运维智能客服需求

一. 需求

（一）客服场景智能问答

1. 客服场景：基于知识库提供精准问答，不过度联想，禁止杜撰答案。
2. 支持多轮对话，用户问题不明确时，引导用户补充描述以明确问题。
3. 回答内容支持文本、已编撰好的 PDF/Word 等格式的文件

（二）转人工

1. 对于知识库不存在的问题，支持无缝转人工客服。
2. 用户如对智能问答结果不满意，或有明确转人工诉求时，支持无缝转人工客服。
3. 转人工时，传递历史对话，避免用户重复描述问题。

（三）知识管理

知识来源支持多种类型：企业微信用户对话、在线文档/表格、PDF、Word、Excel 等。

（四）效果评估与优化

技术指标：运维人员定期进行随机抽样评估，准确率不低于 90%，召回率不低于 70%。

二. 方案设计

基于 Agent + RAG-Anything 构建运维智能客服：以 bge-m3 做嵌入、Qwen3 负责文本生成、Qwen2.5-VL 解析图表，配套 PDF/Docx 导出与人工接管工具；RAG 仅召回，Agent 仅作答，全流程可一键建库、多轮问答、转人工，并以 ≥90 % 准确率、≥70 % 召回率持续评估。

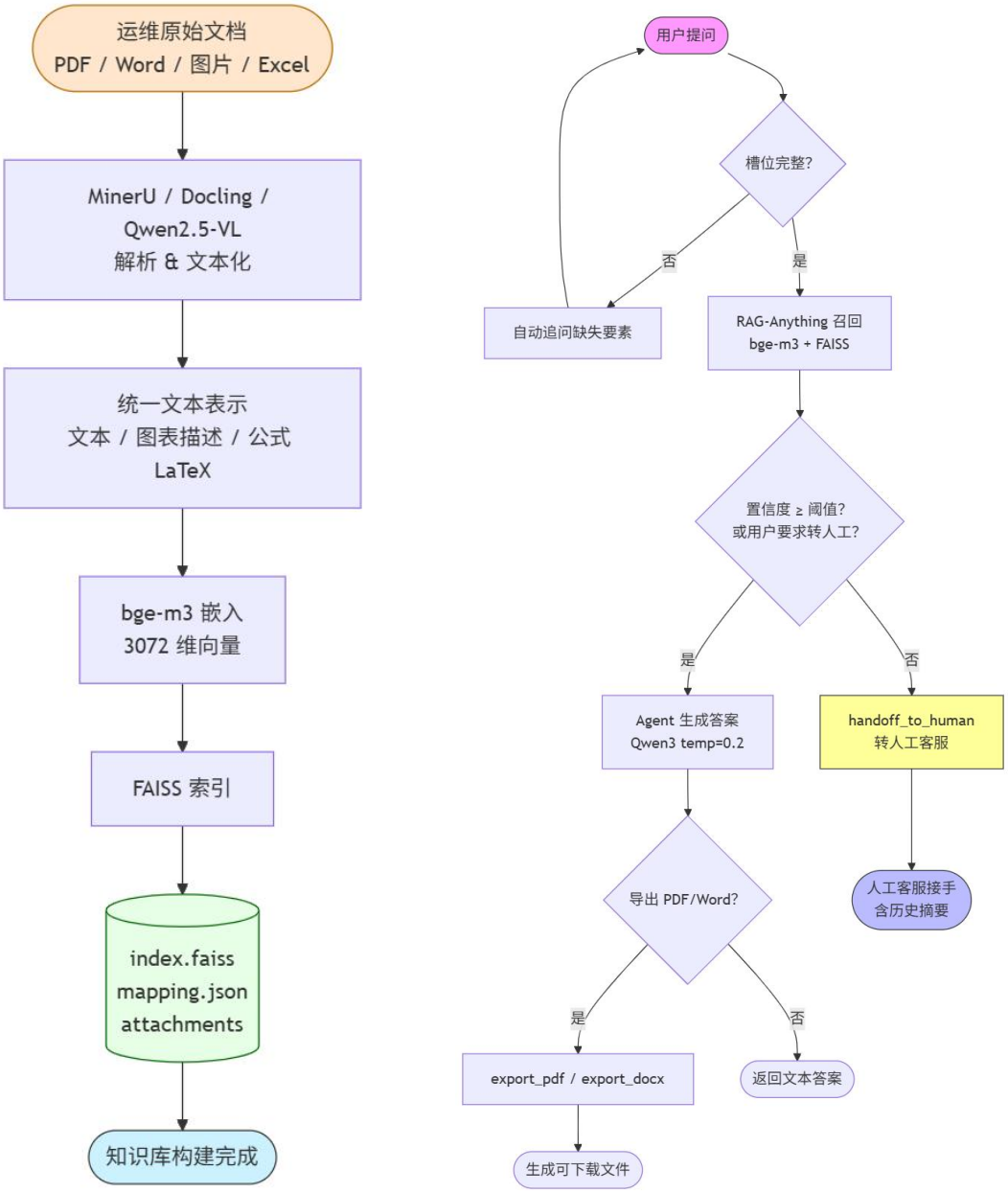

图 整体架构（左：构建知识库，右：使用知识库）

为了保证流程的正确实施，采用 LangGraph 把运维客服抽象成「追问-召回-回答-转人工」四步状态机，缺信息先追问，召回后低置信即转人工，全程记忆上下文。

（一）客服场景智能问答

1. 客服场景：基于知识库提供精准问答，不过度联想，禁止杜撰答案。

只基于知识库，不杜撰，无命中则转人工客服；合理设置 Agent LLM 的 temperature 参数（0.1-0.3），使得模型生成更偏于保守。

2. 支持多轮对话，用户问题不明确时，引导用户补充描述以明确问题。

- 多轮对话

设置合理的滑动窗口大小，保存用户上下文，在新一轮提问时附带上下文信息，达到多轮记忆。

- 自动追问

设计必填槽位，用户输入映射到预定义的必填要素（slot）并实时检查哪些要素为空。实现方法是必填要素写成 JSON schema，让 LLM 模型一次性返回缺字段。具体迭代过程如下：

  - 多轮记忆 → 槽位持续累积。
  - 自动追问 → 缺什么问什么。
  - 不杜撰 → 必填槽位齐后才查知识库。

3. 回答内容支持文本、已编撰好的 PDF/Word 等格式的文件

- 纯文本（默认）

模型返回的 result.text 直接就是带引用的自然语言答案。

- 已编排好的 PDF / Word

设计并封装函数，根据 result 结果构建 export_pdf()和 export_docx()，实现输出为 PDF/Word 等格式的文件，函数要在 result 答案中插入：

  - 引用页码/图表编号
  - 相关图片、表格按阅读顺序排版

生成的文件直接可下载，无需二次编辑。这些都写为函数，并加入 AI Agent 的工具。

（二）转人工

1. 对于知识库不存在的问题，支持无缝转人工客服。

知识库查询没有高于阈值的向量，那么转人工客服程序。人工客服程序是一个人工客服 Function 接口，输入包含智能客服与人工客服的交接文本，输出可以在终端打印相关话语（“人工客服已经接手！”）。该函数也当作一个工具，加入 AI Agent 的工具列表。

2. 用户如对智能问答结果不满意，或有明确转人工诉求时，支持无缝转人工客服。

使用 LLM 分析，用户的问题中出现了“对智能问答结果不满意，或有明确转人工诉求时”的意图时，调用人工客服程序。

3.转人工时，传递历史对话，避免用户重复描述问题。

使用 LLM 将与用户交互历史总结为一段清晰明确的话，传递给人工客服程序，方便人工客服理解前一步智能客服解决用户问题的状态，并进行下一步的工作。

（三）知识管理

使用 RAGAnything 构建知识库，把文本/图像/表格/公式都对齐为文本表示，并把本身记录路径之后作为附件进行保存,再对统一的文本表示做语义分块和嵌入，以此来构建知识库，知识库由一份索引文件（存储向量）、一份映射表（每一个向量 ID 与原始出处）和一组可溯源的附件目录共同构成。

并且支持对于多个文档（可以是不同类型，PDF/EXCEL/WORD/文本），最终都保证落在同一个 faiss.index、同一个知识图谱里，后续问答无需任何改动即可跨文档问答。

具体步骤如下：

- 文档解析阶段（多模态 → 纯文本）

  PDF/图片/表格 → MinerU/Docling 解析

  图像 → 生成 文字描述（由 VLM 生成说明文字，原始文件保存到附件，路径保存到映射表）

  表格 → 转成 Markdown 文本

  公式 → 转成 LaTeX 字符串

- 索引与检索阶段（只剩文本向量）

所有模态都被文本化，再统一喂给 bge-m3 得到向量。

文本块 + 图像描述 + 表格文本 → 3072 维向量。

存入 FAISS → 支持文本/表格/公式的相似度检索。

原始图像文件仅作为附件随结果返回，不参与向量计算，因此后续检索时可一次性完成跨模态相似度计算。

- 查询阶段（只能“按文字搜图”）

用户提问 → bge-m3 编码 → 在文本向量里找相似 →

返回：命中文字块,并附带对应图片路径/表格。

（四）效果评估与优化

技术指标：运维人员定期进行随机抽样评估，准确率不低于90%，召回率不低于70%。

- 准确率 = 正确回答数 / 回答总数

准确率验证 Agent 给出的答案是否在知识库里有直接依据且不杜撰。

- 召回率 = 被检索到的相关块数 / 知识库中全部相关块数

召回率检验知识库里该答的内容是否被成功检出。