RAG的流程大致分为三个步骤：

1. 对于文档的处理。

首先，从pdf（或其他类型的文本比如word）中提取内容，提取出前三页合并并存储为文档摘要；

然后，先存储了整页的内容，再将每一页的文本分块，并生成每一个分块的嵌入向量，记录每个块的所属文档ID、所属知识库ID、页码、内容、向量等，并存储到ES中，分块时要注意确保块之间有重叠，让逻辑更完整。

1. 进行检索的操作。

首先，进行文档检索，检索两次，包括全文检索和语义检索，全文检索实现的是检索指定知识库中与query相关的块，语义检索实现的是根据向量检索到含义相似的字段，两种检索都返回最相关或最相似前50个分块；

然后，将检索得到的结果进行RRF融合，对计算得到的融合分数进行排序，选取了前50个分块作为候选分块；

最后，重排序，对候选分块再次打分排序，进一步提高排序的准确性。

1. RAG聊天。

如果是第一轮的对话，将检索得到的分块内容合并，填入BASIC\_QA\_TEMPLATE，这是一个发送给LLM的模版，然后调用LLM，要求根据检索得到的资料回答问题，如果要进行多轮对话，直接调用LLM基于上下文回答问题。