M3 · 第 9 讲

知识库问答与工作流

答错时，知道该查哪里

阅读约 35 分钟

练习约 30 分钟

产出：一个带引用的问答库

本节目录

开场：明明文档里写了，它却说没有

本课目标

一、RAG 是怎么工作的

二、文档准备四步

三、答错了，按这个顺序查

四、工作流：把问答接进业务

五、权限与日志

六、30 分钟练习

七、五个常见坑

八、检查清单

九、延伸阅读

开场：明明文档里写了，它却说没有

你挂上了十份产品文档，问“退款要几天”。它回答“通常 7–15 个工作日”，可文档里写的是“3 个工作日”。你以为是模型不行，其实多半是**检索没找到那一句**。

知识库问答的成败，八成在文档准备与检索，两成才在回答。

本课目标

- 说清 RAG 的三步：切分、检索、生成；
- 用四步把文档整理成可检索的知识库；
- 会用排查表定位“答错”的原因；
- 设计一条带人工确认的工作流。

一、RAG 是怎么工作的

1. **切分**：把文档切成小段；
2. **检索**：按问题找出最相关的几段；
3. **生成**：把这几段塞进上下文，让模型基于它们回答。

所以答错只有三种可能：没切对、没检到、答歪了。先查前两个，不要一上来改提示词。

二、文档准备四步

1. 清洗

去掉页眉页脚、重复目录、过期版本。同一主题只保留最新版，旧版归档。

2. 切分

按语义切，不按固定字数硬切：

- 一段 300–800 字；
- 一个小标题下的内容尽量在一段；
- 表格单独成段，保留表头。

3. 加元数据

每段带上：文档名、章节、更新时间、适用范围。检索时能按版本过滤，回答时能给出处。

4. 建索引

把“问题—答案”常见说法也放进去，例如“退款多久到账”与“退款时效”是同义词。

三、答错了，按这个顺序查

现象	先查	处理
说“没有相关信息”	检索是否命中	补同义词、换切分
答的是旧版本	元数据与过滤	归档旧文档、加时间过滤
答对一半	切分是否切断	合并相关段落
引用不存在	提示词是否要求引用	强制“只依据材料，给出来源”
答得很泛	问题是否太宽	先反问澄清，再检索

一个硬规则：**回答必须带出处；没有出处就承认不知道。**

四、工作流：把问答接进业务

一条可用的退款咨询流程：

1. 用户提问；
2. 检索知识库，取前 3 段；
3. 置信度低 → 转人工；
4. 置信度高 → 生成回答 + 引用；
5. 敏感操作（退款、改地址）→ 人工确认后再执行；
6. 记录日志：问题、命中段落、回答、是否转人工。

三个设计原则：

- **能查就不猜**：没有材料就转人工；
- **能确认就不自动**：涉及钱与权限，必须有人；
- **能回退就不锁死**：失败可重试，且不重复执行。

> 知识库平台与模型调用的额度、计费方式差异很大（2026-11-20 查询）；选型时点名具体平台核对免费额度类型，多轮调试通常会先耗尽试用额度。

五、权限与日志

项	最低要求
文档权限	按部门/角色隔离，不能全员可见
数据脱敏	手机号、身份证、合同金额
日志	谁问的、检索到什么、答了什么
保留期	按合规要求，默认 90 天

没有日志的问答系统，出了问题无法复盘。

六、30 分钟练习

1. 选 5–10 份同主题文档（5 分钟）；
2. 清洗并切成 300–800 字的段（10 分钟）；
3. 每段加“文档名 + 更新时间”（5 分钟）；
4. 问 5 个真实问题，记录命中与答错原因（10 分钟）。

七、五个常见坑

1. **把整份 PDF 直接上传**：切分不可控；
2. **旧版文档没删**：模型会引用过期信息；
3. **不要求引用**：无法判断它是不是在编；
4. **所有问题都自动答**：高风险场景必须转人工；
5. **没有评估集**：永远不知道改版是变好还是变坏。

八、检查清单

- [ ] 文档已清洗，旧版已归档；
- [ ] 切分按语义，单段 300–800 字；
- [ ] 每段有来源与更新时间；
- [ ] 回答强制引用；
- [ ] 低置信度转人工；
- [ ] 有日志与评估问题集。

九、延伸阅读

- 知识库 1.3 AI Agents（智能体）
- 知识库 2.3 藏经阁：个人专栏
- 知识库 4.4 商业化：行业解决方案

> 下一讲：AI 编程入门。不懂代码，也能做出一个能运行的小工具。