

知识库检索系统设计说明

本文档描述一个本地知识库聊天系统的检索管线设计。系统采用混合检索策略，将向量近邻搜索与全文关键词搜索通过倒数排名融合算法合并，再经过交叉编码器重排序，最终把小块命中扩展为父段落交给大语言模型阅读。

一、切分策略

文档导入时先解析为统一的中间表示，再按文档形态路由到五种切分策略之一。整篇不足一千字符的小文档整体作为单块入库；无标题的纯文本走滑动窗口；标题密集的笔记以小节为块；长章节文档采用父子两级切分。

1.1 父子切分

父子切分的核心思想是"检索用小块，阅读给大块"。子块控制在一千字符以内用于嵌入和匹配，父块上限四千字符，在回答阶段替换命中的子块，让模型读到完整语境。子块永不跨越父块边界，这保证了替换操作的正确性。

1.2 块感知切分

切分器按内容块打包，只有单个块超出预算才会切入块内部。代码块按行断开，表格按行组切分且每个分片重复表头，确保任何分片都不丢失列语义。

二、检索参数

参数	默认值	说明
子块大小	1000 字符	嵌入与匹配的基本单位
父块上限	4000 字符	回答阶段的阅读单位
召回数量	5	融合后送入重排的候选数
重排候选	20	超量取回后重排到召回数量

三、示例代码

```
export function chooseStrategy(profile) {
  if (profile.chars === 0) return "empty";
  if (profile.chars <= CHILD_SIZE) return "whole-doc";
  if (profile.headingless) return "sliding-window";
  return "parent-child";
}
```

评测采用检索专用通道，只计算上下文精确率与召回率两项指标，避免把生成模型的弱点混入信号。第一轮评测显示父块扩展让两项指标同时提升，据此将其设为默认开启。