token:

每本书前10页，总结 喂给LLM 写 tags

存到字符文件中

数据写入：

// 视频：现成api (transcript)

pdf：ocr api (mathpix)

单个文件对应一个db

用户问问题：

LLM 写出跟问题相关的多个tags（联想）

找所有有共同tag的文件

调用chromadb

现成prompt让他找到最契合的"文件夹"

template:

从下面几个词条中找到和问题最相似的一项：

{该层的字符文件}

问题：{输入的问题}

不断重复直到找到所有相关的文件

调用这个文件对应的数据库，similarity search

结果返回LLM，总结

问过的问题存到memory中

每个文件的总结再整体总结

token:

// 如果没找到，就用最相似的信息上网搜