1. Orion-14B-Chat-RAG
https://huggingface.co/OrionStarAI/Orion-14B-Chat-RAG/blob/main/README_zh.md
https://github.com/OrionStarAI/Orion/blob/master/gradio_demo/doc_qa_task/doc_qa.py
猎户星空 rag 版 14B，支持 320k token
难度：1

2.RAGFLOW
https://github.com/infiniflow/ragflow?tab=readme-ov-file
提供一个封装好的 RAG，支持本地模型，自吹无限上下文海底捞针，可以封装成 api
难度：1

3.kotaemon
https://github.com/deepset-ai/haystack
一个简单的 RAG 开源框架，需要自主搭建 pipeline
支持本地模型,相比于下面的 kotaemon 更灵活，稍微有点代码难度，疑似可以零代码打造知识库
可支持文本大小未确定
难度：2.5

4.haystack
https://github.com/Cinnamon/kotaemon
类似于 haystack,是一个 rag 框架，额外提供 UI 界面，完全零代码，可以自定义 pipeline，支持本地大模型
是否提供输出接口待确认
难度：1.5

5.LightRag
https://github.com/HKUDS/LightRAG
和 ragflow 类似，据说不稳定，设计为不支持本地模型，本地模型需要修改源代码！只支持 txt
难度：4

6.txtai
https://github.com/neuml/txtai
和 kotaemon 类似，需要代码搭建框架，非常灵活，疑似输出支持语音合成
难度：3

7.LLM APP
https://github.com/pathwaycom/llm-app
和 kotaemon 类似，优点是支持动态数据库，git 上有详细教程，有 UI,
难度：2.5

8.cognita
类似 hystack,全 UI 框架
难度：2

以上项目，多数 b 站和 youtube 上搜名字有别人做好的演示流程
总结：下面是各类方法搭建完后搭建和编写难度
- orion,ragflow 即下即用，奶奶都会
- cognita，haystack 拉个文科生都会用
- kotaemon, LLM APP 难度适中，值得一试
- txtai 没仔细看， lightrag 狗都不用