# 研发工作量评估

## 文档信息

| 文档版本 | 修订日期 | 修订人 | 修订描述 |
| --- | --- | --- | --- |
| v0.2 | 2025-09-04 | 项目组 |  |

## 1. 项目概述

### 1.1 项目背景

飞机发动机QEC拆装智能化系统（二期）在一期基础上，引入AI与RAG能力，建设四大模块：维修知识库管理、维修规划及资源智能化、生产看板、发动机检测辅助。总体采用分层架构（展示/应用/平台/数据/基础设施），技术栈以 Python/Flask、Next.js、PostgreSQL、向量数据库、Redis、MQ、对象存储为主，Docker/Helm 部署。约束：项目周期≤12个月，需与SAP、文档管理系统等对接，符合企业安全规范。

### 1.2 项目范围

* 在范围：
  + 维修知识库管理（知识接入/解析/分段/索引/检索/问答、元数据、权限与审计、质量评测、外部KB接入、限流等）
  + 智能问答与检索定位（RAG、混合检索、重排、引用可追溯）
  + 生产看板（计划/进度/资源/工位看板、实时刷新）
  + 发动机检测辅助（任务/执行/结果/预警、报告生成）
  + 基础设施与运维（监控、日志、备份、部署）
* 不在范围：SAP内部实现、物流运输流程、硬件采购选型等。

## 2. 工作量评估

说明：以下估算基于设计文档明确的范围与复杂度，按“工作包”拆分，单位为人日（理想工时），未含团队同步与管理损耗（建议新增15%管理/沟通/缓冲系数）。标注“待澄清”的项已在第5章给出假设与缺口。

### 2.1 功能模块划分（概览）

| 模块名称 | 功能描述 | 复杂度 |
| --- | --- | --- |
| 维修知识库管理 | 知识接入/解析/分段/索引、文档与分段维护、权限审计、质量治理 | 高 |
| 智能问答（RAG） | 多轮问答、检索增强、引用追溯、会话/反馈 | 中高 |
| 检索与定位 | 混合检索、父子分段上下文、重排与手册定位 | 高 |
| 元数据管理 | 字段治理、批量标注、筛选过滤 | 中 |
| 外部知识库接入 | 标准化结果合并、降级与熔断、监控告警 | 中 |
| 维修规划及资源智能化 | 计划生成/调整、资源优化、冲突解决（详细待补充） | 高（待澄清） |
| 生产看板 | 计划/进度/资源/工位可视化与实时刷新（详细待补充） | 中（待澄清） |
| 发动机检测辅助 | 检测任务/执行/分析/预警（详细待补充） | 高（待澄清） |
| 平台与运维 | 部署、监控、日志、备份、限流与安全 | 中 |

### 2.2 工作包与估算

| 工作包 | 主要交付物 | 依赖 | 估算(人日) | 优先级 |
| --- | --- | --- | --- | --- |
| 数据集/文档/分段CRUD与状态机、API | Dataset/Document/Segment 全流程、状态机、分页检索 | 基础框架 | 45 | P0 |
| 解析与索引流水线（Worker+队列+VDB） | 解析/分段/Embedding/索引写入、失败重试、进度查询 | 对象存储、VDB | 60 | P0 |
| 混合检索与重排（父子分段合并） | 向量+关键词检索、MMR/启发式/可选Rerank | VDB/FTS | 40 | P0 |
| 元数据字段与批量标注 | 字段CRUD、批量赋值、校验、索引 | RDB | 20 | P1 |
| 检索质量评测与策略版本化 | 测试台、A/B、门禁与回滚 | 检索服务 | 25 | P1 |
| 外部知识库连接器 | 配置/鉴权、标准化合并、超时降级与监控 | 网络、密管 | 25 | P1 |
| 权限与审计、限流 | RBAC/资源级权限、审计日志、令牌桶 | 统一网关 | 20 | P1 |
| 智能问答RAG编排 | 会话/消息/反馈、引用可追溯、流式应答 | 检索服务/LLM | 45 | P0 |
| 维修规划与资源智能化（里程碑版） | 需求梳理Spike(10) + 核心流程实现(50) | SAP/KB | 60 | P2 |
| 生产看板（里程碑版） | 需求梳理Spike(8) + 看板页面/聚合(32) | 规划/实时 | 40 | P2 |
| 发动机检测辅助（里程碑版） | 需求梳理Spike(12) + 检测流程/报告(48) | 设备/KB | 60 | P2 |
| 前端Web（Next.js） | KB管理、分段维护、检索测试台、基础看板 | 后端API | 60 | P0 |
| 部署与运维 | Docker/Helm、监控/日志、备份、环境差异化 | 中间件 | 30 | P0 |
| 性能优化与压测 | 压测场景/脚本、热点缓存、参数调优 | 全链路 | 20 | P1 |
| 安全与合规 | 传输/存储加密、密钥管理、合规审计 | 安全基线 | 15 | P1 |

合计（不含管理缓冲）：约 565 人日；建议加15%缓冲（沟通/集成/不确定性）≈ 650 人日。

说明：规划、看板、检测模块目前详细设计占位较多，估算以“里程碑可用版”为目标，后续按增量迭代补强。

### 2.3 人力资源评估（建议团队）

| 角色 | 人数 | 主要工作 | 参与阶段 |
| --- | --- | --- | --- |
| 项目经理/交付经理 | 1 | 计划/进度/风险/干系人沟通 | 全程 |
| 架构师/Tech Lead | 1 | 架构/技术把关/关键难题攻关 | 设计/开发/性能 |
| 后端工程师 | 4 | API/服务/队列/DB/VDB/检索 | 开发/联调 |
| 前端工程师 | 2 | Next.js页面/交互/看板 | 开发/联调 |
| 算法/LLM工程师 | 1 | Embedding/Rerank/RAG策略 | 设计/开发 |
| 测试工程师 | 2 | 用例/自动化/性能/安全测试 | 测试/验收 |
| DevOps/SRE | 1 | CI/CD、监控、部署、容量 | 搭建/上线 |
| UI/UX（兼职） | 0.5 | 信息架构/交互设计/视觉 | 设计/开发 |
| 数据/ETL（兼职） | 0.5 | 解析/清洗/词表治理支持 | 开发 |

### 2.4 时间评估与里程碑（建议）

| 阶段/里程碑 | 工作日 | 主要交付 | 备注 |
| --- | --- | --- | --- |
| M0 需求/方案冻结 | 10 | 需求清单、接口契约、里程碑计划 | 含跨系统对齐 |
| M1 KB/RAG基础可用 | 50 | KB CRUD/解析索引/混检/问答最小闭环、基础前端、部署监控 | P0完成 |
| M2 质量与治理增强 | 30 | 元数据、质量评测/AB、审计/限流、安全、性能优化 | P1完成 |
| M3 规划与看板Beta | 40 | 规划核心流程、基础看板联动 | P2部分完成 |
| M4 检测辅助Beta与收尾 | 40 | 检测流程与报告、稳定性/文档/验收 | 全量联调 |

总计约170个工作日（团队并行可缩短至约7个月内完成交付）。实际开始/结束日期由立项后排期落表。

## 3. 风险评估

| 风险描述 | 可能性 | 影响 | 应对措施 |
| --- | --- | --- | --- |
| 向量数据库性能与稳定性不足 | 中 | 高 | 选型评估与压测；分片+副本；降级关键词检索；启用缓存 |
| LLM/MaaS接口不稳定或成本超标 | 中 | 中高 | 多供应商备份；流量限额；离线/本地推理兜底；提示词与上下文预算器 |
| 文档解析失败率高/格式多样 | 中 | 中 | 多解析器并行；失败回退人工；解析沙箱隔离与白名单 |
| 与SAP/文档系统对接不确定 | 中 | 高 | 尽早对齐接口契约/沙箱账号；设计适配层与重试机制 |
| 权限与审计合规要求变更 | 低 | 中 | 采用RBAC与审计通用组件；预留策略配置 |
| 模块详细设计缺失导致返工 | 中 | 中高 | 先行Spike澄清需求；里程碑切片交付；预留15%缓冲 |

## 4. 资源需求

### 4.1 硬件/环境（建议基线）

* 开发/测试：2C/8GiB×若干；VDB按量独立
* 预生产：4C/16GiB×3（API/Worker/Web分开），独立PG/Redis/VDB
* 生产：8C/16GiB×5（API/Worker/Web多副本），PG主从、Redis集群、VDB集群；对象存储与备份空间

### 4.2 软件与中间件

* PostgreSQL 15+（含pgvector或外部VDB：Qdrant/Weaviate/Milvus择一）
* Redis、消息队列、全文检索(Elasticsearch/OpenSearch 可选)
* Docker/Helm、Prometheus/Grafana、ELK/Opensearch Logs
* 身份/权限：OIDC/SAML对接、API-Key与RBAC

### 4.3 其他资源

* 外部系统对接沙箱（SAP、文档管理）与测试账号
* 模型/向量服务密钥（含备用供应商）与密钥管理（KMS）
* 维修手册与历史数据样本、词表治理评审机制

## 5. 设计不清晰项与假设

* 维修规划与资源智能化：算法/优化策略与与SAP数据契约待明确（当前估算按“里程碑可用版”预留60人日）
* 生产看板：看板维度与实时通道（WebSocket/轮询/中间件）待确认（预留40人日）
* 发动机检测辅助：设备类型/接入协议/分析算法待明确（预留60人日）
* LLM与Rerank模型选型与配额策略未定（按多供应商可切换假设估算）
* 向量数据库选型（Qdrant/Weaviate/Milvus/Pgvector）与部署规模待压测确认

## 6. 结论与建议

* 建议以“P0最小可用闭环（KB+检索+问答+部署）→ P1质量治理与安全 → P2行业模块Beta”的三段式交付，控制范围与风险。
* 立项即锁定关键选型（VDB/LLM/全文检索），并完成M0压测与对接沙箱准备。
* 采用里程碑与A/B评测门禁机制，确保检索/问答质量达标后再扩展到规划/看板/检测模块。
* 建议全程保留≥15%缓冲与风险储备，避免因外部依赖与不确定性导致工期失控。