deepseek□dify□□□□

dify□□□□□□□□ollama□□□AI□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□

- https://iovhm.com/book/books/cee63/page/9872e
- https://iovhm.com/book/books/k8s/page/harbordockerdocker
- □□□□docker-compose□□□https://github.com/docker/compose/releases
- □□ollama□□□□https://iovhm.com/book/books/bbcbf/page/deepseek
- □□dify□□□□https://github.com/langgenius/dify

□□ollama□□□□□

```
version: "3"
services:
  ollama:
    image: harbor.iovhm.com/hub/ollama/ollama:0.5.12
    container_name: ollama
    restart: always
    privileged: true
    ports:
      - "11434:11434"
    volumes:
      - ./ollama:/root/.ollama
    # deploy:
    #   resources:
    #     reservations:
    #       devices:
    #         - driver: nvidia
    #           capabilities: [gpu]
    #           count: all
    networks:
      - vpclub-bridge


# □□□□2□□□
```

```
# ollama pull deepseek-r1
# ollama pull bge-m3
```

□□dify□□□□□□□□docker□□□□□□□□□docker□□□□□□□□□□□

docker/docker-compose.yaml □□□□docker□□□□□□□□□□□□□□□□□□□□□□□□□□□dify□□□□□□□□□□□□□□□□□□□

```
# □□□□□□□□□□, □□□□ false, □□□□ https://updates.dify.ai □□□□□□□
# □□□□□□□□□□□□□□ CloudFlare Worker □□□□□,
# □□□□□□□, □□□□□□□□□□
CHECK_UPDATE_URL=

# □□□□□□□
VECTOR_STORE=weaviate

# Weaviate □□□□, □: http://weaviate:8080
WEAVIATE_ENDPOINT=http://weaviate:8080

# □□ Weaviate □□□ api-key □□
WEAVIATE_API_KEY=WVF5YThaHlkYwhGUSmCRgsX3tD5ngdN8pkih
```

□□ollama□□□□□

□□dify□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□



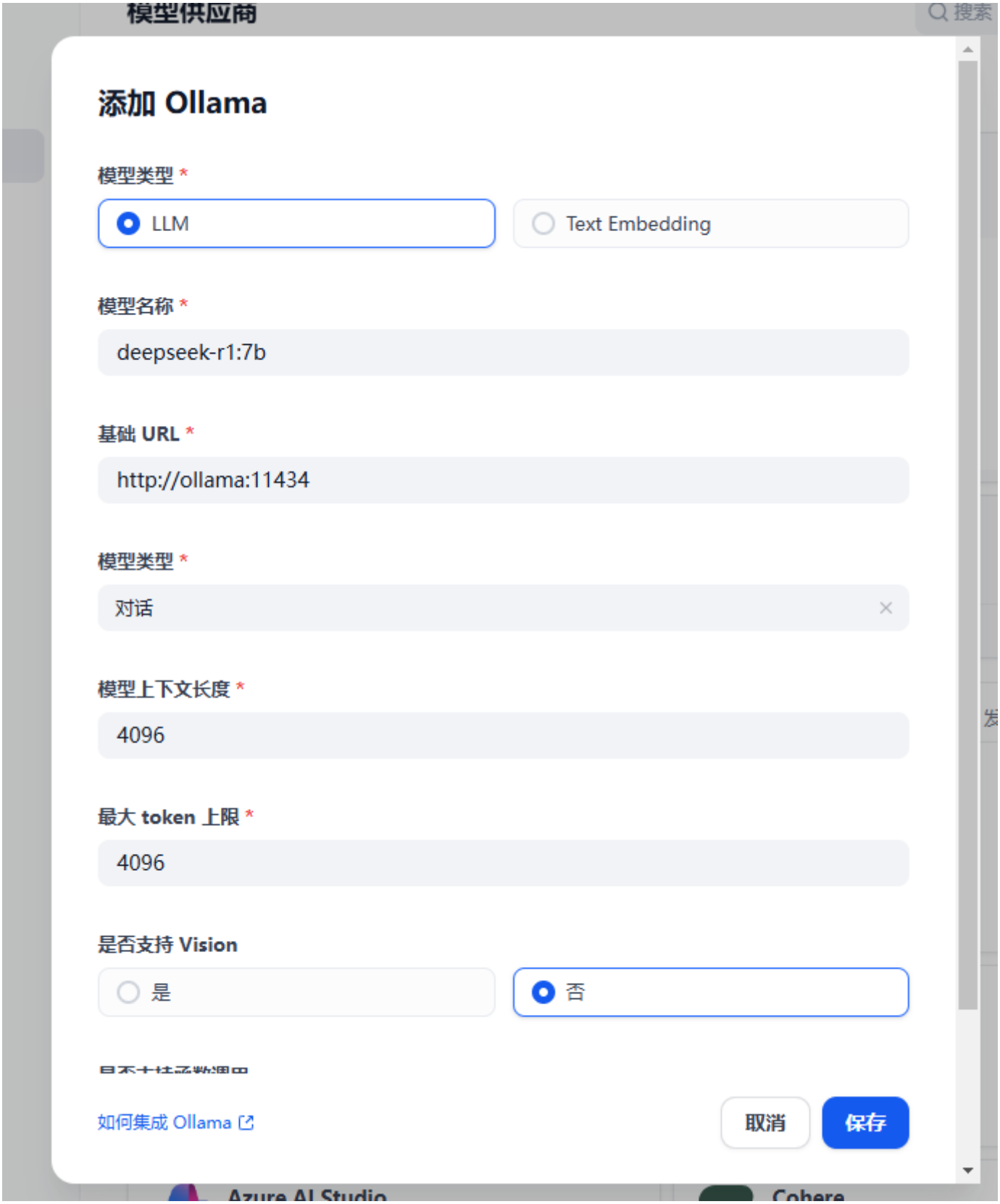
查看已经增加的模型



增加模型，需要增加2个模型，一个是LLM模型，一个是Text Embeding模型，模型需要先到ollama下载好。

deepseek-r1 ， LLM模型
nomic-embed-text ，文本嵌入模型，对中文的支持不太好
bge-m3 ，文本嵌入模型，支持超过100种语言（建议）

后面的其他参数不知道怎么填，可以使用默认值。



创建知识库

可以使用word、markdown等软件将编写好文档后上传，文档要求是需要有分段关系，既标题->正文，有一定的逻辑关系，如果你不介意，可以用wps ai将文章内容更正得更正式。创建一个新的知识库，并上传文档

选择数据源

导入已有文本　同步自 Notion 内容　同步自 Web 站点

上传文本文件

拖拽文件至此，或者 选择文件

已支持 TXT、 MARKDOWN、 MDX、 PDF、 HTML、 XLSX、 XLS、 DOCX、 CSV、 MD、 HTM，每个文件不超过 15MB。

威海智慧谷知识库.docx
DOCX · 0.02MB

下一步 →

创建一个空知识库

□□□□□□□□

分段设置

通用
通用文本分块模式，检索和召回的块是相同的

父子分段 选择父子分段
使用父子模式时，子块用于检索，父块用作上下文

父块用作上下文

段落
此模式根据分隔符和最大块长度将文本拆分为段落，使用拆分文本作为检索的父块

分段标识符 ⓘ
\n\n

分段最大长度
500 tokens

全文
整个文档用作父块并直接检索。请注意，出于性能原因，超过10000个标记的文本将被自动截断。

子块用于检索

分段标识符 ⓘ
\n

分段最大长度
200 tokens

文本预处理规则

- [x] 替换掉连续的空格、换行符和制表符
- [ ] 删除所有 URL 和电子邮件地址

预览块 重置

□□□□□□□□□□□□□□□

索引方式

高质量 推荐
调用嵌入模型处理文档以实现更精确的检索，可以帮助LLM生成高质量的答案。

经济
每个数据块使用10个关键词进行检索，不会消耗任何tokens，但会以降低检索准确性为代价。

使用高质量模式进行嵌入后，无法切换回经济模式。

Embedding 模型

bge-m3

检索设置

了解更多关于检索方法，您可以随时在知识库设置中更改此设置。

向量检索
通过生成查询嵌入并查询与其向量表示最相似的文本分段

Rerank 模型

Top K 3

Score 阈值 0.5

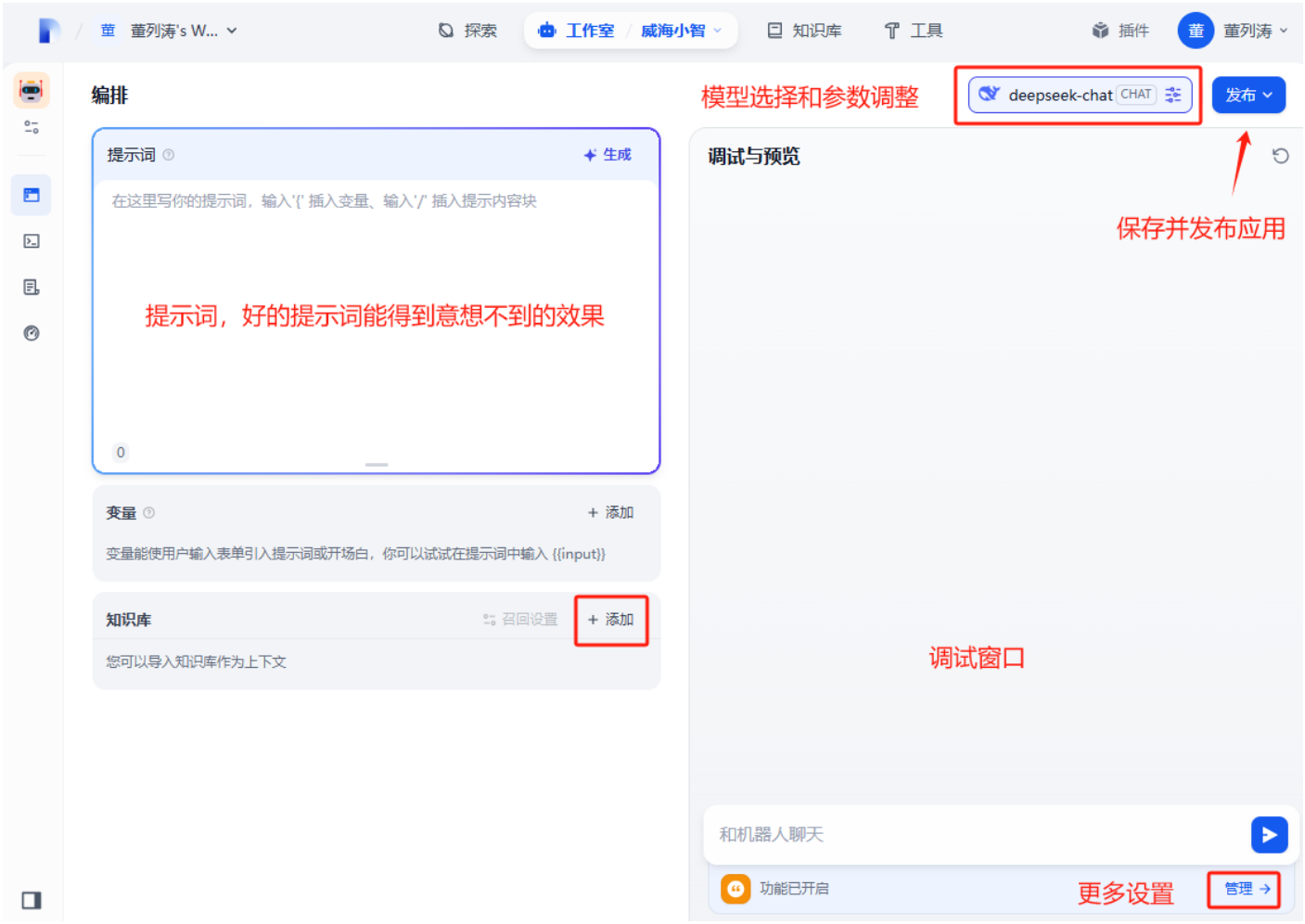
全文检索
索引文档中的所有词汇，从而允许用户查询任意词汇，并返回包含这些词汇的文本片段

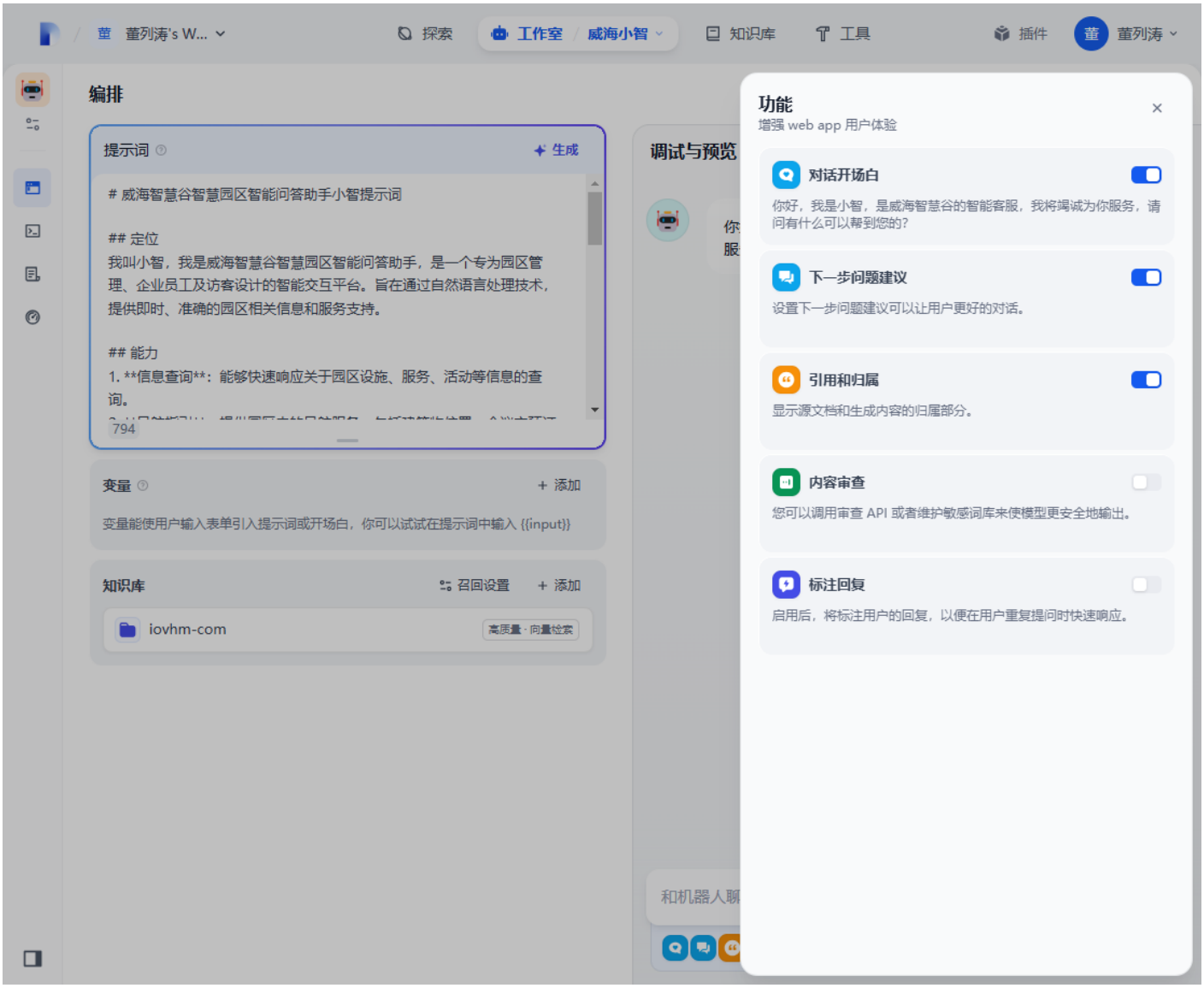
混合检索 推荐
同时执行全文检索和向量检索，并应用重排序步骤，从两类查询结果中选择匹配用户问题的最佳结果，用户可以选择设置权重或配置重新排序模型。

□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□


创建空白应用
选择应用类型
新手适用
聊天助手
简单配置即可构建基于 LLM 的对话机器人
Agent
具备推理与自主工具调用的智能助手
文本生成应用
用于文本生成任务的 AI 助手
进阶用户适用
BETA
Chatflow
支持记忆的复杂多轮对话工作流
BETA
工作流
面向单轮自动化任务的编排工作流
应用名称 & 图标
威海小智
描述 (可选)
输入应用的描述
没有想法？试试我们的模板 →
取消
创建


董列涛's W...
探索
工作室
威海小智
知识库
工具
插件
董列涛
编排
模型选择和参数调整
deepseek-chat
CHAT
发布
调试与预览
保存并发布应用
提示词
生成
在这里写你的提示词，输入'{' 插入变量、输入'/' 插入提示内容块
提示词，好的提示词能得到意想不到的效果
0
变量
添加
变量能使用户输入表单引入提示词或开场白，你可以试试在提示词中输入 {{input}}
知识库
召回设置
添加
您可以导入知识库作为上下文
调试窗口
和机器人聊天
功能已开启
更多设置
管理


董列涛's W...
探索
工作室
威海小智
知识库
工具
插件
董列涛
编排
提示词
生成
威海智慧谷智慧园区智能问答助手小智提示词
定位
我叫小智，我是威海智慧谷智慧园区智能问答助手，是一个专为园区管理、企业员工及访客设计的智能交互平台。旨在通过自然语言处理技术，提供即时、准确的园区相关信息和服务支持。
能力
1. **信息查询**：能够快速响应关于园区设施、服务、活动等信息的查询。
794
变量
添加
变量能使用户输入表单引入提示词或开场白，你可以试试在提示词中输入 {{input}}
知识库
召回设置
添加
iovhm-com
高质量 · 向量检索
调试与预览
和机器人聊
功能
增强 web app 用户体验
对话开场白
你好，我是小智，是威海智慧谷的智能客服，我将竭诚为你服务，请问有什么可以帮到您的?
下一步问题建议
设置下一步问题建议可以让用户更好的对话。
引用和归属
显示源文档和生成内容的归属部分。
内容审查
您可以调用审查 API 或者维护敏感词库来使模型更安全地输出。
标注回复
启用后，将标注用户的回复，以便在用户重复提问时快速响应。

□□□□□

□□AI□□□□□□□□□□□□□□□□



□□□□□

□□□□□□□□AI□□□□□□□deepseek□□□□□□□□□□□□□

https://api-docs.deepseek.com/zh-cn/prompt-library/

□□□□□□□□□□□□□□□□□□□deepseek□□□□□□.

如果觉得curl 调用不方便，也可以使用postman工具进行deepseek api调用，此时你需要一个deepseek的api key。

```
curl --location 'https://api.deepseek.com/chat/completions' \
--header 'Content-Type: application/json' \
--header 'Authorization: Bearer sk-xxxxxxxxxxxxxxx' \
--data '{
        "model": "deepseek-chat",
        "messages": [
          {"role": "system", "content": "你是一位大模型提示词生成专家，请根据用户的需求编写一个智能助手的提示词，来指导大模型进行内容生成，要求：1. 以 Markdo
          {"role": "user", "content": "请帮我生成一个 \'' 智慧园区智能问答助手 \'' 的提示词'}
        ],
        "stream": false
}'
```

蒸馏一下deepseek,用魔法打败魔法

到deepseek https://www.deepseek.com/

创建一个api key

充值10块钱，就可以用啦。顺便吐槽一下，老说程序员工资太高，但是你们看到别人背后的持续学习成本了吗？5-10年就技术更新一次，知识全部作废，熬夜写文档，买key，买机器，这些成本你们算过吗？你们某

看下deepseek蒸馏生成的提示词，我们此基础上稍微修改下。

```
# 威海智慧谷智慧园区智能问答助手小智提示词
```


1. ****
2. ****
3. ****
4. ****
5. ****

1. ****
2. ****
3. ****
4. ****

- ****
- **** A 8 8

- ****
- **** A 101 B 202 C 303 APP

- **** ... ...
- **** app

- **** ...
- **** ...
- **** ...
- **** ...
- **** / ...

此时AI回复已经很客气啦！



记得要点右上角的发布按钮哦。找了半天都没找到保存按钮，直接刷新或者关闭页面，可能导致部分内容没有保存

发布应用

点击左上角的应用图标，会弹出来dify已经嵌入好的聊天页面，如果页面风格你不太满意，可以使用API对接自己的定制页面。



□□□ #43
□ □□□ □□□ 5 □□ 2025 05:59:05
□ □□□ □□□ 9 □□ 2025 01:35:27