主要模块	模块内容
需求分析	客户需求分析
	已有数据盘点
	技术选型
	初步技术方案
数据预处理	解析方案评估与选型
	数据清洗
	复杂文档解析
	数据加工
	多源数据接入
	数据解析效果评估
	数据清洗标准化（可选）
数据向量化	切片策略选择及优化
	嵌入模型选型及优化
	向量数据库选型及结构设计
	数据权限设计
检索与召回	检索结构设计
	排名融合
	Rerank模型选型及优化
生成	Prompt优化
	生成模型选型及调优
	微调数据抽取
	生成模型微调
	微调效果测试
	多轮对话
	原文索引
RAG高级优化	意图识别
	Query优化
	多模态数据输入
	检索策略优化
	召回率优化
	生成质量把控
	Agent能力接入
	推理效率优化
总计	
人天费用总计（万）：	

	平台软件原型设计
	UI设计
平台软件开发	前端界面开发
	后端数据库设计
	系统日志及审查
	API接口开发与接口文档
	前后端对接联调
系统集成	多源数据接入
	现有系统对接
	缓存与性能优化
	模型适配
	工具适配
测试与部署	分布式部署
	性能测试
	环境部署

具体工作	复杂度细分-人天	
	基础	中等
	2	3
	4	7
	2	2
RAG整体框架设计	2	5
	2	4
数据去噪、实体消歧、数据去重	10	20
PDF解析、图片解析、公式解析		10
数据摘要、数据结构化	3	8
		7
		5
	2	3
	2	3
嵌入模型微调、训练数据集构建	4	9
Milvus、Qdrant等	2	3
		3
混合检索：HNSW检索与BM25	2	3
	2	3
Rerank模型训练、微调	4	9
	1	2
Temperature调优、置信度调优	2	3
		10
Lora、Q-Lora微调		50
		10
	3	3
关联文档	3	3
LLM意图识别、专有模型训练		2
问题改写、同义词扩展、HyDE		5
		10
多路交叉召回		5
Top-K/相似度阈值调优、去重机制设计		7
后验修正、逻辑校验		
		5
	52	222
5000/人天	26	111

复杂	算力需求
5	
14	
3	
10	
7	
40	
20	>=20G
15	>=16G
15	
10	
8	
5	
16	>=16G
5	
5	
5	
5	
16	>=16G
3	
5	
20	
80	>=40G
20	
3	
3	
10	>=16G
20	>=16G
30	
10	
14	
8	
待定	
10	
440	
220	

备注
复杂场景按需评估，最少40人天
1. 解析PDF，主要使用MinerU，配置要求：CPU（最少4核心）、内存（最少32G及以上）、硬盘（最少20G SSD）、显卡（GPU显存>6G、已在Ascend 910B 上测试） 2. 图片识别，暂定选用olmOCR，配置要求：GPU（最少20G ）、硬盘（最少 30G SSD）
暂定模型 qwen2.5-7B-Instruct
暂定模型：BAAI/BGE-m3，未公开配置信息
暂定模型：BAAI/bge-reranker-v2-m3，未公开配置信息
1. Qwen-32B模型：LoRA（32-36G），QloRA（16-18G） 2. Qwen-72B模型：LoRA（72-80G），QloRA（36-40G）
暂定模型 qwen2.5-7B-Chat
暂定模型 qwen2.5-7B-Chat

场景描述	
数据复杂度	数据规模
	数据特征
需求复杂度	功能需求
	个性化定制
验收标准精度	问答准确率

基础
一家中小型企业希望构建一个基于FAQ的知识问答机器人，主要服务于内部员工的常见问题解答。企业已有的FAQ数据存储在Excel或简单数据库中，数据来源单一且结构较为固定，语言统一为中文。
数量较少（<1万条） 纯Word文档 / 纯文字PDF文档。结构化数据，数据标准无需清洗
基于RAG构建一个基本问答系统，实现检索与生成模块的最简对接。 采用开源模型及默认配置，无多轮对话支持，用户体验上主要满足单一轮问答。接口对接简单，无复杂权限管理。
精度要求低(70%左右即可)

中等
一家中型企业或机构需要为客服部门构建一个覆盖内部知识与产品手册的问答系统。此系统不仅需要整合多种格式的数据（PDF、Word、数据库内容），而且需要具备较强的语义匹配能力，提供比传统FAQ更智能的检索结果。
1万 – 10万 多模态文档（含图表的PDF）。存在一定非结构化信息，需进行格式转换、数据清洗和实体标准化
搭建中等级别的RAG智能问答系统，实现检索生成联动升级。 系统支持初步的多轮对话和上下文保持，同时提供分页查询和基础API接口，涉及简单权限管理，满足常规企业应用需求。
有一定精度要求(80%左右)

复杂
面向法律或医疗领域的垂直知识问答平台，需整合来自各大部门或机构的海量数据，包括结构化数据、半结构化报告、非结构化病历或法规文档。系统支持多语言、多轮对话和上下文追踪，确保答案的精准性和实时性，是一个定制化、深度优化的应用。
>10万条
数据存在多种格式和语言，不同领域数据差异大，需进行深度清洗、格式转换、知识图谱构建与实体对齐
构建定制化、高精度的RAG智能问答平台，面向行业垂直应用。 系统支持多轮连续对话、上下文记忆、动态知识更新及知识图谱融合，同时具备完善的API开放、权限管理、日志监控和安全防护机制。
精度要求极高(>90%)

RAG成本人		
对应 评估维度	**主要模块**	**模块内容**
场景复杂度	需求分析	客户需求分析
		已有数据盘点
		技术选型
		初步技术方案
数据复杂度	数据预处理 （人天待重评估）	解析方案评估与选型
		数据清洗
		复杂文档解析
		数据加工
		多源数据接入（可选）
		数据解析效果评估
		数据清洗标准化（可选）
	数据向量化	切片策略选择及优化
		嵌入模型选型及优化
		向量数据库选型及结构设计
		数据权限设计（可选）
精度要求	检索与召回	检索结构设计
		排名融合
		Rerank模型选型及优化
	生成	Prompt优化
		生成模型选型及调优
		微调数据抽取（可选）
		生成模型微调（可选）
		微调效果测试（可选）
		多轮对话（可选）

		原文索引（可选）
	RAG高级优化	意图识别
		Query优化
		多模态数据输入
		检索策略优化
		召回率优化
		生成质量把控（可选）
		Agent能力接入（可选）
		推理效率优化（可选）
总计		
人天费用总计（万）：5000/		

平台软件开发	平台软件原型设计
	UI设计
	前端界面开发
	后端数据库设计
	系统日志及审查
	API接口开发与接口文档
	前后端对接联调
系统集成	多源数据接入
	现有系统对接
	缓存与性能优化
	模型适配
	工具适配
测试与部署	分布式部署
	性能测试
	环境部署

人天标准参考

具体工作	复杂度细分-人天			整体评估
	基础	中等	复杂	
	2	3	5	
	4	7	14	
	2	2	3	
RAG整体框架设计	2	5	10	
	2	4	7	
数据去噪、实体消歧、数据去重	10	20	40	
PDF解析、图片解析、公式解析		10	20	
数据摘要、数据结构化	3	8	15	
		7	15	
		5	10	
	2	3	8	
	2	3	5	
嵌入模型微调、训练数据集构建	4	9	16	
Milvus、Qdrant等	2	3	5	
		3	5	
混合检索：HNSW检索与BM25	2	3	5	
	2	3	5	
Rerank模型训练、微调	4	9	16	
	1	2	3	
Temperature调优、置信度调优	2	3	5	
		10	20	
Lora、Q-Lora微调		50	80	
		10	20	
	3	3	3	

关联文档	3	3	3	
LLM意图识别、专有模型训练		2	10	
问题改写、同义词扩展、HyDE		5	20	
		10	30	
多路交叉召回		5	10	
Top-K/相似度阈值调优、去重机制设计		7	14	
后验修正、逻辑校验			8	
			待定	
		5	10	
	52	222	440	
人天	26	111	220	

需求场景	规程库
需求概述	包含包含调规和运规，以文本文档形式存储，一共13本，要求做交互式应答
数据复杂度	中等
场景复杂度	基础
精度要求	复杂
具体人天	2
	4
	2
	2
	4
	20
	10
	8
	5
	3
	9
	3
	2
	2
	4
	1
	2
	3

3
2
5
10
5
7
118
59

缺陷库	事件库
字段明确，包含时间、站点、隐患内容（可能包含图片/视频存储）、处置过程。资料存储在专属系统中，可以表格格式导出。缺陷记录共几千条，存在缺陷情况相似，描述不同的情况，人工没有对缺陷进行正确归类。要求做交互式应答，例如描述缺陷、同类缺陷在哪些站出现过、过往缺陷的处理记录、维修人员描述缺陷，模型给出处理方案、罗列历史同类缺陷等，最好具有统计分析功能，数据可以图表的形式展现	包含时间、站点、事件内容，客户可以提供2024年的全部事件记录，要求做交互式应答
基础	中等
中等	中等
中等	中等
2	2
4	4
2	2
2	2
2	4
	20
3	8
	7
	5
3	3
3	3
9	9
3	3
3	3
3	3
3	3
9	9
2	2
3	3
3	3

3	3
2	2
5	5
10	10
5	5
7	7
5	5
96	
48	

报告库	安全库	算力需求
包含技术报告、事故案例、学习报告等，要求做交互式应答	安全制度每年更新，内容雷同但不相似。要求模型学习过往制度以及作业指导书内容后，能根据最新的安全制度生成相应的作业指导书，作业指导书能以WORD文档形式呈现	
		>=20G
		>=16G
		>=16G
		>=16G
		>=40G

		>=16G
		>=16G

备注
复杂场景按需评估，最少40人天
1. 解析PDF，主要使用MinerU，配置要求：CPU（最少4核心）、内存（最少32G及以上）、硬盘（最少20G SSD）、显卡（GPU显存>6G、已在Ascend 910B 上测试） 2. 图片识别，暂定选用olmOCR，配置要求：GPU（最少20G ）、硬盘（最少 30G SSD）
暂定模型 qwen2.5-7B-Instruct
暂定模型：BAAI/BGE-m3，未公开配置信息
暂定模型：BAAI/bge-reranker-v2-m3，未公开配置信息
1. Qwen-32B模型：LoRA（32-36G），QloRA（16-18G） 2. Qwen-72B模型：LoRA（72-80G），QloRA（36-40G）

暂定模型 qwen2.5-7B-Chat
暂定模型 qwen2.5-7B-Chat