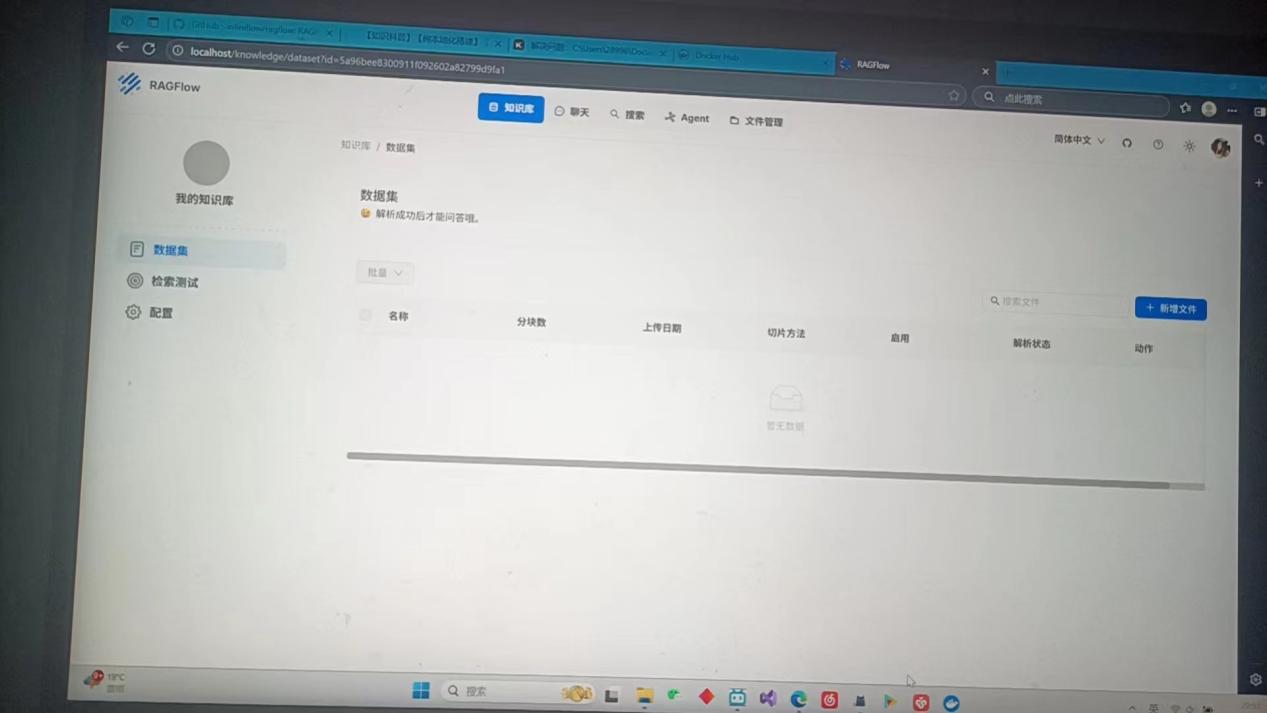
除了微调大模型，也可以加上RAG技术

【知识科普】【纯本地化搭建】【不本地也行】DeepSeek + RAGFlow构建个人知识库 <https://www.bilibili.com/video/BV1WiP2ezE5a/?share_source=copy_web&vd_source=4f37e0a31081d092a2e265bd80ab2da7>

知识库成品如下：



# 四问四答

一、为什么不直接使用网页版DeepSeek?

·需求：绝对的隐私保护和个性化知识库构建

·场景：希望大模型能根据规章制度来回答问题，一般需要上传相关附件；但仍然可能面临的问题有：

1.数据隐私问题：联网使用大模型数据隐私性无法得到绝对保证；

2.上传文件的限制问题：网页版AI对于文件上传的数量、大小一般有限制并且通常需要付费；

3.仅通过附件扩展上下文功能有限；每次在新对话中提问相关问题时，仍需要重新上传附件；修改删除对话中已有的附件困难。

二、如何实现网页版DeepSeek不能实现的需求？

·隐私保护

通过对话大模型的本地部署解决隐私问题

·个性化知识库构建，

使用RAG技术（Retrieval-Augmented Generation，检索增强生成）构建个人知识库。为此我们需要:

1.本地部署RAG技术所需要的开源框架RAGFlow；

2.本地部署Embedding大模型（或者直接部署自带Embedding模型的RAGFlow版本）。

三、为什么要使用RAG技术？RAG和模型微调的区别

·大模型的幻觉问题

·微调技术和RAG技术：

微调：在已有的预训练模型基础上，再结合特定任务的数据集进一步对其进行训练，使得模型在这一领域中表现更好；

RAG：在生成回答之前，通过信息检索从外部知识库中查找与问题相关的知识，增强生成过程中的信息来源，从而提升生成的质量和准确性。

共同点：都是为了赋予模型某个领域的特定知识，解决大模型的幻觉问题。RAG的原理

·检索（Retrieval）：当用户提出问题时，系统会从外部的知识库中检索出与用户输入相关的内容；

·增强（Augmentation）：系统将检索到的信息与用户的输入结合，扩展模型的上下文，再传给生成模型；

·生成（Generation）：生成模型基于增强后的输入生成最终的回答。由于这一回答参考了外部知识库中的内容，因此更加准确可读。

四、什么是Embedding？为什么除了DeepSeek、RAGFlow外还需要“Embedding模型”？

·检索（Retrieval）的详细过程:

准备外部知识库：外部知识库可能来自本地的文件、搜索引擎结果、API等等。

通过Embedding（嵌入）模型，对知识库文件进行解析。Embedding 的主要作用是将自然语言转化为机器可以理解的高维向量，并且通过这一过程捕获到文本背后的语义信息（比如不同文本之间的相似度关系）。

通过 Embedding模型，对用户的提问进行处理；用户的输入同样会经过嵌入处理，生成一个高维向量。

拿用户的提问去匹配本地知识库：使用这个用户输入生成的这个高维向量，去查询知识库中相关的文档片段。在这个过程中，系统会利用某些相似度度量（如余弦相似度）去判断相似度。

简而言之，Embedding模型是用来对上传的附件进行解析的。

·模型的分类：Chat模型、Embedding模型等