

# AI 大模型开发工程师 垂直领域的智能在线搜索平台

讲师：李希沅


# 目录

- 1 智能在线搜索平台需求分析
- 2 智能在线搜索平台方案设计
- 3 智能在线搜索平台代码落地
- 4 智能在线搜索平台项目总结

# **1 智能在线搜索平台需求分析**





# 01、大模型不够“聪明”

大模型	数据截止时间
GPT-3.5	2021年9月
GPT-4	2021年9月

 ChatGPT

截至我最后一次更新知识库的时候，即2022年，我并没有关于清华智谱公司出品的GLM4大模型的具体信息。不过，如果这是在我之后的事情，可能GLM4大模型是指该公司开发的一系列基于GLM（Generalized Linear Models）理论的模型。这些模型可能在处理各种不同类型的数据时具有广泛的应用，涵盖正态分布、二项分布、泊松分布和多项分布等情况。

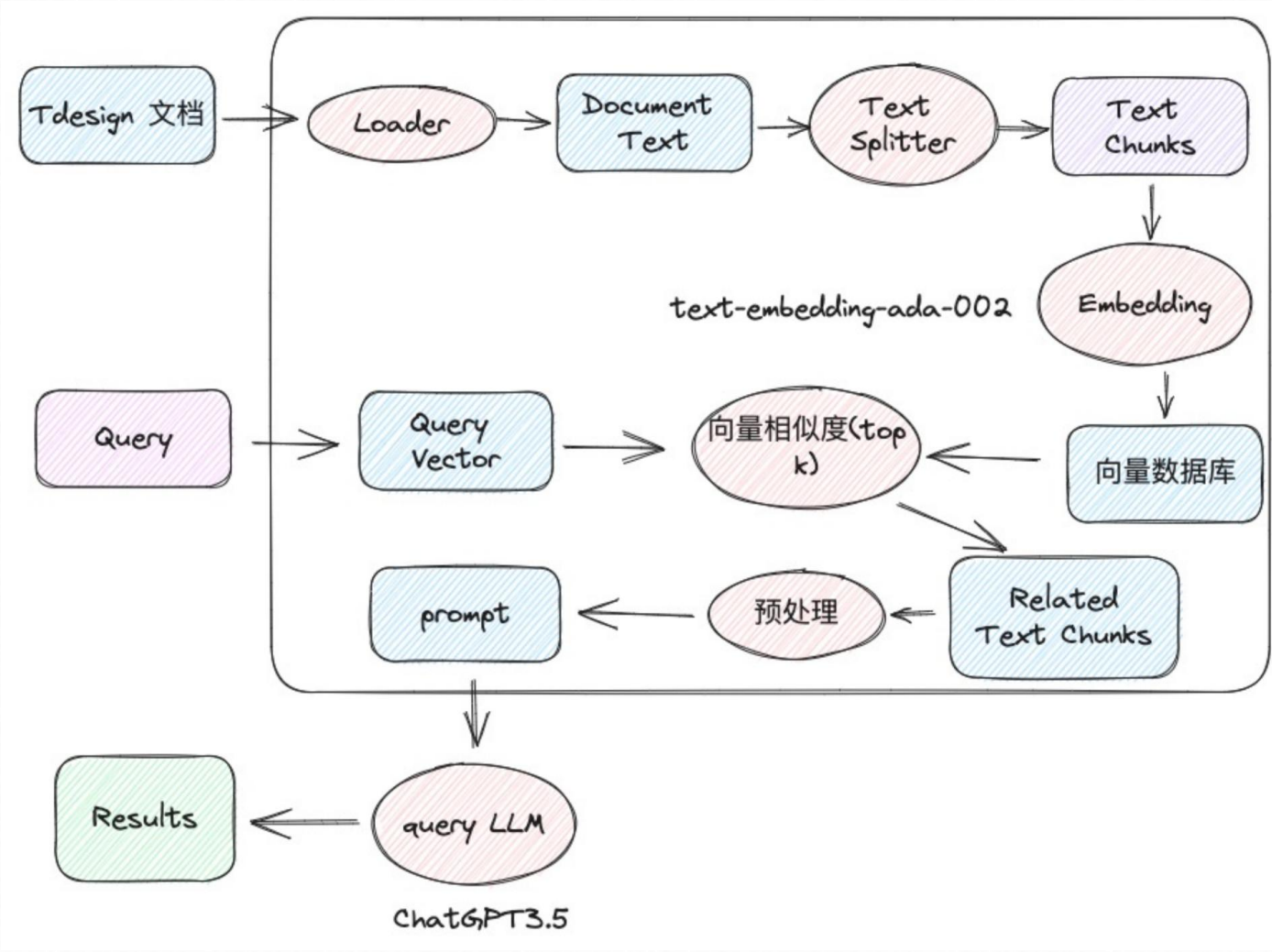
如果你有关于GLM4大模型的具体信息或者需要了解更多内容，我建议你直接查阅清华智谱公司发布的相关资料或者联系他们的技术支持团队以获取详细信息。



RAG

微调

RAG: Retrieval-Augmented Generation (知识检索增强生成)

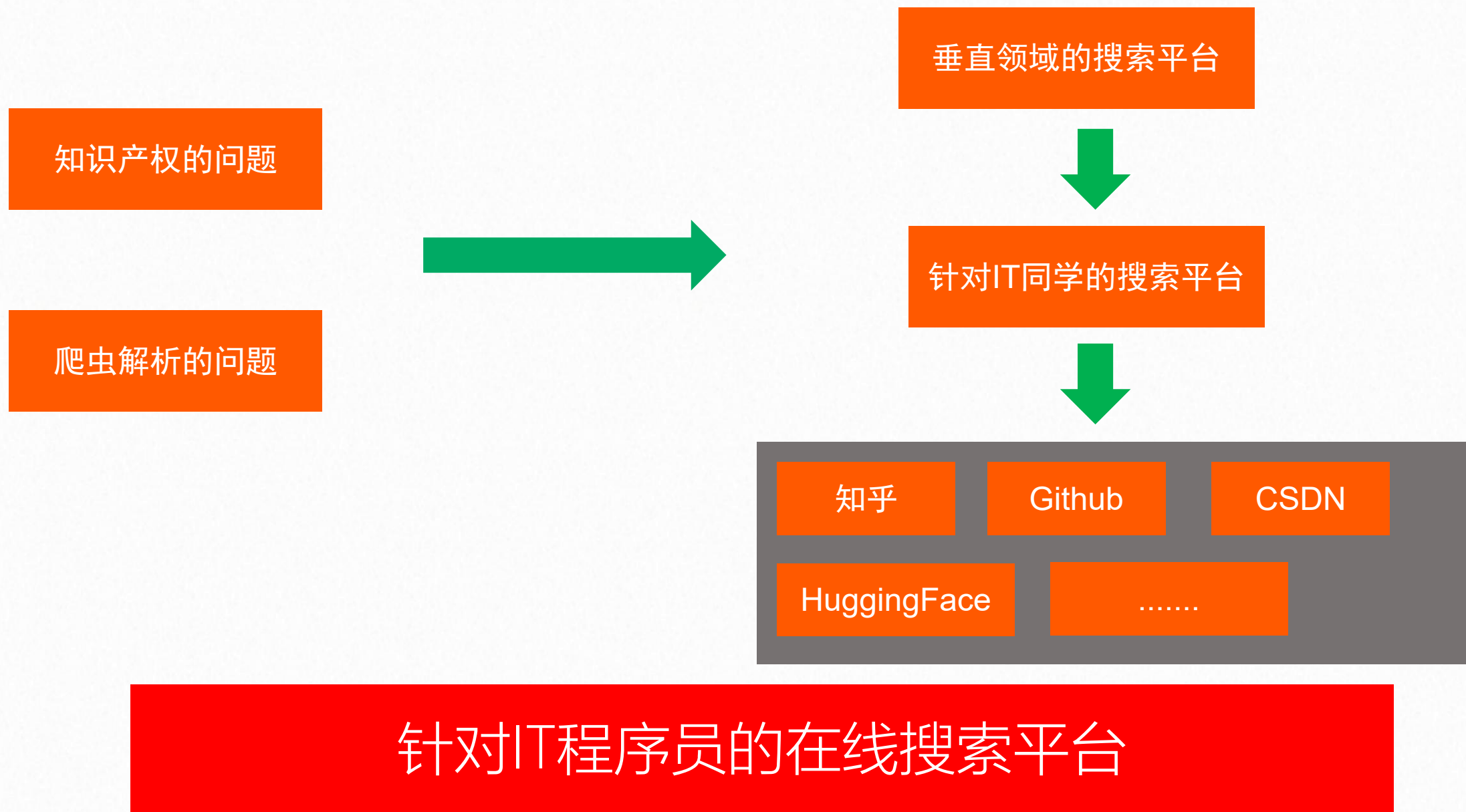




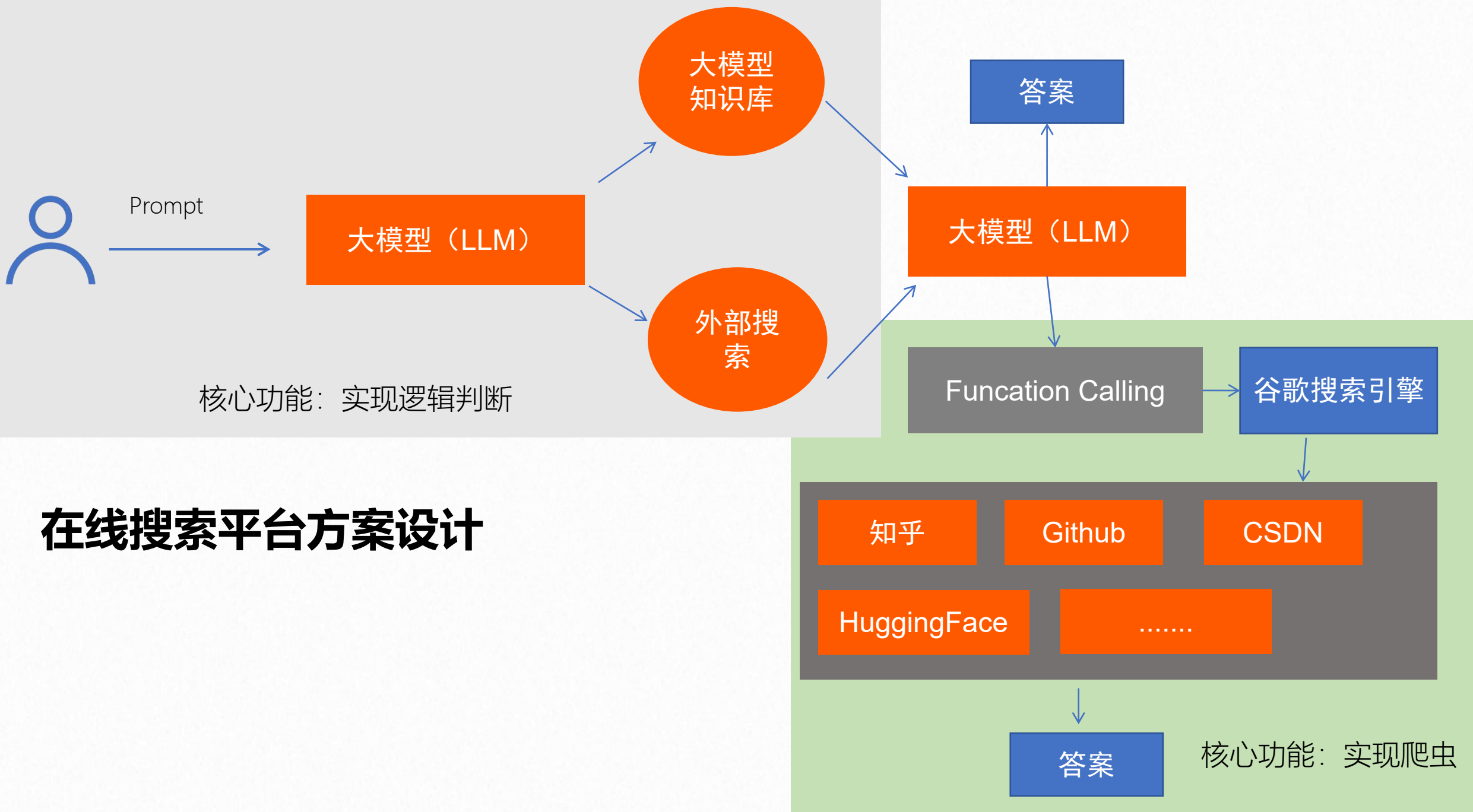
## 02、在线搜索平台需求分析



### 03、通用的搜索引擎？



## **② 智能在线搜索平台方案设计**



# 在线搜索平台方案设计



# 01、技术选型

功能模块	技术选型
大模型	GPT大模型/GLM4大模型/其他模型
搜索技术	谷歌云API、爬虫
性能优化	提示词工程

```
messages=[
    {"role": "system", "content": "你是一个乐于助人的小助手"},
    {"role": "user", "content": "你好，什么是深度学习？"},
]

chat_response = client.chat.completions.create(
    model="models/Qwen1.5-7B-Chat",
    messages=messages
)

print(chat_response.choices[0].message.content)
```

```
from zhipuai import ZhipuAI
client = ZhipuAI(api_key="")
messages = []
tools = [
    {
        "type": "function",
        "function": {
            "name": "get_flight_number",
            "description": "根据始发地、目的地和日期，查询对应日期的航班号。",
            "parameters": {
                "type": "object",
                "properties": {
                    "departure": {
                        "description": "出发地",
                        "type": "string"
                    },
                    "destination": {
                        "description": "目的地",
                        "type": "string"
                    },
                    "date": {
                        "description": "日期",
                        "type": "string"
                    }
                }
            }
        }
    }
]

我们想查询2024年1月20日从北京前往上海的航班。我们向模型提供这个信息：
```

```
1 messages = []
2 messages.append({"role": "user", "content": "帮我查询从2024年1月20日，从北京出发前往上海的航班"})
3 response = client.chat.completions.create(
4     model="glm-4", # 填写需要调用的模型名称
5     messages=messages,
6     tools=tools,
7 )
8 print(response.choices[0].message)
9 messages.append(response.choices[0].message.model_dump())
```

# 01、技术选型-版本

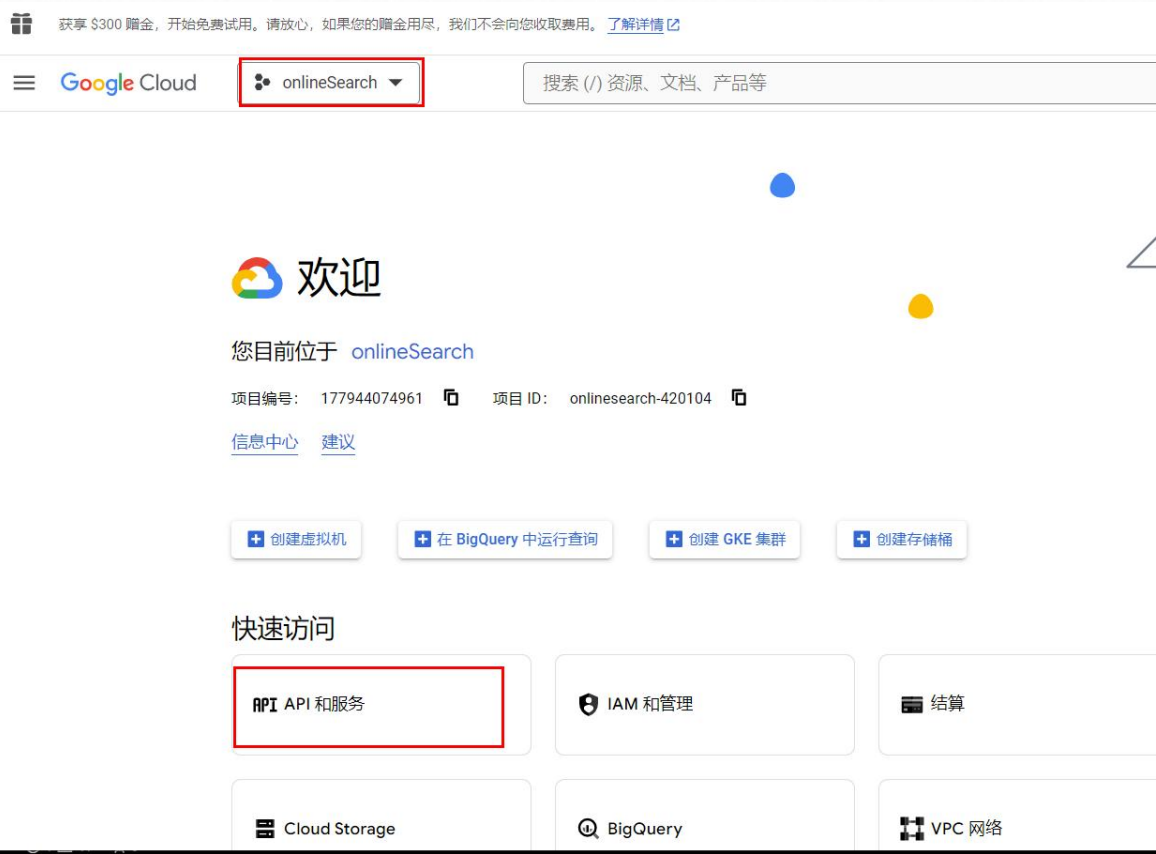
```
root@autodl-container-cb2246bbbe-1e711273:~# pip show zhipuai
Name: zhipuai
Version: 2.0.1
Summary: A SDK library for accessing big model apis from ZhipuAI
Home-page: https://open.bigmodel.cn/
Author: Zhipu AI
Author-email: None
License: None
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: httpx, cachetools, pyjwt, pydantic
Required-by:
root@autodl-container-cb2246bbbe-1e711273:~# pip show openai
Name: openai
Version: 1.8.0
Summary: The official Python library for the openai API
Home-page: None
Author: None
Author-email: OpenAI <support@openai.com>
License: None
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: anyio, typing-extensions, tqdm, sniffio, distro, httpx, pydantic
Required-by:
root@autodl-container-cb2246bbbe-1e711273:~#
```

# 02、Google Cloud平台注册

地址: <https://console.cloud.google.com/>

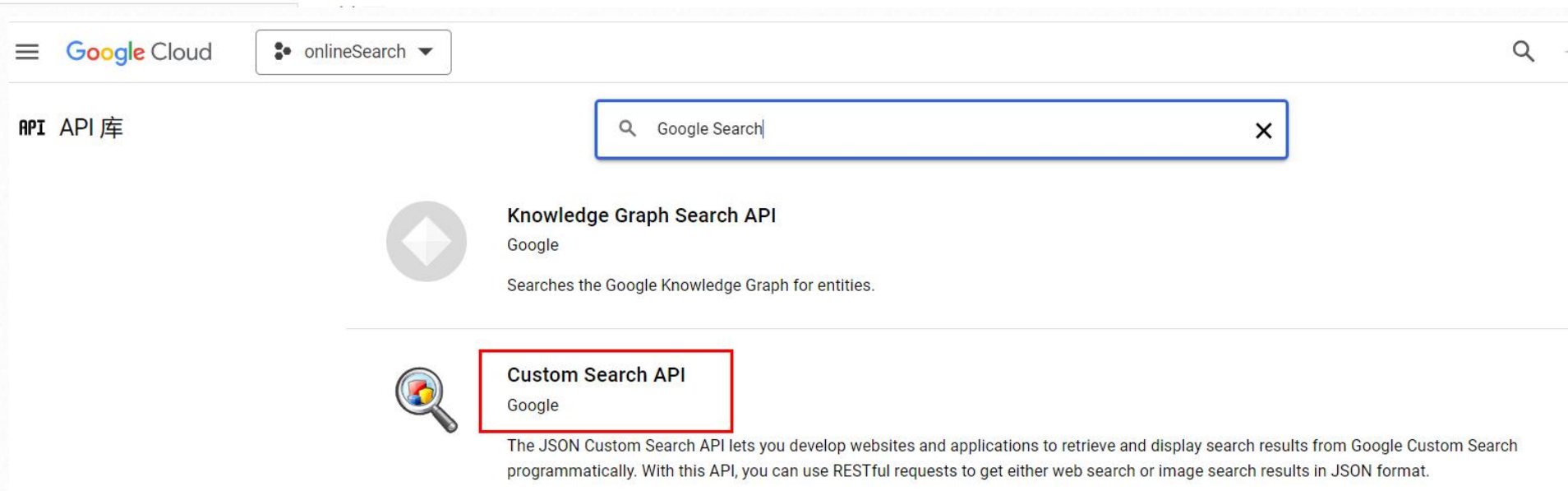


# 03、Google Cloud平台注册







## 04、Google Cloud平台注册






# 05、Google Cloud平台注册

 获享 \$300 赠金，开始免费试用。请放心，如果您的赠金用尽，我们不会向您收取费用。[了解详情](#)



onlineSearch ▾

[← 产品详情](#)



### Custom Search API

[Google](#)

Retrieve and display search results from Google Custom Search programmatically.

启用

试用此 API

概览

价格


文档

支持

相关产品

## 概览

The JSON Custom Search API lets you develop websites and applications to retrieve and display search results from Google Custom Search programmatically. With this API, you can use RESTful requests to get either web search or image search results in JSON format.



onlineSearch ▾


搜索 (/) 资源、文档、产品等

API

[← API/服务详情](#)

■ 停用 API

返回“已启用的 API 和服务”



### Custom Search API

Searches over a website or collection of websites

提供方: Google

服务名称	类型	状态
customsearch.googleapis.com	公共 API	已启用

指标

配额和系统限制

凭据

费用

选择图表

4 个图表

过滤条件:

Credentials

未指定、匿名和Google Cl...

Methods

List、SiteRestrictList、Ge...

流量 (按响应代码)

# 06、Google Cloud平台注册

API

← API/服务详情 ■ 停用 API

您会在配额即将达到上限时收到提醒。点击行中的 ⋮ 即可开始，您也可以点击“了解详情”查看文档。

了解详情

当前用量 > 90%

7 天峰值用量 > 90%

所有配额和系统限制

0

0

6

查看配额和系统限制

查看配额和系统限制

过滤条件 输入属性名称或值

<input type="checkbox"/>	名称	类型	维度 (例如位置)	值	当前用量百分比	↓
<input type="checkbox"/>	Queries per day	配额		100		
<input type="checkbox"/>	Queries per minute	配额		无限制	-	
<input type="checkbox"/>	Site Restricted Queries per day	配额		100		
<input type="checkbox"/>	Site Restricted Queries per minute	配额		无限制	-	
<input type="checkbox"/>	Queries per minute per user	配额		100	- ?	
<input type="checkbox"/>	Site Restricted Queries per minute per user	配额		100	- ?	

https://console.cloud.google.com/apis/api/customsearch.googleapis.com/quotas?project=onlinesearch-420104

← → ↺

console.cloud.google.com/apis/api/customsearch.googleapis.com/quotas?project=online

📁 获享 \$300 赠金，开始免费试用。请放心，如果您的赠金用尽，我们不会向您收取费用。了解详情

☰ Google Cloud

onlineSearch ▾

搜索 (/) 资源、文档、产品等

API API 和服务 ■ 停用 API

已启用的 API 和服务

库

凭据

OAuth 权限请求页面

页面使用协议

您会在配额即将达到上限时收到提醒。点击行中的 ⋮ 即可开始，您也可以点击“了解详情”查看文档。

了解详情

7 天峰值用量 > 90%

所有配额和系统限制

0

6

查看配额和系统限制

值

	类型	维度 (例如位置)	值	当前月
	配额		100	
	配额		无限制	-
s per day	配额		100	
s per minute	配额		无限制	-

# 07、Google Cloud平台注册

获享 \$300 赠金，开始免费试用。请放心，如果您的赠金用尽，我们不会向您收取费用。[了解详情](#)

Google Cloud

onlineSearch

搜索 (/) 资源、文档、产品等

API

凭据

+ 创建凭据

删除

恢复已删除的凭据

创建凭据，以访问您已启用的 API。[了解详情](#)

请务必使用关于您的应用的信息配置 OAuth 同意屏幕。

API 密钥

名称

创建日期

↓

没有要显示的 API 密钥

OAuth 2.0 客户端 ID

名称

创建日期

↓

没有要显示的 OAuth 客户端

服务账号

电子邮件

没有要显示的服务账号

获享 \$300 赠金，开始免费试用。请放心，如果您的赠金用尽，我们不会向您收取费用。[了解详情](#)

Google Cloud

onlineSearch

搜索 (/) 资源、文档、产品等

API

凭据

+ 创建凭据

删除

恢复已删除的凭据

创建凭据，以

请务必使用关于您的应用的信息配置 OAuth 同意屏幕。

API 密钥

名称

创建日期

↓

没有要显示的 API 密钥

OAuth 2.0 客户端 ID

名称

创建日期

↓

没有要显示的 OAuth 客户端

API 密钥

使用简单的 API 密钥来标识您的项目，以便检查配额和访问权限

OAuth 客户端 ID

征得用户同意，让您的应用有权访问用户数据

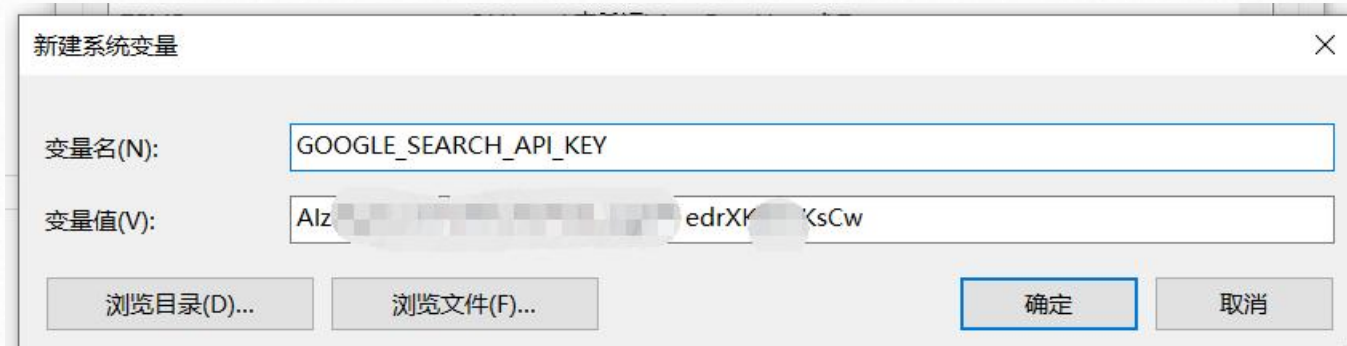
服务账号

利用机器人账号实现服务器到服务器的应用级身份验证

帮我选择

询问几个问题以帮助您决定要使用的凭据类型

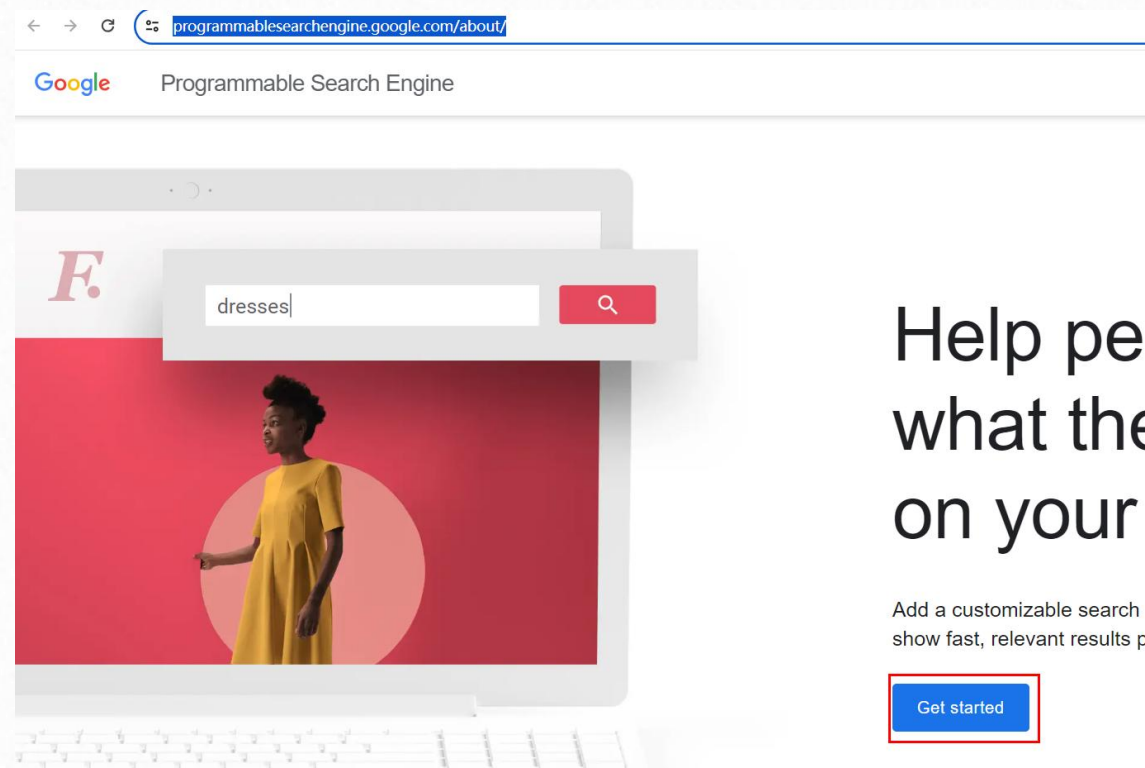
## 08、Google Cloud平台注册





# 09、创建可编程的搜索引擎

地址: <https://programmablesearchengine.google.com/>





# 10、创建可编程的搜索引擎

搜索设置 ⓘ

☐ 图片搜索

☐ 安全搜索



进行人机身份验证



点击“创建”即表示您同意 [服务条款](#).

创建



新的搜索引擎已创建

复制以下代码，并将其粘贴到网站的 `<body>` 部分，也就是您希望在其中显示搜索框和搜索结果的位置。

```
<script async src="https://cse.google.com/cse.js?cx=cd478d">
</script>
<div class="gcse-search"></div>
```



预览

自定义

新建系统变量

变量名(N):

CSE\_ID

变量值(V):

8a8f

浏览目录(D)...

浏览文件(F)...

确定

取消

# 11、创建可编程的搜索引擎



## 新的搜索引擎已创建

复制以下代码，并将其粘贴到网站的 `<body>` 部分，也就是您希望在其中显示搜索框和搜索结果的位置。

```
<script async src="https://cse.google.com/cse.js?cx=[ID]">
</script>
<div class="gcse-search"></div>
```

[预览](#)[自定义](#)

cse.google.com/cse?cx=[ID]

GLM4



glm4

glm 4 智谱

glm4官网

glm 4 开源

glm 4 github

glm 4发布

glm 4.0

glm4 zhipu

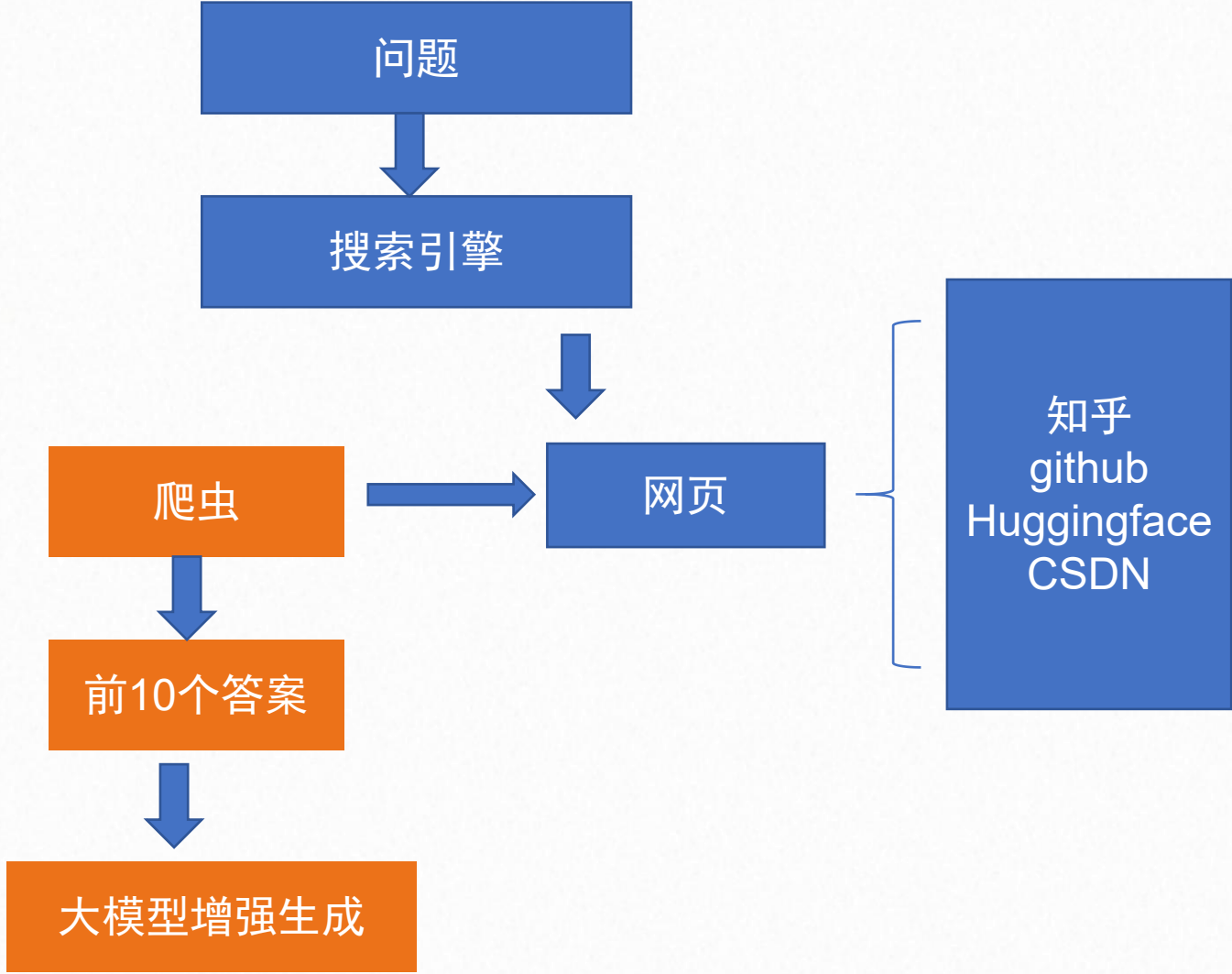
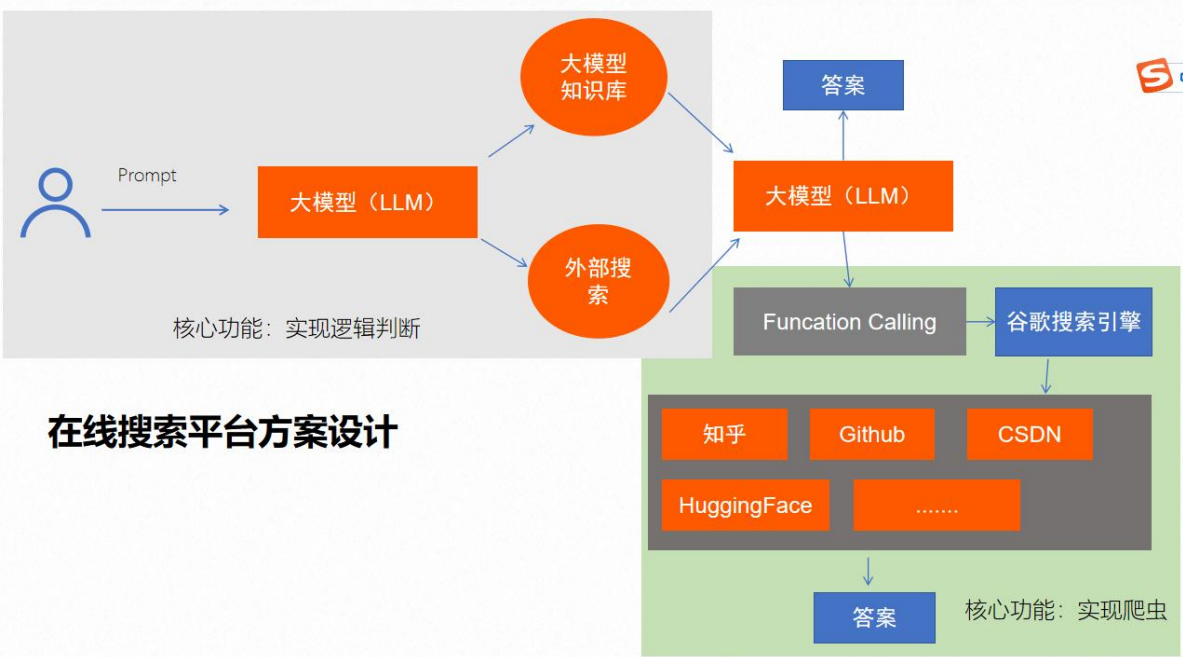
glm 4 langchain

glm4发布会

由 Google 带来优化体验

### **3 在线搜索平台代码落地**

# 01、完成在线搜索思路分析





## 02、Google搜索引擎功能测试

```
import requests

# Step 1.构建请求
url = "https://www.googleapis.com/customsearch/v1"

# Step 2.设置查询参数（还有很多参数）
params = {
    'q': search_term,          # 搜索关键词
    'key': google_search_key,  # 谷歌搜索API Key
    'cx': cse_id               # CSE ID
}

# Step 3.发送GET请求
response = requests.get(url, params=params)

# Step 4.解析响应
data = response.json()
```

Name	Description
c2coff	是否开启中文搜索模式，默认0表示开启，1表示停止
cr	在指定的国家地区进行搜索，即搜索返回结果只源于该地区
dateRestrict	只搜索某特定时间段内的内容
lr	搜索特定语言内容的内容
num	搜索返回的结果个数，默认为10个，有效数值为1-10
siteSearch	搜索结果必须包含某网址
siteSearchFilter	搜索结果必须排除某网址



# 03、完成知乎网站搜索之question页面

问题的主页

zhihu.com/question/639753877

知乎 首页 知乎知学堂 发现 等你来答 考公备考全流程

人工智能 大模型 LLM (大型语言模型) AI先行者 智谱

智谱发布 GLM-4 基座大模型，性能效果如何？

量子位：国产GPTs来了，基于智谱第4代大模型！模型性能均达GPT-4九成以上显示全部

关注问题 写回答 邀请回答 好问题 12 7 条评论 分享

92 个回答 默认排序

平凡

英语等 2 个话题下的优秀答主

+ 关注

95 人赞同了该回答

我简单试了一下，确实跟GPT4几乎一模一样的技术栈。

我上传了这么一张图片

1. Introduction

2. What are the capabilities of GPT-4?

2.1 Does GPT Use Deep Learning?

2.2 What Is the Downside of GPT?

3. How Many Parameters Does GPT-4 Have?

赞同 95

21 条评论

分享

收藏

喜欢

收起

# 04、完成知乎网站搜索之question/answer页面

某个问题的特定回答

zhihu.com/question/633793553/answer/3364933262

知乎 首页 知乎知学堂 发现 等你来答 十大良心赚钱软件

互联网 人工智能 大数据 AI技术 大模型

哪些 AI 大模型比较有前途?

关注问题 写回答 邀请回答 好问题 添加评论 分享 ...

查看全部 7 个回答



雨飞

大连理工大学 计算机科学与技术硕士

+ 关注

4 人赞同了该回答

目前我比较看好的大模型有 [kimi chat](#)，今天智普发布 [glm4](#)，看起来也是未来可期。

[Kimi chat](#) 我使用已经很长时间了，主要是使用它的[PDF读取](#)和网页读取的功能，这个大模型支持的上下文是可以达20万字。从我的实际体验下来，应该是比较好用的。

### KimiChat优势

从下图可以看到，能够有效的读取网上链接里PDF的内容，并进行整理。我使用的提示词比较简单，但kimi chat还是将文中重点的内容给总结出来了。



赞同 4

1 条评论

分享

收藏

喜欢

收起

# 05、完成知乎网站搜索之topic页面

话题页面

知乎

首页 知乎知学堂 发现 等你来答

十大良心赚钱软件

#

GLM

暂无话题描述

关注话题

管理

分享

讨论 精华 视频 等待回答

切换为时间排序

GLM-4函数调用：让AI帮你查成绩会发生什么，浙理工版

Helio

什么都会，什么都不懂。

在科技日新月异的今天，人工智能已经逐渐成为我们生活的一部分。其中，GLM-4 是一款由智谱AI训练的大型语言模型，它采用了清华大学KEG实验室提出的GLM模型结构。而基于GLM-4开发的智能体功能，更... [阅读全文](#)

赞同 1

添加评论

分享

收藏

举报

国产大模型GLM-4 能力接近ChatGPT4和Claude 2.1

李孟聊AI

大数据 | 人工智能 | AI生成 | SolidUI发起人

本文以GLM-4 发布功能作为基准对比ChatGPT4，Claude-2测试。输入测试用例是GLM-4提供，用专业性打败专业性才有趣！以多模态理解，代码解释器，工具调用，逻辑推理方向测试。 1月16日...



# 06、设置需要排除的地址

对搜索引擎设置：<https://programmablesearchengine.google.com/>，设置排除的网址

www.zhihu.com/collection 类网站（收藏列表），  
www.zhihu.com/people 类网站（个人主页），  
www.zhihu.com/column 类网站（个人文章主页），  
www.zhihu.com/topic 类网站(主题)

所有搜索引擎

删除添加

<input type="checkbox"/>	名称	角色	公开网址	上次更新时间 ↓
<input type="checkbox"/>	google-search	所有者		2024年4月12日, 13:30

每页行数10第 1-1 行 (共 1 行) <>

要搜索的网站

删除添加

您没有可供搜索的网站。

要排除的网站

删除添加

您没有已排除的网站。

上传/下载 XML 文件形式的注释

上传/下载 XML 文件形式的搜索引擎上下文

要排除的网站

删除添加

网址包含

清除过滤条件应用过滤条件

<input type="checkbox"/>	网站	上次更新时间
<input type="checkbox"/>	www.zhihu.com/topic*	2024年4月15日, 16:13
<input type="checkbox"/>	www.zhihu.com/column*	2024年4月15日, 16:13
<input type="checkbox"/>	www.zhihu.com/people*	2024年4月15日, 16:13
<input type="checkbox"/>	www.zhihu.com/collection*	2024年4月15日, 16:13

每页行数5第 1-4 行 (共 4 行) <>

## 07、爬虫之Cookie获取

1. 手动登录：首先，你需要在浏览器中手动访问目标网站并登录知乎。
  2. 开发者工具：登录后，打开浏览器的开发者工具。在 Chrome 和 Firefox 中，你可以右键点击页面，然后选择“检查”或“检查元素”。
  3. 网络标签：在开发者工具中，转到“网络”或“Network”标签。
  4. 刷新页面：在开发者工具打开的情况下，刷新页面。这会捕获所有页面加载过程中的网络请求。
  5. 查找请求：在捕获的请求列表中，找到主要的请求（通常是顶部的第一个，或者是与你的目标URL匹配的请求）。
  6. 查看请求头：点击这个请求，然后查找“请求头”或“Request Headers”部分。
  7. 复制Cookie：在“请求头”部分，你应该能看到一个名为 `Cookie` 的字段。你可以直接复制这个字段的值。
  8. 使用在爬虫中：将复制的 `Cookie` 值用于你的爬虫代码中的请求头。
- 需要注意的是`Cookie` 通常有有效期，过了有效期它可能会过期。如果你的爬虫在某个时间点突然无法工作，你可能需要重新获取新的 `Cookie`。



# 08、爬虫之Cookie获取

知乎

首页 知乎知学堂 发现 等你来答

计算机今年究竟炸到什么程度

互联网 人工智能 大数据 AI技术 大模型

哪些 AI 大模型比较有前途?

关注问题 写回答 邀请回答 好问题 添加评论 分享

查看全部 7 个回答

雨飞

大连理工大学 计算机科学与技术硕士

4 人赞同了该回答

目前我比较看好的大模型有 [kimi chat](#)，今天智普发布 [glm4](#)，看起来也是未来可期。

[Kimi chat](#) 我使用已经很长时间了，主要是使用它的[PDF读取](#)和[网页读取](#)的功能，这个大模型支持的上下文是可以达20万字。从我的实际体验下来，应该还是比较好用的。

KimiChat优势

从下图可以看到，能够有效的读取网上链接里PDF的内容，并进行整理。我使用的提示词比较简单，但kimi chat还是将文中重点的内容给总结出来了。

赞同 4 1 条评论 分享 收藏 喜欢 收起

DevTools is now available in Chinese! Always match Chrome's language Switch DevTools to Chinese Don't show again

Elements Console Sources Network Performance Memory Application

Preserve log Disable cache No throttling

Filter Invert 1 Fetch/XHR More filters

2

3

Headers Payload Preview Response Initiator Timing Cookies

Request Headers

answer\_up?question\_tok...

concerned\_followers?lim...

rewards?include=data...

relationship?desktop=true

entity\_word?type=answe...

feeds?include=data%5B\*

draft?include=question...

user-permission

relation

related

inviters

relationship?desktop=true

xiao-mao-tun-yu-48?incl...

53 / 205 requests 34.9 kB / 22

:authority: www.zhihu.com

:method: GET

:path: /commercial\_api/banners\_v3/answer\_up?

:scheme: https

Accept: \*/\*

Accept-Encoding: gzip, deflate, br, zstd

Accept-Language: zh-CN,zh;q=0.9

Cookie: q\_c1=3c10baf5bd084b3cbfe7eece648ba243|1704976541000|1704976541000; \_zap=086350b3-1588-49c4-9d6d-de88f3faae03; d\_c0=AGCYfYvO\_RePTvjfB1kZwPLeke\_N5AM6nwo=|1704949678; \_xsrf=qR1FJHIZ9dvYhhoj4SUj43SAIBUwPOqm; \_\_snaker\_\_id=wNWnamijKBi0kzkl; n\_c1=d44e397edh6740859a7d2a0d4155bfab117065097530001170650

Console What's new

Highlights from the Chrome 121 update

@font-palette-values in Elements

The Elements panel now supports and shows a new section in Styles for the @font-palette-values at-rule.

Source map improvements

# 09、爬虫之user-agent

1. Chrome (Windows 10):

Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36

2. Firefox (Windows 10):

Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:89.0) Gecko/20100101 Firefox/89.0

3. Safari (macOS):

Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.0.3 Safari/605.1.15

4. Edge (Windows 10):

Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36

Edg/91.0.864.59

5. Chrome (Android):

Mozilla/5.0 (Linux; Android 10; SM-A205U) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.120 Mobile

Safari/537.36

6. Safari (iPhone):

Mozilla/5.0 (iPhone; CPU iPhone OS 14\_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/14.1 Mobile/15E148

Safari/604.1

# 10、爬虫之获取PATH

以下是使用Google Chrome浏览器查看和获取XPath的步骤，但其他现代浏览器（如Firefox、Edge等）的操作方式也类似：

## 1. 打开网页:

- 在Chrome中打开你想要查看的网页。

## 2. 打开开发者工具:

- 右键点击页面上的任意位置，然后选择“检查”（或“Inspect Element”）。
- 或者，你可以使用快捷键 `Ctrl + Shift + I`（Windows/Linux）或 `Cmd + Option + I`（Mac）。

## 3. 使用元素选择器:

- 在开发者工具的左上角，你会看到一个鼠标指针图标，这是“选择元素”工具。点击它。
- 然后，将鼠标悬停在页面上的元素上。你会看到元素被高亮显示，并在开发者工具的“Elements”面板中显示其HTML结构。

## 4. 获取XPath:

- 当你找到想要的元素后，在“Elements”面板中右键点击它。
- 在弹出的菜单中，选择“Copy” > “Copy XPath”。这将元素的XPath复制到剪贴板。

## 5. 粘贴XPath:

- 接下来即可将复制的XPath粘贴到你的代码或其他地方。



# 11、爬虫之获取PATH

← → ↻ 🌐 zhihu.com/question/640050703

知乎 首页 知乎知学堂 发现 等你来答 普通人该如何抓住AI这个风口

h1.QuestionHeader-title 674 × 32 芯片 (集成电路) 模型制作

智谱 AI 宣布基座大模型 GLM-4 发布, 该大模型有何功能?

1月16日下午消息, 智谱AI首届技术开放日上, 智谱AI宣布发布新一代基座大模型GLM-4。据介绍, GLM-4的整体性能相比上一代大幅提升, 逼近GPT-...显示全部

关注问题 写回答 邀请回答 好问题 添加评论 分享

你可以邀请下面用户, 快速获得回答 搜索你想邀请的人

委托刘看山邀请 开启后为你智能邀请潜在的回答者 开启邀请

太电AI TA 已帮助 23 人获得解答 邀请回答

AI百科 互联网行业 从业人员 邀请回答

更多推荐结果

DevTools is now available in Chinese! Always match Chrome's language Switch DevTools to Chinese Don't show again

Elements Console Sources Network Performance Memory Application >> 3 1 ⚙️ ⋮ ×

1

2

3

<div class="QuestionStatus"></div>

<div data-za-detail-view-path-module="QuestionDescription" data-za-extra-module="{card":{"content":{"type":"Question","token":"640050703}}}">

<div class="QuestionHeader">

<div class="QuestionHeader-content">

flex

<div class="QuestionHeader-main">

<div class="QuestionHeader-tags"></div> flex

<h1 class="QuestionHeader-title">智谱 AI 宣布基座大模型 GLM-4 发布, 该大模型有何功能?</h1> == \$0

<div class="css-1uovyp5"></div>

<div class="css-1rpg5c6"></div>

<div class="css-dojgf2"></div>

<div class="css-4cffwv"></div>

flex

<div></div>

<div class="QuestionHeader-side"></div>

</div>

<div class="QuestionHeader-footer"></div>

Filter :hov .cls + - [🔍] [🔍]

element.style {

.QuestionHeader-title { main-questi...54a52.css:1 color: #191b1f; line-height: 32px; }

.QuestionHeader-title { main-questi...54a52.css:1 font-size: 22px; font-weight: 600; margin-bottom: 3px; margin-top: 12px; }

h1, h2, h3 { main.216a26...6d113.css:1 font: inherit; margin: 0; }

h1 { user agent stylesheet display: block; font-size: 2em; margin-block-start: 0.67em; margin-block-end: 0.67em; margin-inline-start: 0px; margin-inline-end: 0px; font-weight: bold; }

div.QuestionHeader-main h1.QuestionHeader-title

Console What's new ×

Highlights from the Chrome 121 update

@font-palette-values in Elements

The Elements panel now supports and shows a new section in Styles for the @font-palette-values at-rule.

Source map improvements

67% 1.3 K/s 0.4 K/s W



# 12、完成github网站搜索之排除网址

要排除的网站

删除

添加

网址包含

清除过滤条件

应用过滤条件

<input type="checkbox"/>	网站	上次更新时间
<input type="checkbox"/>	<a href="#">github.com/topic*</a>	2024年4月17日, 13:12
<input type="checkbox"/>	<a href="#">www.zhihu.com/topic*</a>	2024年4月15日, 16:13
<input type="checkbox"/>	<a href="#">www.zhihu.com/column*</a>	2024年4月15日, 16:13
<input type="checkbox"/>	<a href="#">www.zhihu.com/people*</a>	2024年4月15日, 16:13
<input type="checkbox"/>	<a href="#">www.zhihu.com/collection*</a>	2024年4月15日, 16:13

每页行数

5

第 1-5 行 (共 5 行)

<

>

# 13、github token获取

如何获取这些项目的readme文档。这里我们不再需要使用爬虫，而是简单调用Github API即可获得完整的项目readme文档。而使用Github API，则首先需要获取Github token。Github token全称为Github Personal Access Token，用于在调用Github时进行身份验证，其本质上就是一串字符串形式的加密密钥，也就相当于OpenAI API Key

地址：<https://github.com/settings/tokens>

Personal access tokens (classic) Generate new token ▼

Need an API token for scripts or testing? Generate a personal access token for quick access to the [GitHub API](#).

Personal access tokens (classic) function like ordinary OAuth access tokens. They can be used instead of a password for Git over HTTPS, or can be used to [authenticate to the API over Basic Authentication](#).

编辑系统变量

变量名(N): GITHUB\_TOKEN

变量值(V): ghp\_...v3oGl6C

浏览目录(D)...

浏览文件(F)...

确定

取消

New personal access token (classic)

Personal access tokens (classic) function like ordinary OAuth access tokens. They can be used instead of a password over HTTPS, or can be used to [authenticate to the API over Basic Authentication](#).

Note

readme\_check 任意写

What's this token for?

Expiration \*

90 days

The token will expire on Tue, Jul 16 2024

Select scopes

Scopes define the access for personal tokens. [Read more about OAuth scopes](#). 无需勾选任何选项

☐ repo

Full control of private repositories

☐ repo:status

Access commit status

☐ repo\_deployment

Access deployment status

☐ public\_repo

Access public repositories

☐ repo:invite

Access repository invitations

☐ security\_events

Read and write security events

☐ workflow

Update GitHub Action workflows

☐ write:packages

Upload packages to GitHub Package Registry

☐ read:packages

Download packages from GitHub Package Registry

Personal access tokens (classic) Generate new token ▼ Revoke all

Tokens you have generated that can be used to access the [GitHub API](#).

Make sure to copy your personal access token now. You won't be able to see it again!

✓ ghp\_vsI1V0LJqy...32ww3oGl6C

Delete

Personal access tokens (classic) function like ordinary OAuth access tokens. They can be used instead of a password for Git over HTTPS, or can be used to [authenticate to the API over Basic Authentication](#).

# 14、完成HuggingFace网站搜索

The screenshot shows the Hugging Face website interface. The browser address bar displays `huggingface.co/THUDM/chatglm-6b`. The page header includes the Hugging Face logo and a search bar. The main content area shows the model `THUDM/chatglm-6b` with a like count of 2.77k. Below the model name, there are tags for `Transformers`, `PyTorch`, `Chinese`, `English`, `chatglm`, `glm`, `thudm`, `custom_code`, `Inference Endpoints`, `arxiv:2103.10360`, and `arxiv:2210.02414`. The `Model card` tab is selected, showing the model's name `ChatGLM-6B` and links to its `Blog`, `Github Repo`, `Twitter`, and `GitHub` repositories. A section titled `Join our Slack and WeChat` is also present. The `Files and versions` tab is visible, showing the model's architecture and parameters. The `Community` tab shows 106 members. The `Downloads last month` section shows 26,319 downloads. The `Spaces using THUDM/chatglm-6b` section lists 217 spaces, including `multimodalart/ChatGLM-6B`, `yentinglin/Taiwan-LLaMa2`, `eson/tokenizer-arena`, `qingxu98/academic-chatgpt-beta`, `EmbeddedLLM/chat-template-generation`, `meval/multilingual-chatbot-arena-leaderboard`, `yhavinga/dutch-tokenizer-arena`, `Ameaou/academic-chatgpt3.1`, `wangkui/ChatGLM-6B-Int4-API-OpenAI-Compatible`, `CodeTed/chinese_spelling_error_correction`, `MoDongbao/ChatGPT_for_Academic_Releases1`, `zefanwang/gpt-academic`, `Here123/play`, `HMinions/ws`, `caojiachen1/chatgpt-webui`, `gordonchan/embedding-m3e-large`, `ChallengeHub/Chinese-LangChain`, `OptimalScale/Robin-7b`, `ysharma/ChatGLM-6b_Gradio_Streaming`, and `DrSong/ChatGLM-6B-ChatBot`. The page also includes a section for `Experience the larger-scale ChatGLM model at chatglm.cn`.

我们发布了 ChatGLM2-6B, ChatGLM-6B 的升级版本, 在保留了初代模型对话流畅、部署门槛较低等众多优秀特性的基础之上, 引入了更强大的性能、更长的上下文、更高效的推理等升级。

**介绍**

ChatGLM-6B 是一个开源的、支持中英双语问答的对话语言模型, 基于 `General Language Model (GLM)` 架构, 具有 62 亿参数。结合模型量化技术, 用户可以在消费级的显卡上进行本地部署 (INT4 量化级别下最低只需 6GB 显存)。ChatGLM-6B 使用了和 `ChatGLM` 相同的技术, 针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练, 辅以监督微调、反馈自助、人类反馈强化学习等技术的加持, 62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。ChatGLM-6B 权重对学术研究**完全开放**, 在填写[问卷](#)进行登记后**亦允许免费商业使用**。

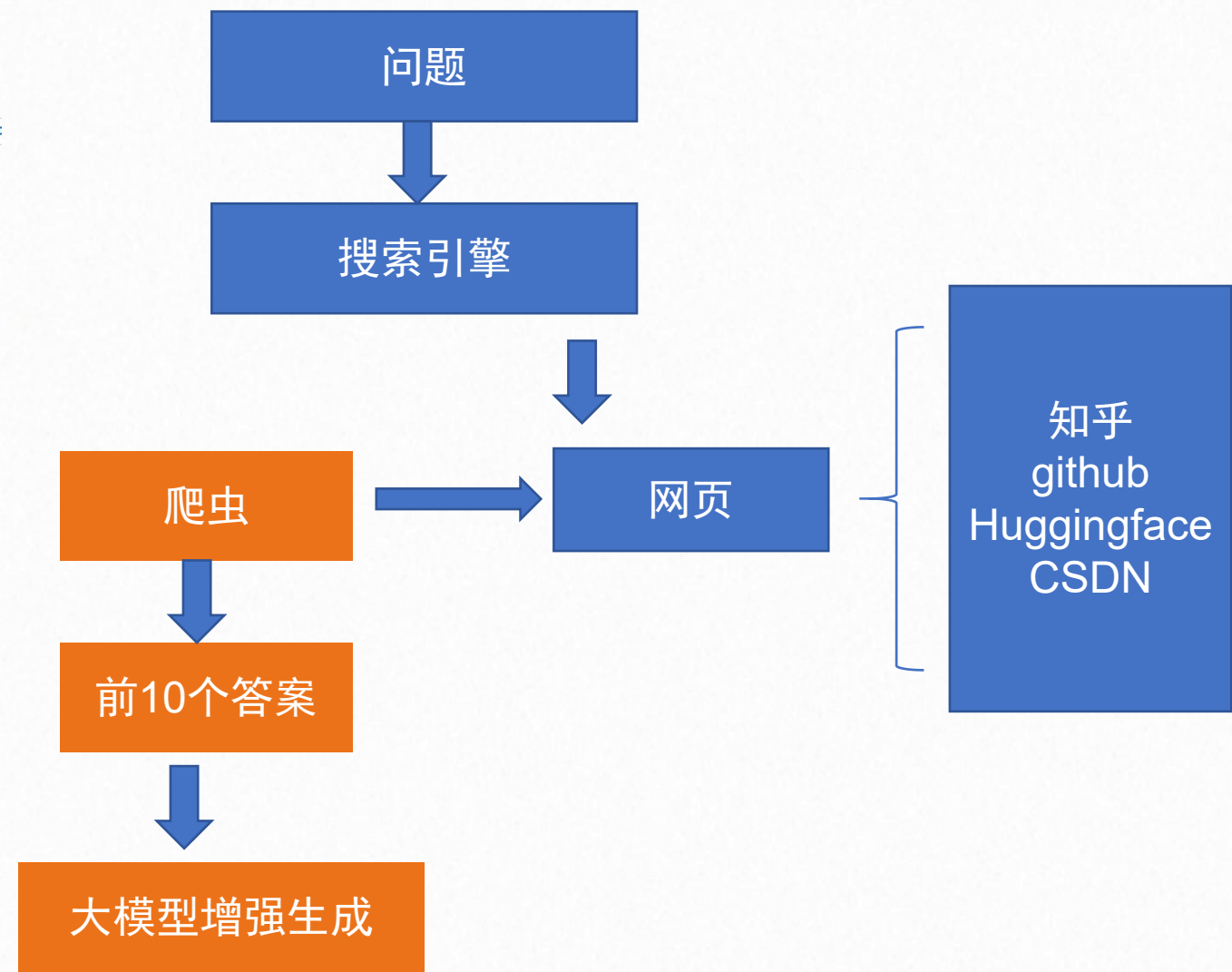
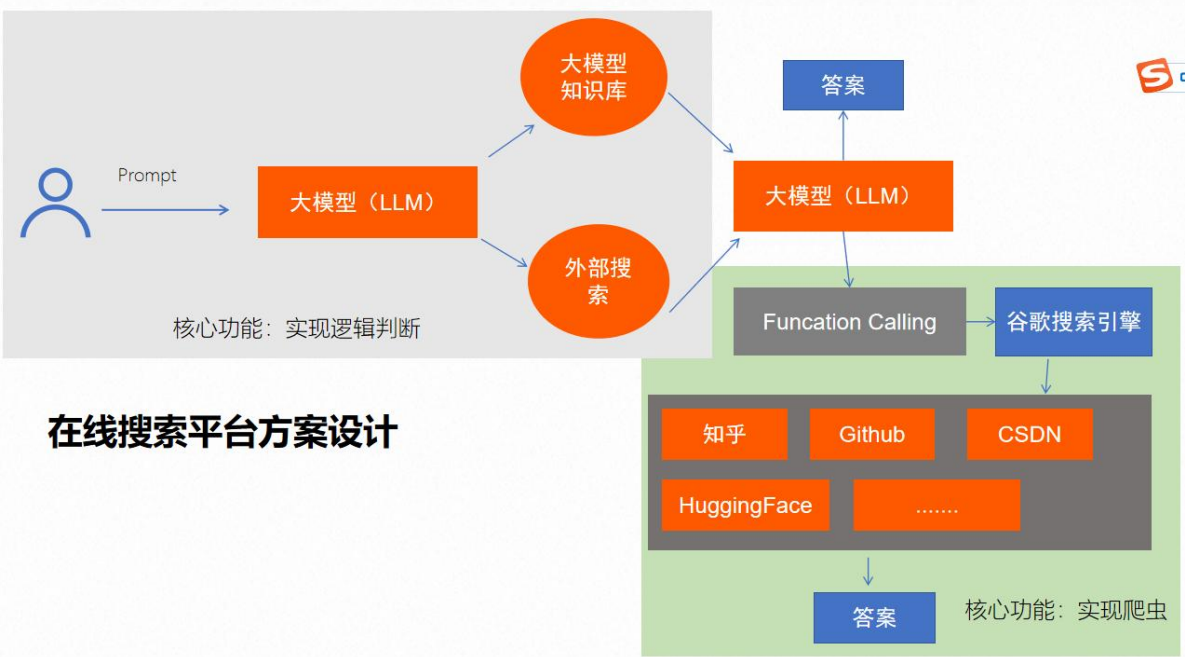
<https://huggingface.co/spaces/Here123/play> en bilingual language model based on `General Language Model (GLM)`.



## **4 在线搜索平台项目总结**



# 01、在线搜索项目总结



## 02、在线搜索平台项目总结

大模型具有不稳定性

功能强大，更考验的是思路

项目中更多是传统技术

## 03、未来展望

打造针对程序员的智能助手（IT智能助手）



能帮你写代码

能帮你检查错误

能帮你答疑解惑

能帮你了解IT技术

**关注视频号：玄姐谈AGI**  
助力数字化人才提升 AIGC 能力



玄姐谈 AGI 



扫一扫二维码，关注我的视频号