



个人摘要

2027 届智能科学与技术本科生，独立完成两个从 0 到 1 的 LLM 应用：**个人知识库**（RAG 检索问答）与 **Coding Agent** 执行引擎，覆盖需求拆解、方案设计、编码、测试到部署。能对 **AI 生成的代码** 做甄别、调试与补强，用调用轨迹与执行计划定位根因。

核心能力

LLM 多供应商接入

Agent 工具调用

RAG 混合检索

RRF 与重排

Prompt 约束

上下文预算

文档解析与 OCR

AI 代码审查

回归测试

Python

FastAPI

SQLite / sqllite-vec

Git

Docker Compose

断点续跑

项目经历

Whence

2026.03 - 至今

个人知识库与问答系统（RAG 检索问答）

github.com/wawapian/Whence

Python 3.11 / FastAPI / SQLAlchemy / SQLite / sqllite-vec / FTS5 / BGE-M3 / BGE-Reranker / RapidOCR / Docker Compose

- 入库与分块**：统一接入 **Word、PDF、Excel 与图片**；PDF 走文字层并按坐标抽表，仅空页触发 **OCR**；按标题分节分块（每块不超过 **600 字、重叠 60 字**），整节正文留作父级上下文并在回答时回填，避免长文档被切碎后语义断裂；表格按行入索引、表头前置，修复「统计类问得到、普通问答查不到」的召回盲区。
- 混合检索**：**向量与 FTS5 (trigram)** 双路召回，候选扩到 $\text{top}_k \times 3$ ，经 **RRF (k=60)** 融合后交叉重排；两级检索先在材料层取前 8 条、再在命中材料内取细节块；对已失效事实做双时态过滤，防止旧值污染答案。
- Agent 问答**：把前置意图图分类换成 **4 轮、6 个工具的 tool-loop**，模型只决定调哪个工具，统计类一律由代码精确计算；提示词固定「结论必须引用材料、材料不足输出 [信息不足]、实体原样引用」；时间窗与实体名做**参数守卫**（首次打回、二次强制纠正并注明口径）；本地推理不可用时直接返回 **503**，不伪装成「没有找到」。
- 归因与回归**：沿调用轨迹定位到跨月份缓存误命中，实测余弦 **0.9749** 让「上个月」命中了「这个月」的答案，阈值由 0.95 上调至 **0.98**，并用 $\cos=0.9749$ 的构造向量把该行为钉成回归测试；对 FTS5 做 **MATCH 语法中和**（逐词加引号、词内片段 OR），避免中文标点触发语法报错，异常时降级为纯向量召回。
- 性能与稳定性**：把「**KNN join 必须走索引**」用 **EXPLAIN QUERY PLAN** 钉成测试（截获真实生产 SQL，断言 SEARCH ... USING INDEX 且无全表扫描，另加反向钉子防止断言空转）；把嵌入计算移出写事务，让 SQLite 写锁从秒级压到毫秒级，修掉「大文件入库期间删除报 500」；后台队列分双车道，记录侧单工人保序、文档侧 4 路并行。仓库含 **80 个测试文件、611 个用例**。

min

2026.03 - 至今

本地 Coding Agent 执行引擎

github.com/wawapian/min

Python / 工具调用 / 上下文管理 / JSON、JSONL / pytest

- 执行闭环**：把「指令生成、参数解析、权限与路径校验、执行、结果记录、回灌模型」串成统一循环，落地 **6 个工具**（列文件、读文件、搜索、写文件、定点修改、执行命令）；抽象 **Ollama、OpenAI、Anthropic、DeepSeek** 四类客户端并统一解析 **JSON 与 SSE**，不依赖 Agent 框架自己实现 tool-loop；参数校验逐工具手写，例如定点修改要求 `old_text` 在文件中恰好出现一次，路径经 **commonpath 校验** 防止逃出工作区。
- 安全边界**：**三档审批 (ask / never / auto)** 加只读模式与任务取消，审批等待超时自动拒绝；最近两次完全相同的工具调用直接拒绝；**risky 工具** 执行前后各做一次 **全仓 sha256 快照**。
- 上下文管理**：按估算 token 给**指令、记忆、相关记忆、历史**四段分配预算（默认 **3400**），超限时按固定优先级逐段削减并记录 **reduction_log**；最近工具结果的可见长度随预算动态伸缩，旧结果按工具类型（shell、read、search、patch）摘要压缩。
- 记忆与续跑**：记忆分**工作记忆、事件笔记、文件摘要**三层，长期主题落到 `.min/memory/` 的索引与主题文件，只注入与当前文件指纹一致的摘要；断点续跑区分**五态**（无断点、完全有效、部分过期、工作区不匹配、schema 不匹配），依据文件 **sha256 新鲜度**、11 项运行时身份与工作区指纹判定能否恢复。
- 稳定性与验证**：`read_file` 返回**总行数、读取范围与截断续读位置**，避免大文件截断误判；工具步数耗尽时追加「仅允许终答」的收尾轮，防止空转；自建 **12 任务、5 类**的确定性评测框架（fixture 隔离、模型回放、verifier 子进程断言、artifact 记录），配 **128 个测试**覆盖工具校验、审批、取消与恢复路径。

教育背景

四川轻化工大学 | 智能科学与技术 (本科)

2023.09 - 2027.06

技能与证书

AI 应用开发: LLM 多供应商接入, Agent 工具调用, RAG 混合检索, RRF 融合, 重排序, Prompt 约束, 结构化输出**AI 协作与交付**: AI 生成代码审查, 参数校验与边界补强, 回归测试, 审批与只读兜底, 调用轨迹归因**工程基础**: Python, FastAPI, Pydantic, SQLAlchemy, SQLite, sqllite-vec, FTS5, Git, Docker Compose, CI **语言**: CET-4