曹灏天

1207912646@qq.com
13708254512
GitHub仓库
个人网站

教育经历

四川大学锦江学院 人工智能 · 大四在读 2023/09 - 2027/05

- 主修课程：人工智能导论 · 神经网络与深度学习 · 机器学习 ·数据挖掘 · 数据分析与应用 · Python 语言面向对象程序设计 ·云计算与虚拟化技术 · 基于平台的软件开发实践
- 求职意向：AI 应用开发（RAG / Agent）· Vibe Coding · AI 实施交付（FDE）实习生 ｜ 可提前到岗长期实习

其他经历 中国人民解放军 ｜ 义务兵役 2020/09 - 2022/09

连续两年获评"四有优秀士兵"

项目经历

我的详细信息

技术栈实现由 AI 编码代理完成，我负责描述需求、定规格、设计验收、用数据判断结果。

RAG问答助手 个人独立项目 2026/07 - 至今

https://rag.caohaotian.top/

Python · FastAPI · LangChain · ChromaDB · BGE / bge-reranker · GLM-4-Flash · SSE · Vue3

- 搭了一条完整的检索问答管线：先做多轮指代消解改写问题，再用 BM25 和向量双路召回做 RRF 融合，最后交给 CrossEncoder 精排并按分数阈值拒答。三段各自能独立开关，出问题能快速定位是哪一段的问题。
- 为了避免"改完不知道是变好还是变坏"，建了 6 套评测集，按 7:3 冻结切成训练集和盲测集，种子固定可复现。调参只用训练集，对外只报盲测数字。目前盲测 14 题，Hit@5 75.0%、MRR 0.61。
- 消融做下来有两个反直觉的结论，都写进了 README：在当前库规模下，混合检索相比纯向量没有收益（Hit@5 持平、MRR 反而略降）；查询改写的净收益也存疑。
- 上线后先补可观测性（延迟分位数、失败分类、分环节 Token 记账），数据出来才看清瓶颈在上游免费模型限流，不在检索管线本身。于是分三层治理：调用节流、限流时自动切备用模型、语义缓存（相似度阈值 0.92，只缓存带来源的回答，知识库一变更就失效）。未命中延迟从 70.7s 降到 5.7s，命中缓存约 10ms。
- 兜底和安全做了双向评测：16 题里拒答率 75%、误杀 0%。没有来源的答案会被自动拦下转人工工单，人工补答后一键回流知识库和评测集。另做过一轮对抗式安全审查，13 项指控中 11 项属实已修复，沉淀 27 条回归测试进 CI（会话越权、路径穿越、XSS、提示词注入），剩余 2 项不成立，附了验真依据。

校园失物招领智能匹配系统 个人独立项目 2026/07 - 至今

https://lost.caohaotian.top/

Python · FastAPI · SQLAlchemy · Vue3 · YOLOv8 · CLIP · Docker · pytest

- 为防止恶意冒领，我设计以下硬规则：双方各持一个 10 秒时效的验证码、比对用恒时比较防时序猜测、连错 5 次锁定、数据库行锁防并发重复确认、角色由服务端推导防越权冒充。交接全程审计留痕，出事能追溯。
- 失物的描述太杂，关键词搜索难以准确，采取七维加权匹配（品类、数量、颜色、成色、地点、关键词、时间），并对品牌和数量冲突加重罚，解决了"捡到 iPhone 却匹配到丢失华为"这类误配。自建 40 对人工标注数据做评测，F1 从 62.9% 提到 78.0%、召回从 55% 提到 80%，CI 里设了 F1 低于 76 就阻断合并。
- 性能方面，图像识别从同步执行改成异步任务架构（任务表 + 幂等键防重复消费 + 失败重试 + 死信队列 + 崩溃恢复）后，压测 QPS 从 5.4 提到 74.4、p95 从 1800ms 降到 68ms。匹配列表查询下推数据库后，单请求 SQL 从 630 条降到 5 条，平均延迟从 165ms 降到 18ms。