徐硕

北京大学 PEKING UNIVERSITY

LLM 应用后端 / 后端开发

xushuo_1471@foxmail.com

GitHub | 个人主页

教育经历

北京大学 2021.09 - 2026.07

计算机科学与技术 本科　信息科学技术学院

项目经历

多租户 RAG 知识库后端 | FastAPI, PostgreSQL/pgvector, Redis, MinIO | GitHub

- 把租户、角色和文档 ACL 条件下推到全文与向量检索 SQL，避免召回后再过滤；索引任务支持版本更新、幂等重试和进程中断恢复
- FastEmbed BGE 在公开 SciFact 数据集（5,183 篇文档、188 条标注查询）上取得 Recall@5 0.865；本地 5 万 chunk、20 并发检索客户端 p95 144ms，跨租户隔离测试通过

LLM 评测与 Prompt 回归平台 | FastAPI, SQLAlchemy, Vue 3, GitHub Actions | GitHub

- 将数据集、Prompt、模型参数和评测器保存为不可变版本，每条结果可回溯到完整输入；任务中断后只重跑未完成样本，避免重复调用模型服务
- 质量、成本、延迟和权限回归分别判定，服务调用失败或超时单独记为基础设施问题；GitHub Actions 每晚运行回归并保存报告，前端可对比基线与三组候选配置

安全业务工作流 Agent | FastAPI, PostgreSQL, Redis, MCP | GitHub

- 模型只提交固定结构的操作提案，真正的 RBAC 与租户策略由服务端执行；退款等高风险操作持久化等待独立审批人确认
- 用持久化检查点、outbox 和幂等键处理超时、取消与请求重放；公开演示展示审批、提示注入和重复请求下的状态变化

TinyLlama CUDA 推理引擎 | C++20, CUDA, CMake, cuBLAS | GitHub

- 作为推理系统补充，完成 TinyLlama-1.1B CPU FP32 参考路径、CUDA FP16 和 Hybrid W8A16；RTX 3080 Laptop 上 Batch 4 吞吐为 Batch 1 的 2.905 倍，W8A16 峰值显存降低 40.4%

专业技能

- **编程语言**：Python、SQL、C++
- **应用开发**：FastAPI、PostgreSQL/pgvector、Redis、MinIO、Docker、Vue 3
- **工程工具**：Linux、Git、GitHub Actions、OpenTelemetry、Prometheus
- **专业基础**：数据结构与算法、操作系统、编译原理、计算机体系结构、数据库
- **英语能力**：CET-4 582，熟练阅读英文技术文档与论文