

展映 AI 本地服务安装、使用说明

当前企业对 AI 需求越来越高，如何创建一个企业 AI 工作流程，用 AI 辅助参与信息管理，提升效率，同时软件工具更可控，做到低成本、本地化、大规模使用是提升企业信息效率的一个更好的方式。深圳展映在这个领域一直投入大量资源帮助企业获得更多 AI 落地及应用。如何能在成本和效率、体验上、产品使用可控找到最适合的产品及服务，是我们深圳展映科技一直探索的方向。现在我们推出了本地 AI 服务，希望通过我们免费的合作伙伴系列产品为大家带来更好的 AI 服务能力。同时也欢迎使用我们展映在线 AI 服务：<https://ai.zyinfo.pro>。

本次 AI 服务功能比较强大，而且不需要高端 GPU 也能在 windows 系统上运行，需要较少的内存及 CPU 计算，并且能有较快的 AI 回复速度，较好的质量，满足我们的正常场景下 检索、文章生成、AI 工作流等场景的使用。后续我们一直也会推进相关产品的更新、推荐最好的产品及新的动态，欢迎关注我们的微信公众号：展映科技，获取最新的动态。也欢迎与我们客服互动~

特别说明：部分合作方产品来自市场公开产品，都会选择至少有数十万用户、客户的产品供应方。我们自有的软件、服务都经过认真的安全审核，尽可能不使用第三方的系统工具，保障服务质量和安全。每个软件包我们都查阅了背后的开发者和部分代码及用户反馈，付出了较多的成本和精力，为的是给企业用户提供更多一层保障。

好的，现在开始了解下如何安装及体验最新的本地 AI 服务吧~

小提示：本文档最新更新地址：<https://docs.qq.com/doc/DS3Bsa1ZUWU1JYVhS>

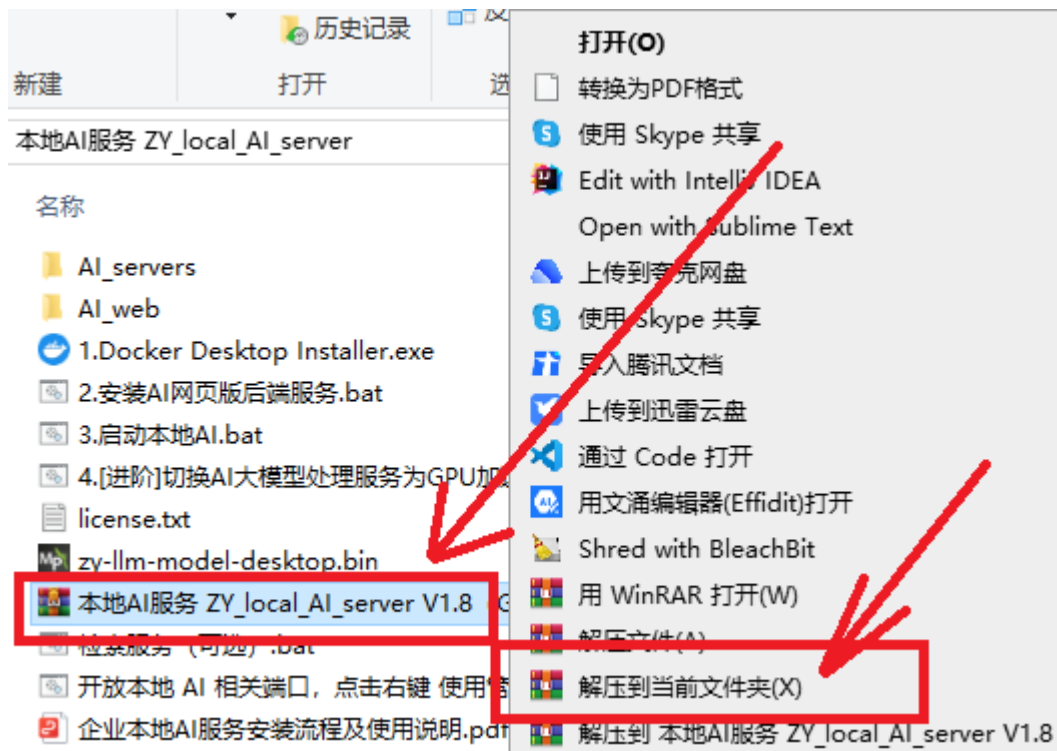
本说明文档相关产品的下载：

链接：<https://pan.baidu.com/s/1k5jnh36kcU0kOEE3IDivMQ?pwd=5566> 提取码：5566

或链接：<https://pan.quark.cn/s/8d413221daff>

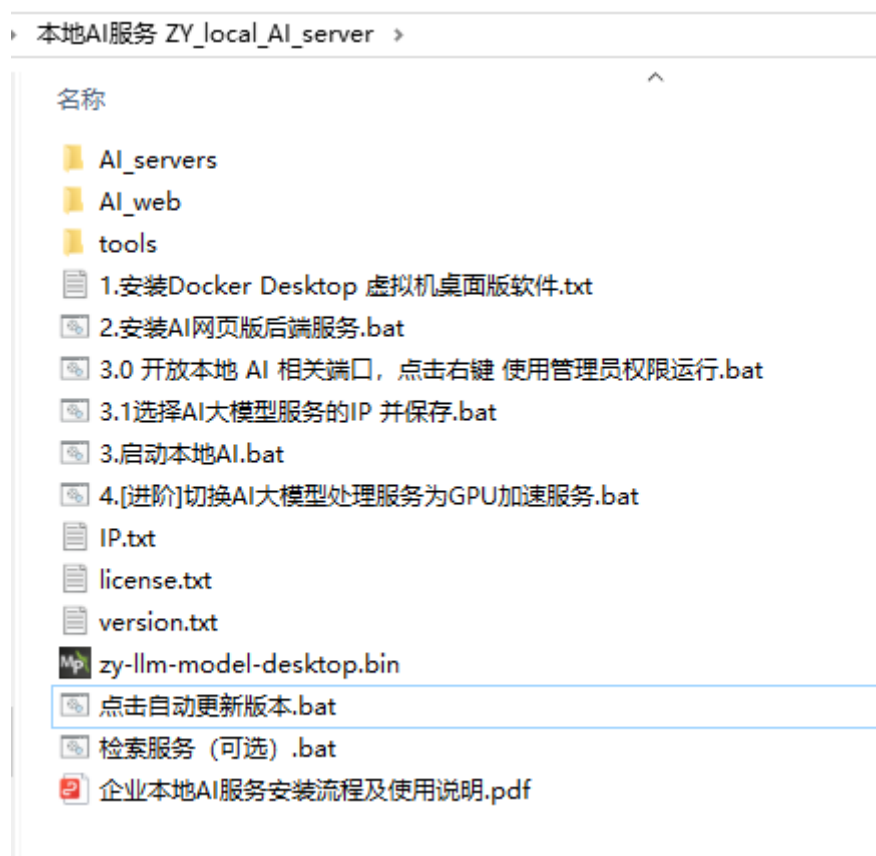
安装流程：

先下载安装包 V1.7、V1.8，由于 V1.8 是更新包。需要覆盖 V1.7 的内容。V1.7 解压后，将 V1.8 rar 移动到文件夹中，点击解压到当前文件夹：



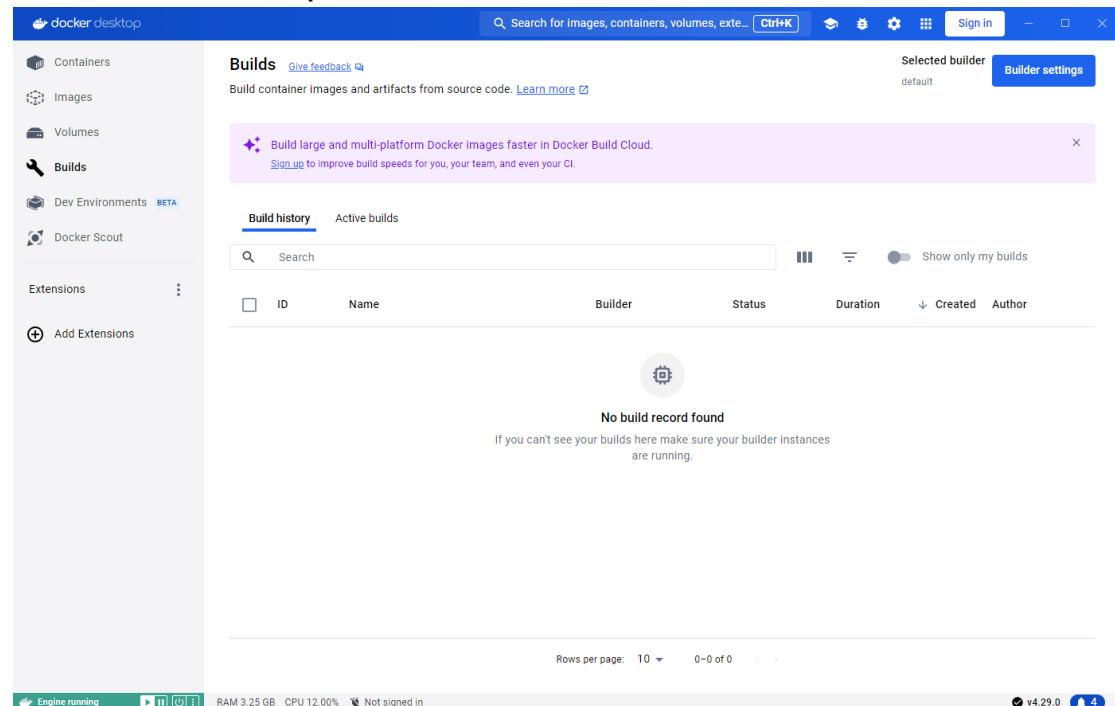
解压到当前文件夹、覆盖当前文件夹文件

(后续更新可点击自动更新.bat)



AI 服务软件包资源如图所示 (按文件名从小到大排序)

1、安装 Docker Desktop Installer.exe



或者前往：<https://docs.docker.com/desktop/install/windows-install/>获取最新版本。
安装完毕后点击运行，可看到类似如上界面。

2、启动 Docker engine

注意看上图左下角，是否是 **running** 正在运行。

3、点击 安装 AI 网页版后端服务.bat

后续后端升级也可直接点击那个

如果无法拉取新的镜像，可能需要配置国内源。

若实在无法使用，可能需要外网环境。可联系我们远程提供合作伙伴的解决方案：
微信 **youkpan** 。 Email: tel_pan@126.com

4、点击 “3.0 开放本地 AI 相关端口，点击右键 使用管理员权限运行”

5、点击 "3.1 选择 AI 大模型服务的 IP 并保存.bat"

```
C:\WINDOWS\system32\cmd.exe
"请打开 Docker Desktop 查看Docker引擎是否启动完成。AI服务是否启动完毕，确认
请回车继续"
Press any key to continue . . .
1. 192.168.56.1
2. 192.168.11.1
3. 192.168.188.1
4. 169.254.147.120
5. 192.168.18.223
6. 172.20.80.1
请选择你网卡局域网的IP地址（输入对应数字）：5
选择的IP地址是：192.168.18.223
ECHO is off.
OPENAI 兼容的接口 请填写：http://192.168.18.223:7900/v1
ECHO is off.
Press any key to continue . . .
```

配置 AI 模型服务 IP 示例（选择你的无线网卡或有线网卡 IP）

你可以在命令行输入 `ipconfig` 查看你的本地网卡 局域网地址确认。（比如无线网卡，有线网卡等，但不能是虚拟网卡）

如图：本地兼容 openAI 服务的接口是：`http://192.168.18.223:7900/v1` （示例）

6、点击 启动本地 AI.bat

后续可直接运行此命令行，可点击右键发送到桌面快捷方式

会自动打开几个网站，最后那个是本地的 AI 服务页面了。

也可以手动打开 <http://127.0.0.1>。

同时也欢迎体验我们展映科技的 AI 网页版服务 <https://ai.zyinfo.pro> 以及流程图（智能视频剪辑：<https://docs.qq.com/doc/DS3BIbE9JcWZ6TkNM>）。

此时会出现一个这样的窗口，**请不要关闭它**：

```
our site is : http://zyinfo.pro ,
contact us :youkpan@gmail.com or wechat :youkpan.

llm_load_vocab: mismatch in special tokens definition ( 3528/122753 vs 259/122753 ).
llm_load_tensors: ggml ctx size = 0.18 MiB
llm_load_tensors: CPU buffer size = 1646.66 MiB
.....
llama_kv_cache_init: CPU KV buffer size = 1440.00 MiB
{"tid":"12652","timestamp":1714445929,"level":"INFO","function":"main","line":3015,"msg":"model ok"}
{"tid":"12652","timestamp":1714445929,"level":"INFO","function":"update_slots","line":1807,"msg":"slots are empty"}
-
```

这是本地 AI 服务的后端程序，需要保持运行

在浏览器中，<http://127.0.0.1> 的网址下，会出现如下界面：

设置管理员账户

管理员拥有的最大权限，可用于创建应用和管理 LLM 供应商等。

邮箱

用户名

密码



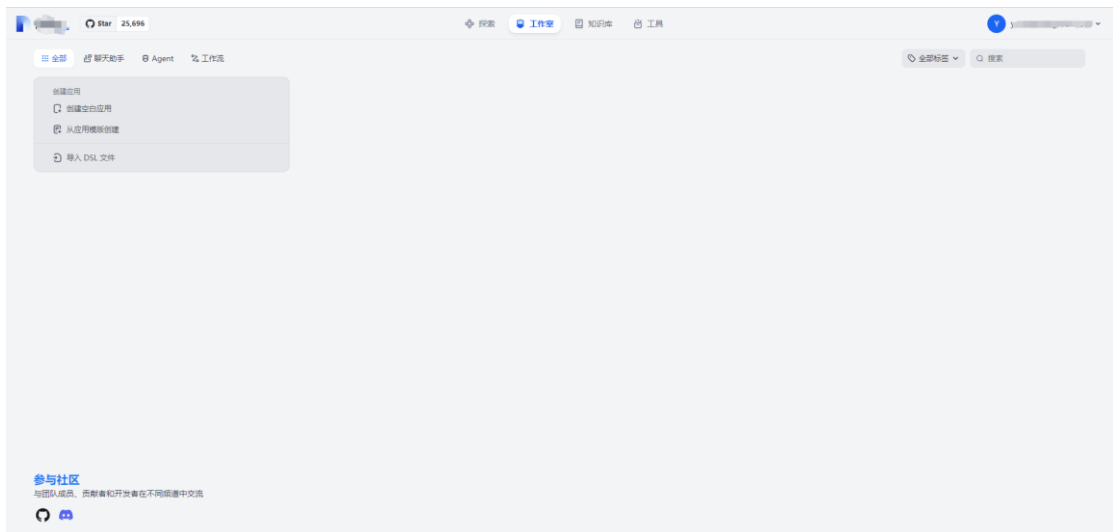
密码必须包含字母和数字，且长度不小于8位

设置

启动 Dify 社区版之前, 请阅读 [GitHub 上的 开源协议](#)

设置 AI 服务的管理账户

注册后，再登录即可使用了。



进入后，可看到如下工作页面



点击右上角 设置，开始设置我们本地 AI 服务



选择模型供应商



找到 openAI 适配的接口

添加 OpenAI-API-compatible OpenAI-API-compatible

模型类型 *

☒ LLM ☐ Text Embedding

模型名称 *

zhangying-AI-llm

API Key

no-key

API endpoint URL *

http:// 192.168.18.223:7900/v1

Completion mode

对话

模型上下文长度 *

4096

最大 token 上限 *

4096

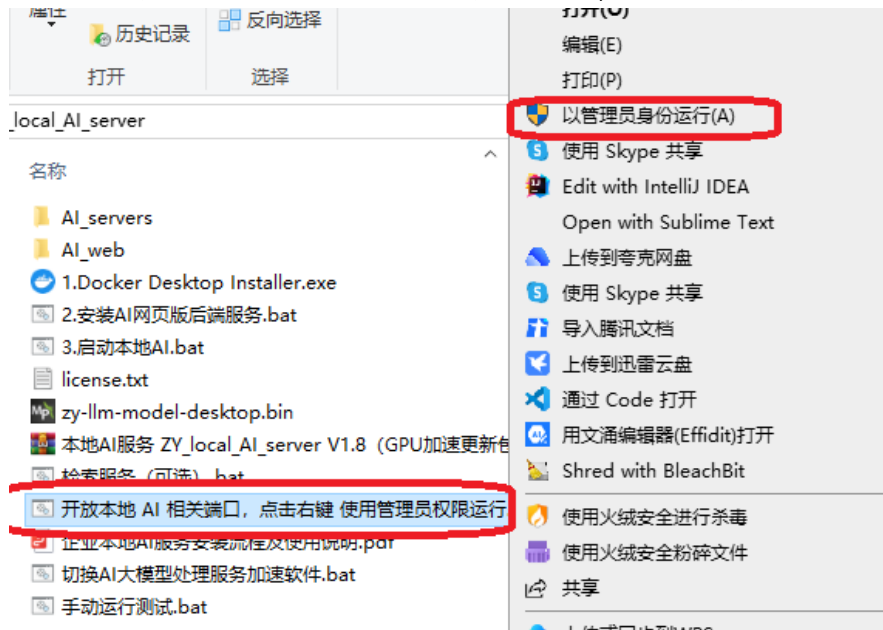
模型填写 zhangying-AI-llm, key 填写 no-key。

API 的 IP 替换成你自己的选择的 IP 信息。

比如 `http://192.168.18.223:7900/v1` (注意网址中间不能有空格, 替换红色部分)

点击保存即可, 第一次配置时, 后端服务会启动, 可能需要点时间, 若无响应, 再试试。

若无法连接可以先尝试开放 windows 系统端口（7900,7901 已提供自动化命令）：



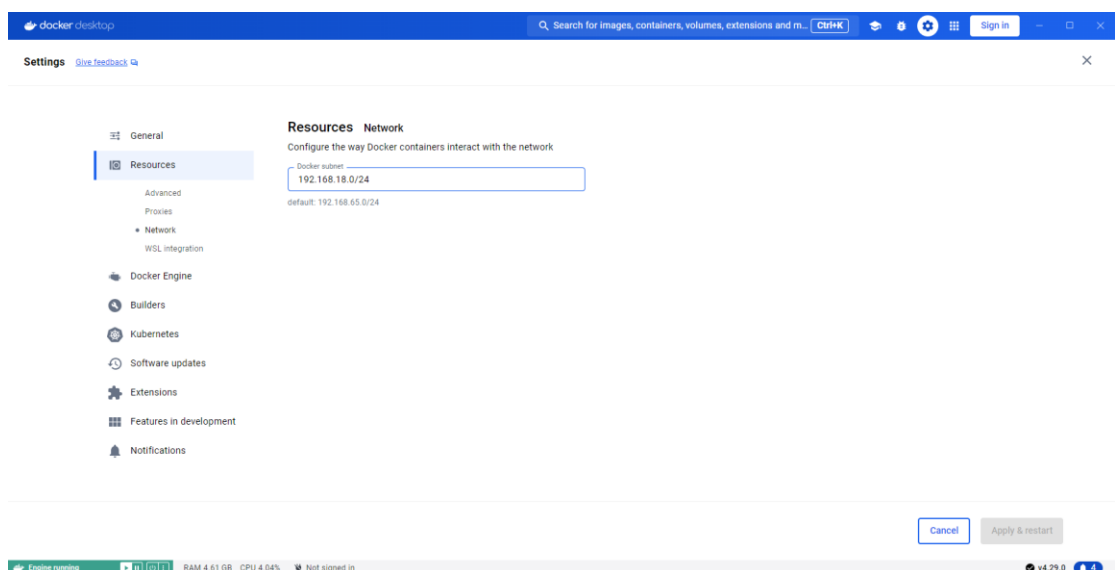
找到：开放本地 AI 相关端口，点击右键 使用管理员权限运行.bat
点击右键用管理员方式运行

可使用：https://v.stylee.top:8899/llm_api/test/test/qwen/v1 代替测试

推荐使用我们的 AI 服务在线接口，质量更高，还能选择 30 多款高级 AI 服务模型：

[展映智慧助手大模型接口 LLM API](#)

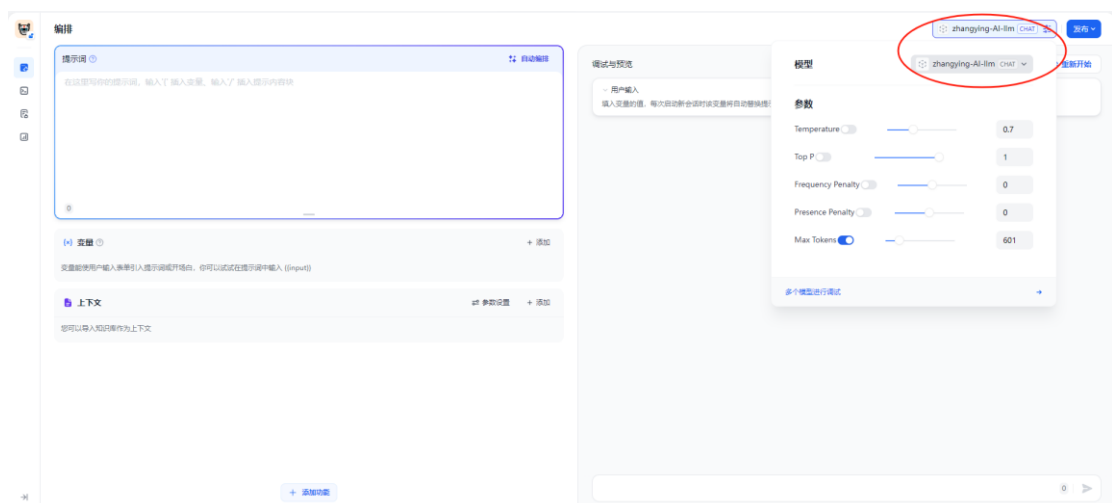
如果还是不行，基本不出现这个问题了，
也可设置 docker desktop 与内部容器的网络，点击右上角设置：



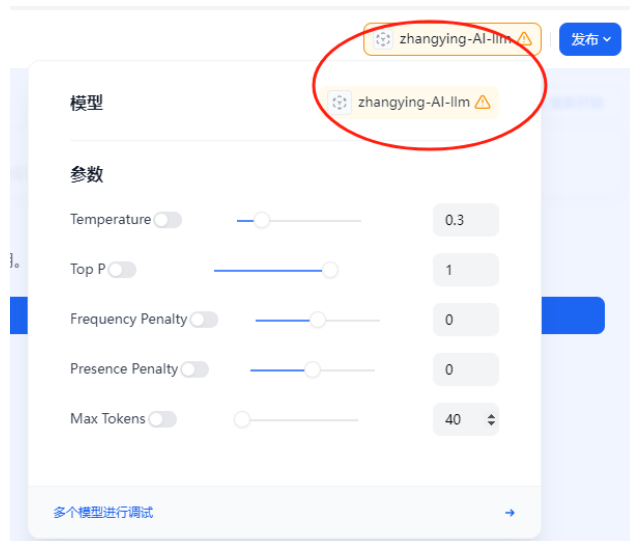
将其设定为与你兼容的网络试试
比如 将：192.168.50.0/24 改为：192.168.18.0/24
假设我们 IP 是 192.168.18.223



接着，我们创建一个聊天对话试试



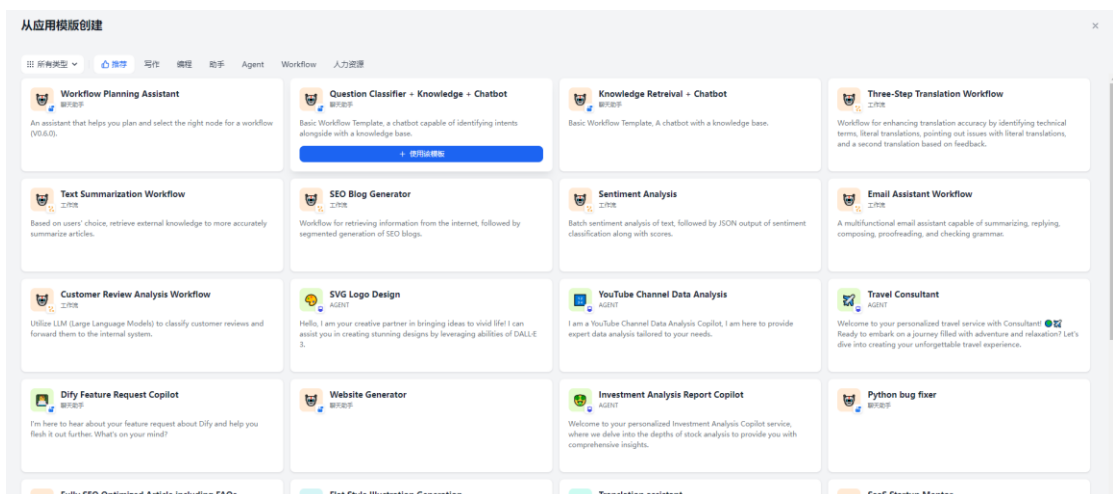
在右上角 选择你的 AI 服务，



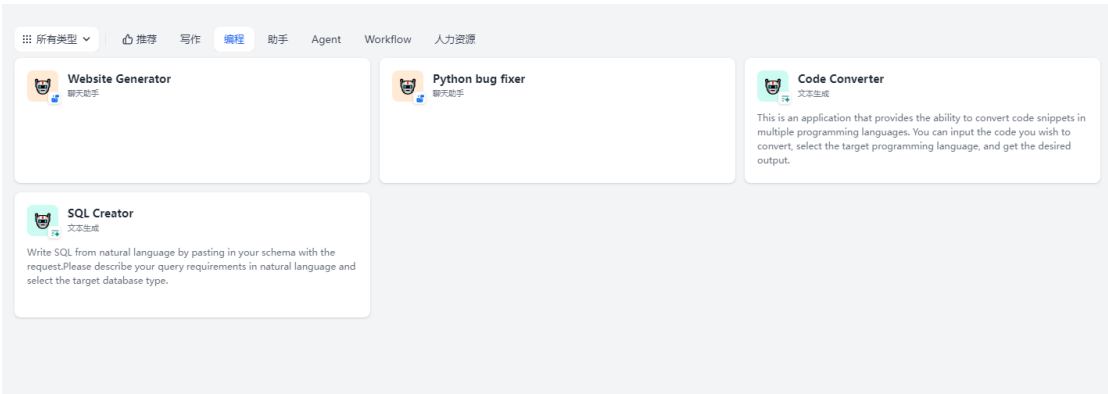
这样填可能比较简单好用，最后一个选项是用来设定每次 AI 输出长度。一般在 600 以内，它如果输出不够再和它说 继续。



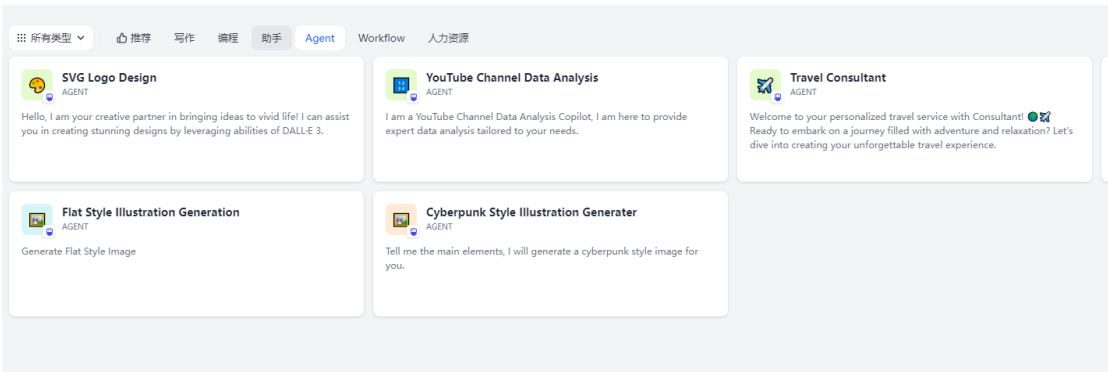
在右侧输入 你好，就能看到回复了！
在你本地运行的 AI 服务正常工作啦！



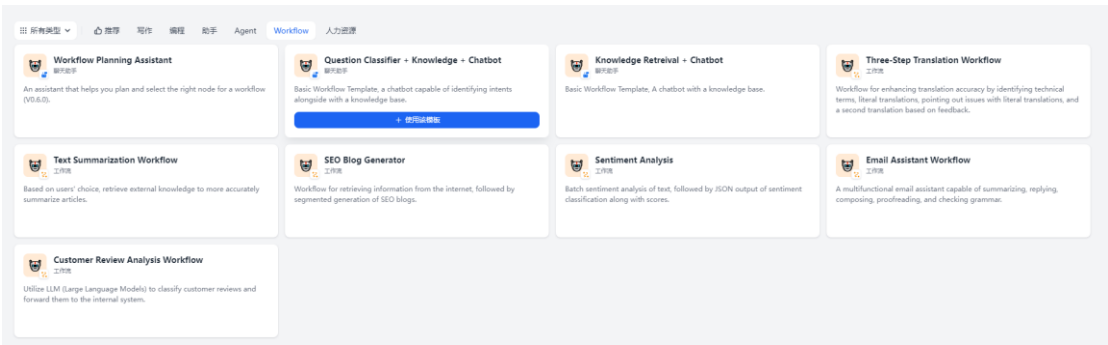
接下来可以自行体验里面的产品吧，有多种模板



编程、协作 助手



高级版助理



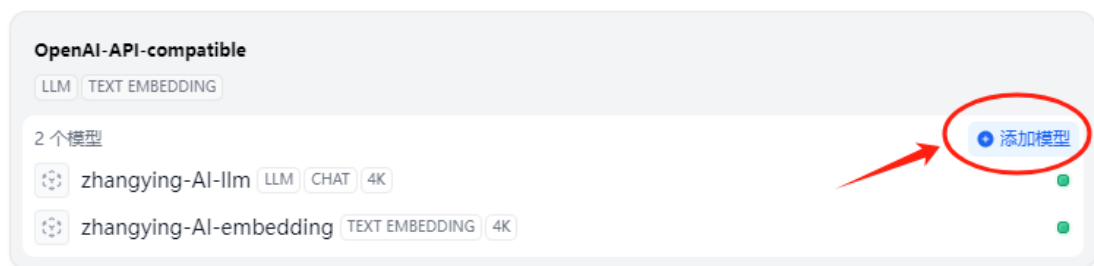
工作流等



还可以添加本地知识库（信息都保存在本地，需要开启本地检索接口）

添加本地信息检索接口：

添加本地信息检索接口，点击设置：



双击打开：检索服务（可选）.bat

添加 OpenAI-API-compatible

OpenAI-API-compatible

模型类型 *

LLM

Text Embedding

模型名称 *

zhangying-AI-embedding

API Key

在此输入您的 API Key

API endpoint URL *

http://192.168.18.223:7901/v1

模型上下文长度 *

4096

取消

保存

您的密钥将使用 PKCS1_OAEP 技术进行加密和存储。

示例 模型填写 zhangying-AI-embedding，key 填写 no-key。

填写完毕后，回到原来知识库界面

创建知识库

1 选择数据源

2 文本分段与清洗

3 处理并完成

文本分段与清洗

分段设置

自动分段与清洗

自动设置分段规则与预处理规则，如果了解这些参数建议选择此项

自定义

自定义分段规则、分段长度以及预处理规则等参数

索引方式

高质量 推荐

调用系统默认的嵌入接口进行处理，以在用户查询时提供更高的准确度

请先完成模型供应商的 API KEY 设置。前往设置

经济

使用离线的向量引擎、关键词索引等方式，降低了准确度但无需花费 Token

执行嵌入预估消耗 0 tokens

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。

倒排索引

倒排索引是一种用于高效检索的结构。技术语组织，每个术语指向包含它的文档或网页

Top K

3

预处理文档

test

预估分段数

1

13

按默认设置试试

← 创建知识库

1 ✓ 选择数据源

2 ✓ 文本分段与清洗

3 处理并完成

 知识库已创建

我们自动为该知识库起了个名称，您也可以随时修改

知识库名称

test.txt...

嵌入已完成 🔥 经济模式 · 预估消耗 0 tokens

test.txt

100%

分段规则

自动

分段长度

500

文本预定义与清洗

自动

前往文档 →

导入本地文档数据测试

文档

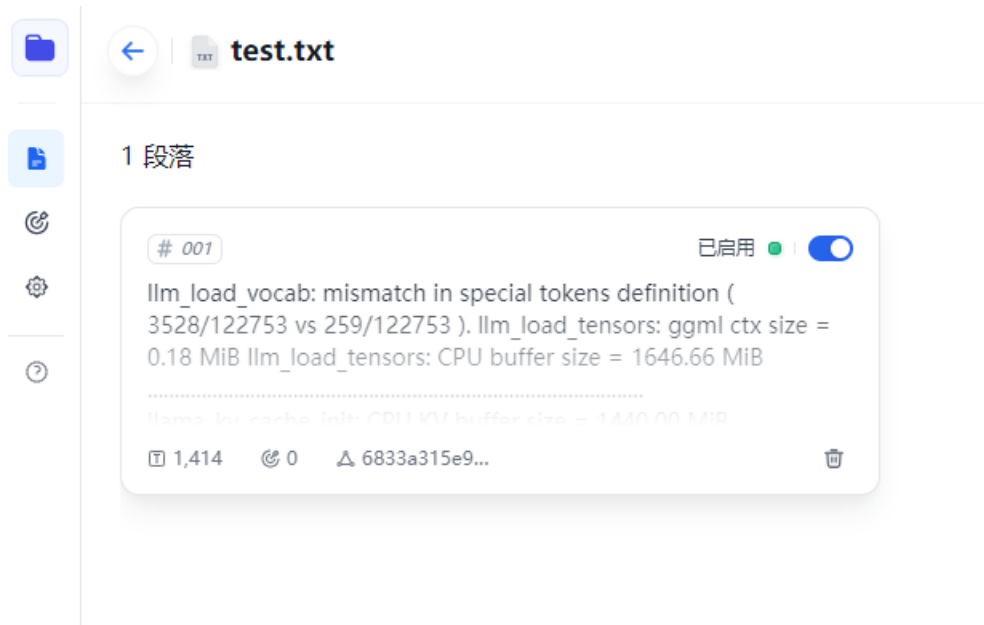
知识库的所有文件都在这里显示。整个知识库都可以链接到 Dify 应用或通过 Chat 操作进行索引。

🔍 搜索

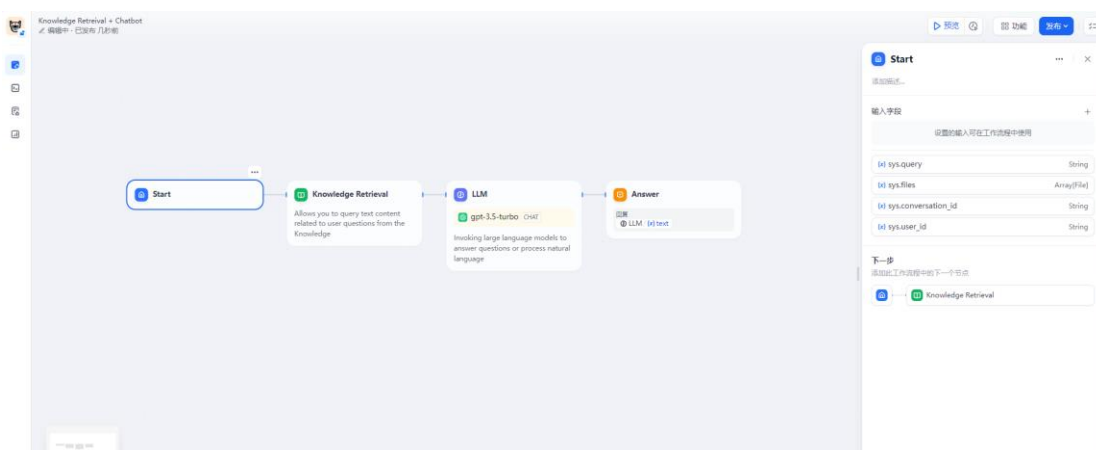
+ 添加文件

#	文件名	字符数	回顾次数	上传时间	状态	操作
1	test.txt	1.4k	0	2024-04-30 11:42	可用	🔵 ...

查看文件列表



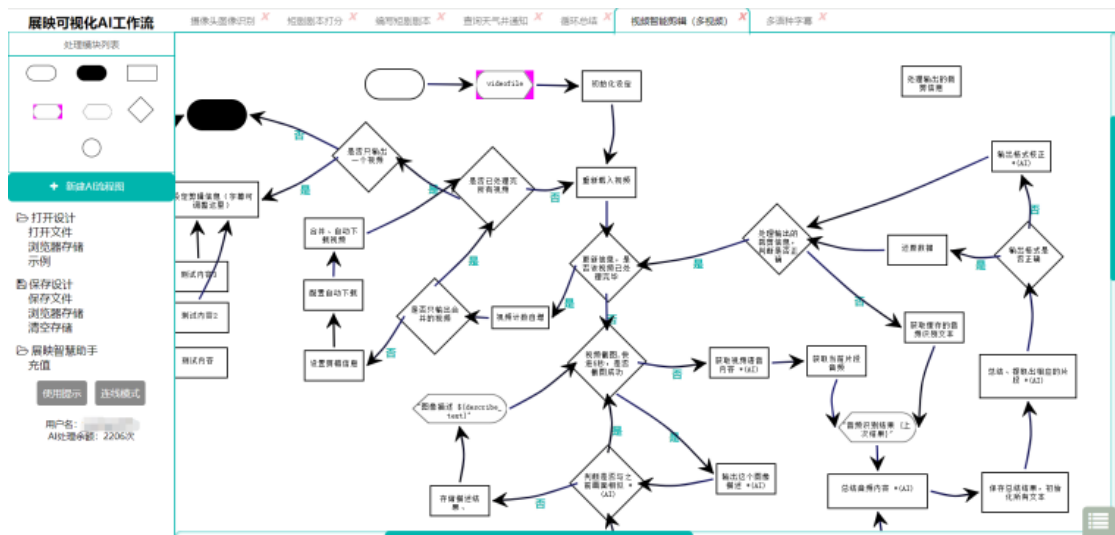
即可查看到相关信息



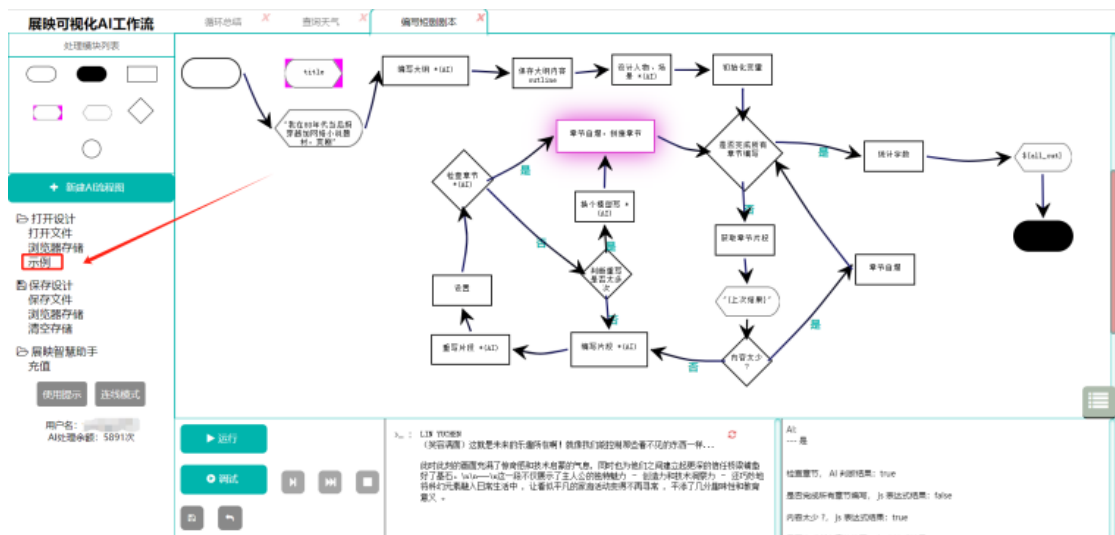
同时也可体验最新版的 AI workflow
让 AI 一步步按照你的流程自动执行，帮助你处理重复中等难度的工作。

现在开始自由体验吧！

推荐使用我们的展映 **AI 在线服务**（可视化 AI 流程工作室）：



展映的 AI workflows 视频自动剪辑
再也不担心拍摄、存放太多视频素材了，可以放心拍~



AI 自动创作

欢迎参阅：

[展映可视化 AI workflows 设计工具 使用说明](#)

欢迎联系我们：

深圳展映科技

<http://zyinfo.pro>

Email: tel_pan@126.com

微信: youkpan