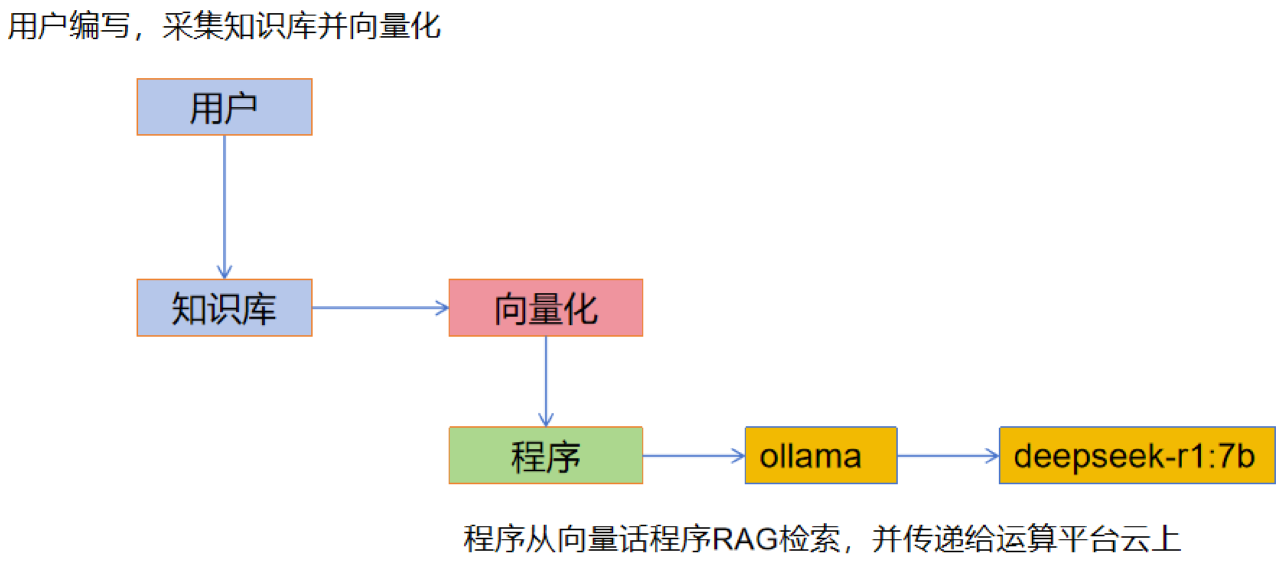
deepseek开发流程

deepseek的工作路径

根据公开资料显示，LLM的工作方式如下。



- 人工部分：

采集和创建知识库，可以是文档，结构化数据等，需要进行段落分割，存放到向量数据库。

- 程序部分

可以是dify等一类的平台，也可以是使用OpenaAI，http接口调用。初创和小公司推荐使用dify。

- 模型运行部分

这个就不用说了，全国会搞模型的也没几个，可以调ollama运行的本地模型，也可以调用云上模型。

- 问题，向量化归哪个环节呢

向量化是归于人工部分还是程序部分呢，知识库被向量化的好坏，对模型有阵非常大的影响，我们可以选择手工向量化，也可以使用程序自动向量化

学习路径

既然我们已经弄清楚了deepseek的工作路径，第一步要做的事情就是准备知识库。

采集知识

```
import requests
import json

bas_url = "http://localhost:3002/v1/scrape"


headers = {"Content-Type": "application/json"}


req_data = {
    "url": "http://www.eweihai.gov.cn/art/2025/3/10/art_159136_5310185.html",
    "formats": ["markdown", "links"],
    "includeTags": [".page-bd.article-bd"],
    "onlyMainContent": True,
}


response = requests.post(bas_url, headers=headers, data=json.dumps(req_data))
print(response.json())
```

构建知识库

编程方式连接到ollama运行模型

```
import openai

base_url = "http://192.168.0.11:11434/v1"
api_key = "sk-"


## 阻塞式
response = client.chat.completions.create(
    model="deepseek-r1:1.5b",
```

```
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Hello!"},
    ],
    stream=False,
)

print(response.choices[0].message.content)



## 流式
# response = client.chat.completions.create(
#     model="deepseek-r1:1.5b",
#     messages=[
#         {"role": "system", "content": "You are a helpful assistant."},
#         {"role": "user", "content": "Hello!"},
#     ],
#     stream=True,
# )

# for chunk in response:
#     # 检查块中是否有内容
#     if chunk.choices and chunk.choices[0].delta.content:
#         print(chunk.choices[0].delta.content)
```

版本号 #7
由 董列涛 创建于 10 三月 2025 13:09:16
由 董列涛 更新于 18 九月 2025 03:45:38