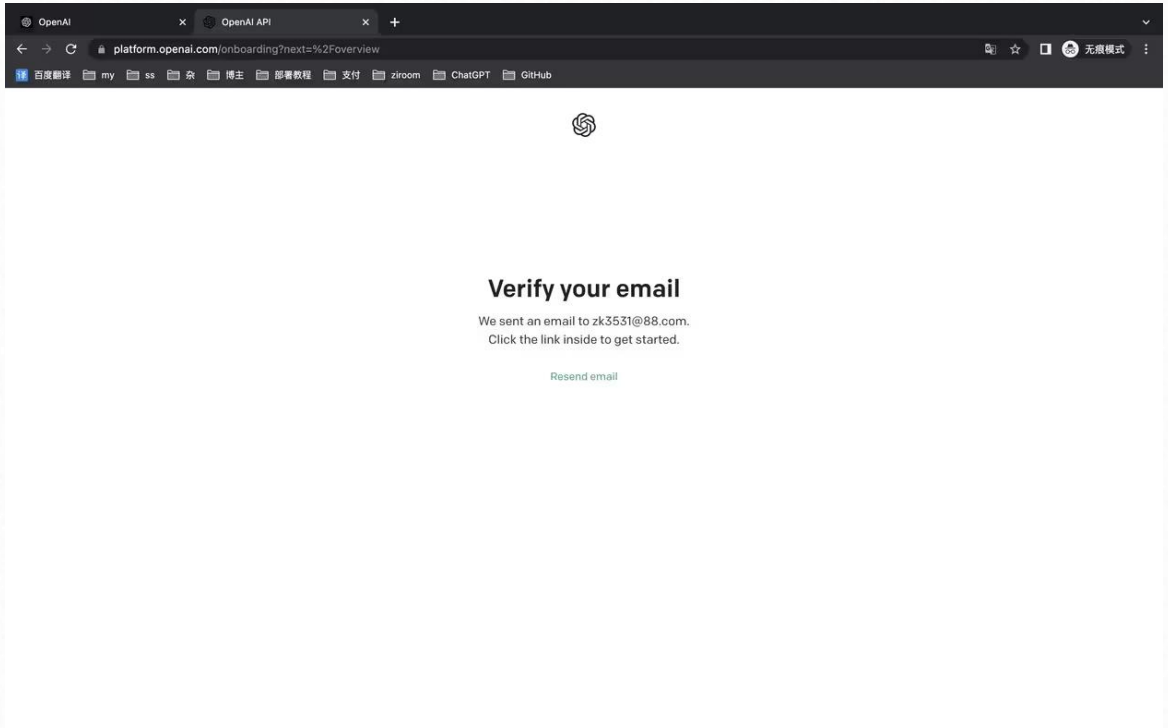
可以参考：https://www.bilibili.com/read/cv23758827/

02、OpenAI账户注册

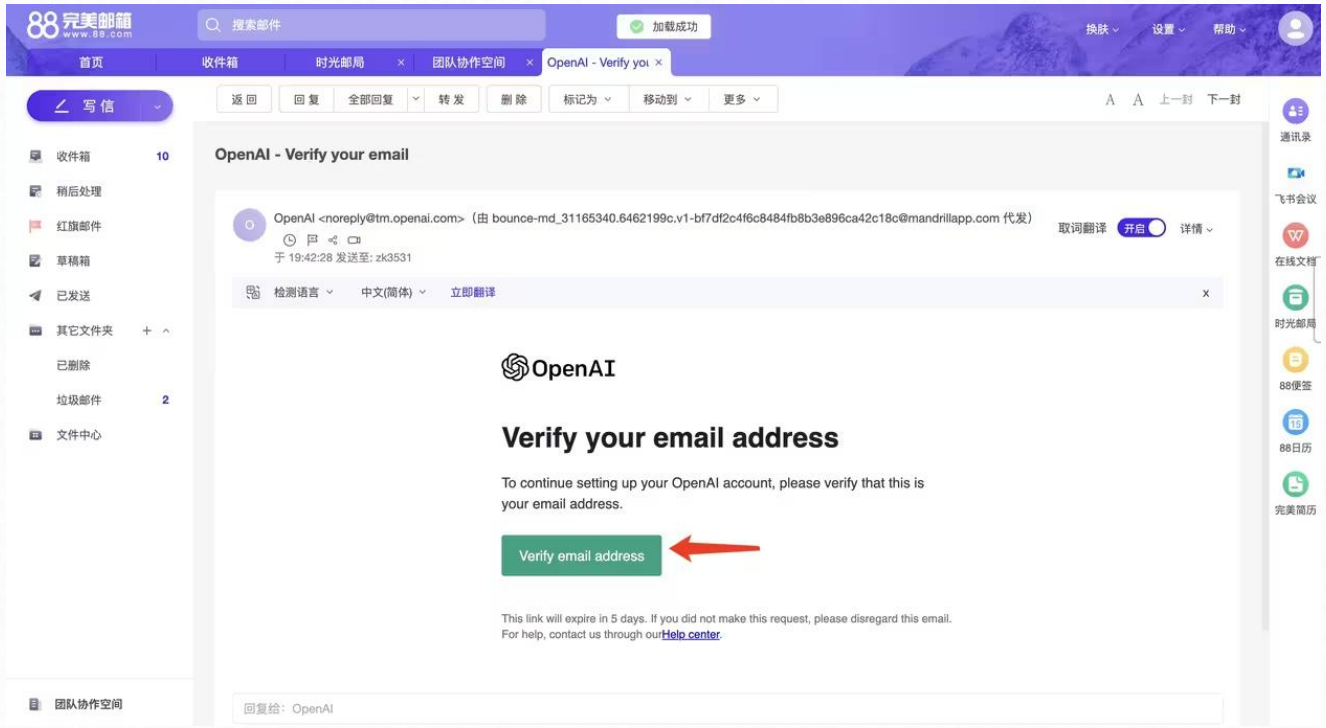


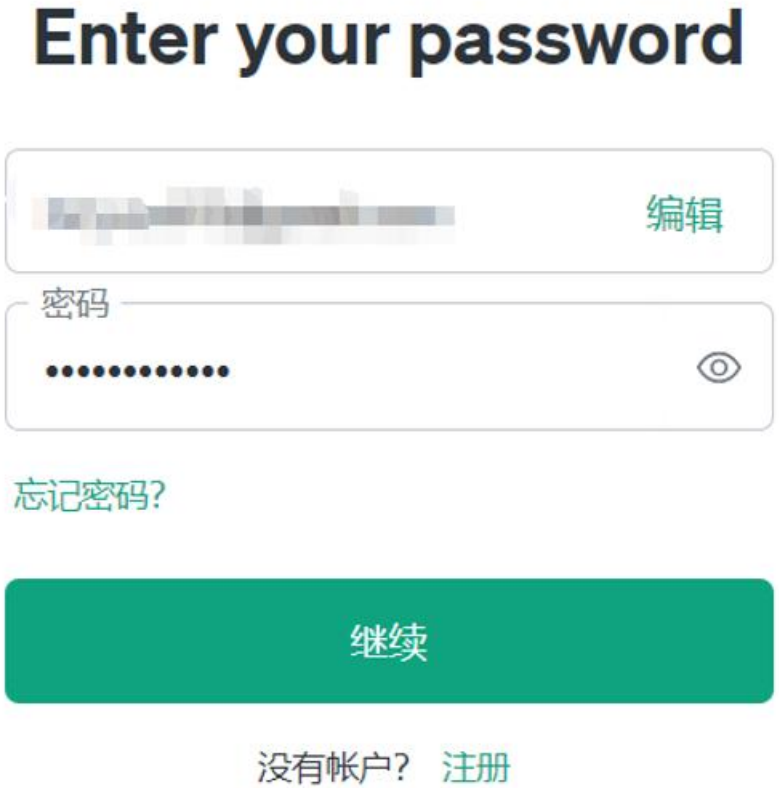
03、OpenAI账户注册





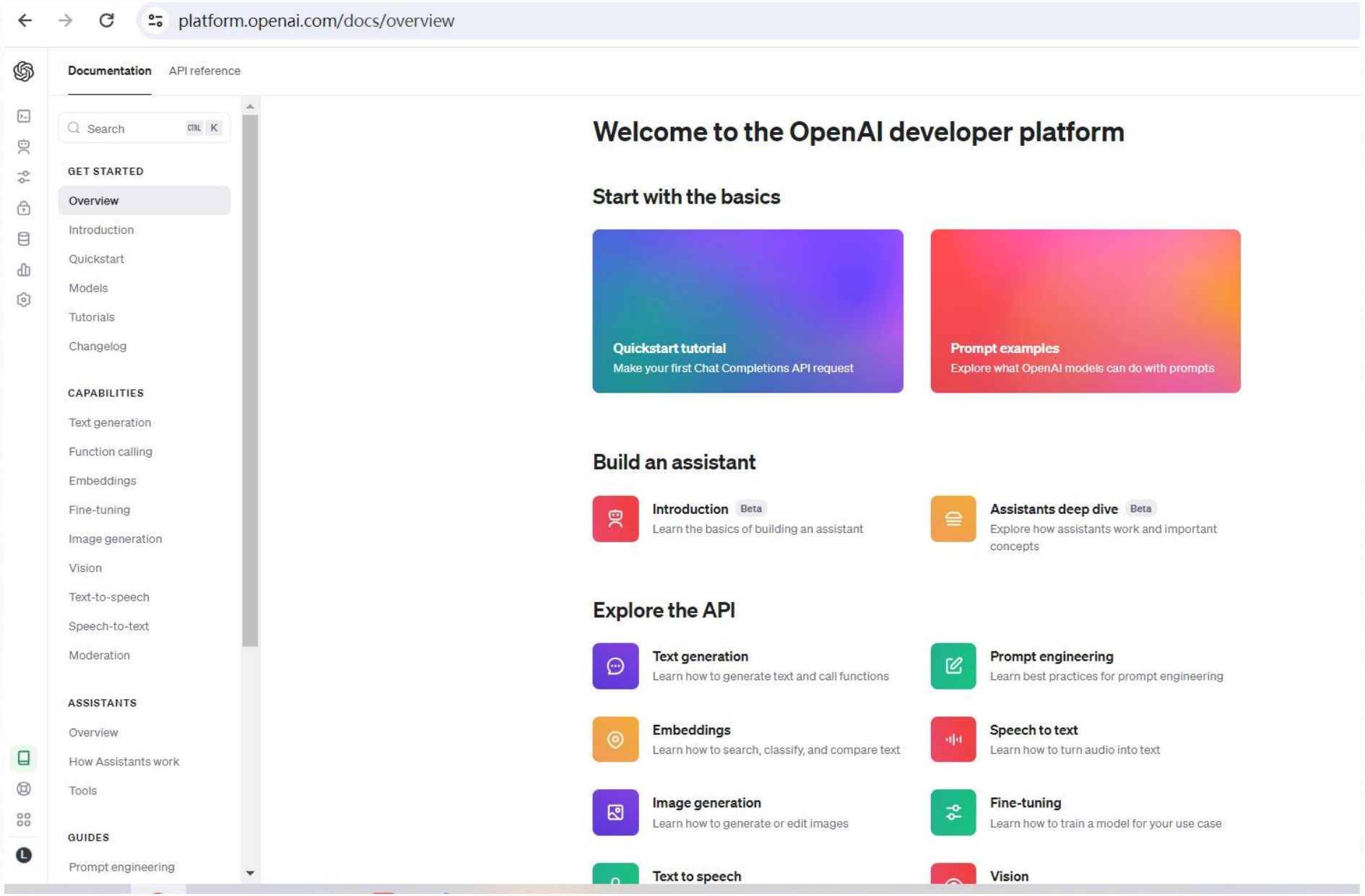
04、OpenAI账户注册





2 OpenAI 官网介绍

01、官网介绍


platform.openai.com/docs/overview
Documentation
API reference
Search
CTRL K
GET STARTED
Overview
Introduction
Quickstart
Models
Tutorials
Changelog
CAPABILITIES
Text generation
Function calling
Embeddings
Fine-tuning
Image generation
Vision
Text-to-speech
Speech-to-text
Moderation
ASSISTANTS
Overview
How Assistants work
Tools
GUIDES
Prompt engineering
Welcome to the OpenAI developer platform
Start with the basics
Quickstart tutorial
Make your first Chat Completions API request
Prompt examples
Explore what OpenAI models can do with prompts
Build an assistant
Introduction Beta
Learn the basics of building an assistant
Assistants deep dive Beta
Explore how assistants work and important concepts
Explore the API
Text generation
Learn how to generate text and call functions
Prompt engineering
Learn best practices for prompt engineering
Embeddings
Learn how to search, classify, and compare text
Speech to text
Learn how to turn audio into text
Image generation
Learn how to generate or edit images
Fine-tuning
Learn how to train a model for your use case
Text to speech
Vision

3 OpenAI 费用计算

01、OpenAI费用计算

API费用计算：https://openai.com/pricing

ChatGPT费用计算：https://openai.com/chatgpt/pricing

Model	Input	Output
gpt-4-0125-preview	$10.00 / 1M tokens	$30.00 / 1M tokens
gpt-4-1106-preview	$10.00 / 1M tokens	$30.00 / 1M tokens
gpt-4-1106-vision-preview	$10.00 / 1M tokens	$30.00 / 1M tokens

Vision pricing calculator

150 px by 150 px = $0.00085

■ Low resolution

《红楼梦》有731017个字

02、免费的Token的限制

How do these rate limits work?

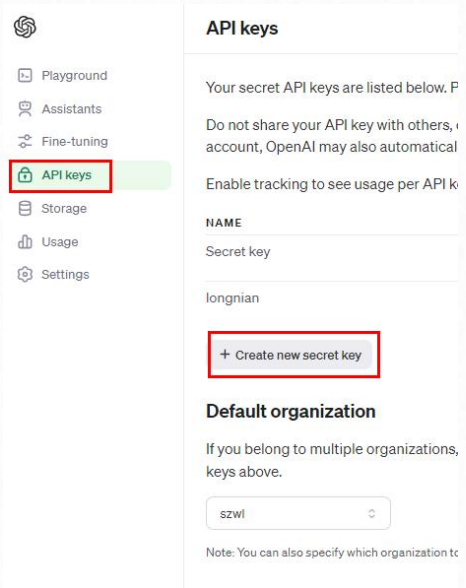
Rate limits are measured in five ways: **RPM** (requests per minute), **RPD** (requests per day), **TPM** (tokens per minute), **TPD** (tokens per day), and **IPM** (images per minute). Rate limits can be hit across any of the options depending on what occurs first. For example, you might send 20 requests with only 100 tokens to the ChatCompletions endpoint and that would fill your limit (if your RPM was 20), even if you did not send 150k tokens (if your TPM limit was 150k) within those 20 requests.

This is a high level summary and there are per-model exceptions to these limits (e.g. some legacy models or models with larger context windows have different rate limits). To view the exact rate limits per model for your account, visit the limits section of your account settings.

MODEL	RPM	RPD	TPM
gpt-3.5-turbo	3	200	40,000
text-embedding-3-small	3	200	150,000
whisper-1	3	200	-
tts-1	3	200	-
dall-e-2	5 img/min	-	-
dall-e-3	1 img/min	-	-

4 OpenAI Key的获取与配置

01、OpenAI Key的获取与配置



方案一：自己注册

科学上网　　搞不定，找助理

国外手机　　使用里面的平台：https://www.bilibili.com/read/cv23758827

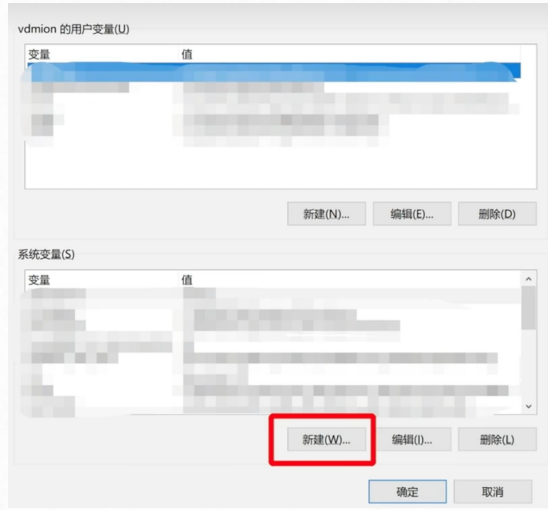
国外信用卡　　只能自己想办法了

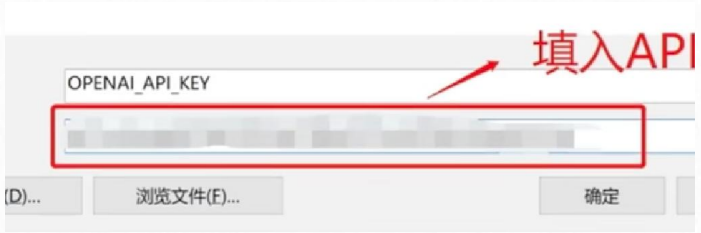
方案二：求助淘宝

方案三：找代充平台

02、API keys配置

Windows电脑





Mac或者Linux

1.首先打开终端

输入：vi ~/.bashrc ，进入配置文件中，

2.添加以下代码：

```
export OPENAI_API_KEY='你的 OpenAI API 密钥'
```

5 OpenAI 大模型总览

01、GPT大模型概览

Models

Overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4 and GPT-4 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

02、历史上的大模型

GPT-3 Legacy

GPT-3 models can understand and generate natural language. These models were superseded by the more powerful GPT-3.5 generation models. However, the original GPT-3 base models (`davinci` , `curie` , `ada` , and `babbage`) are current the only models that are available to fine-tune.

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`text-curie-001`	Very capable, faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
`text-babbage-001`	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019
`text-ada-001`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019
`davinci`	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality.	2,049 tokens	Up to Oct 2019
`curie`	Very capable, but faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
`babbage`	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019
`ada`	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019

停止维护了

谢谢观看