杜旭嘉

AI 应用工程师 · RAG · Agent · 后端工程化

15035925107 2041487752dxj@gmail.com GitHub 作品集 国际站

求职方向：AI 应用工程师 · LLM 应用工程师 · 后端工程师（AI 工程方向）

个人简介

主要做 RAG、Agent Runtime、业务后台和多端系统，习惯把 AI 能力接进真实业务流程，而不是只停在单接口问答。熟悉 Python · FastAPI、Java · Spring Boot、React · Next.js，能独立完成方案设计、核心开发、联调、测试和上线。

工程上更关注可追溯、可恢复、可回放和可验证：检索结果要能看到证据来源，异步任务失败后要能恢复，支付与设备回调要能幂等处理，关键路径要有测试和运行指标。

核心能力

- **AI 应用开发**：熟悉 RAG、LangGraph、工具调用、工作流编排、多通道接入和知识库治理，能把问答流程接到查询、提交、转人工等业务动作
- **后端系统设计**：熟悉认证、支付、任务流、后台管理、数据迁移和性能治理，常用 FastAPI、Spring Boot、PostgreSQL、Redis、ClickHouse
- **质量与发布**：习惯用 pytest、Vitest、Playwright、lint · build、健康检查、回归测试和回滚方案控制上线风险
- **可验证结果**：在中国联通项目中将活动统计接口从 **20s+ 降到 4s**，完成 **300+ 表、3 亿+ 记录**迁移，部分核心查询从 **10s+ 降到 500ms**

工作经历

独立开发者

2025.12 - 至今

AI 应用 · 全栈项目

RentBox 共享擦窗宝小程序

Spring Boot · React · uni-app

- **支付设备一致性**：共享柜业务里支付、免押、取机、归还和柜机回调可能乱序到达，设计订单生命周期服务与状态机，用行锁、状态历史和审计日志把“是否可开锁”收敛到服务端状态真相
- **回调幂等与补偿**：接入微信支付 v3、微信支付分和柜机 Web API 时，把重复通知、超时重试和人工补偿纳入同一套 Redis/JDBC 幂等、分布式锁、回调回放链路，降低重复扣款、错发开锁和状态倒退风险

论文检索任务平台

FastAPI · React · PostgreSQL

- **共享表任务边界**：在甲方 Worker 只承诺读写共享 tasks 表、不能接入用户系统的约束下，补齐用户侧任务创建、状态同步、结果下载和额度展示，并用 LISTEN/NOTIFY、SSE 与轮询兜底推送检索进度
- **终态契约与扣减防重**：把额度扣减放到任务成功完成后执行，通过 Alembic check constraint/trigger 约束终态回退和成功任务字段，再用 SELECT ... FOR UPDATE、唯一额度事件和并发测试避免重复扣减

智能客服运行时

Agent Runtime · FastAPI

- **业务型客服 Runtime**：将客服从单轮 RAG 问答扩展为可接文本、语音/RTC、宿主页面和业务查询的运行时，统一渠道接入、宿主桥接、核心引擎、业务增强、插件平台和模型适配边界
- **路由与宿主解耦**：基于 intent_stack、page_context、business_objects 做上下文路由，低置信或高风险场景转人工；通过 7 类插件和 AuthBridge 复用宿主鉴权，避免把行业工具写死在主链路

微信智能助手

LangGraph • Electron • RAG

- **桌面自动化稳定性**：针对 Windows 微信版本、权限和消息通道不稳定的问题，抽象 BaseTransport 收敛接入差异，并用 LangGraph 将首响应链路与事实提取、向量写回、画像更新等后台成长任务解耦
- **长期运行治理**：围绕 2 秒回复预算构建 SQLite 短期记忆、运行期向量记忆和导出语料 RAG，并补齐 /api/status、/api/metrics、配置审计、备份恢复和离线 eval，让故障定位与质量回归可验证

中软国际

2025.04 - 2025.09

企业知识库 RAG · Agent 工作台 · 后端 · 全栈工程师

- **证据优先的 RAG**：企业知识库不能只生成“像答案的答案”，需要同时说明依据、拒答边界和检索诊断；设计结构检索、全文检索、向量检索三路召回，结合 query rewrite、加权 RRF 与 rerank 返回 citations、grounding_score、trace_id
- **可恢复运行与回归门禁**：将 Gateway 问答流程迁移到 LangGraph，支持 checkpoint、interrupt/resume 和人工澄清；建设 ingest、知识治理、retrieve/debug、Prompt Injection 防护与 smoke-eval/regression gate，把问答质量纳入发布验证

国家骨科临床研究中心

2024.08 - 2024.10

SubMed 医学论文检索小程序 · 后端开发实习生

- **医学检索闭环**：面向医生按疾病、术式和研究方向找 PubMed 文献的场景，参与小程序接口设计与联调，覆盖 AI 搜索、语音输入、论文详情、推荐、收藏、深度分析和分享
- **持续跟踪信息流**：对接 /api/articles/ai_search、/api/deep_analysis、/api/recommend_paper、订阅配置和推送历史，将一次性搜索扩展为可按关键词、影响因子和中科院分区持续追踪的科研信息流

中国联通陕西省分公司

2024.05 - 2024.08

触点赋能中心 · 陕西运营平台 · 数据中台 · 后端开发实习生

- **触点赋能中心**：在活动统计接口响应慢、运营侧需要高频看数的背景下，改造对外 API 二次授权认证，并通过 OLTP/OLAP 分离和 ClickHouse 聚合分析，将活动统计接口从 **20s+ 降到 4s**
- **陕西运营平台**：开发能力管理页面与后端逻辑，处理请求/响应加解密；维护陕西号码复通接口，并完成按模型扫描号码、自动开机的二次开发，支撑运营流程自动化
- **数据中台**：在不中断业务的前提下完成 **300+ 表、3 亿+ 记录**迁移和一致性校验，用 xxl-job 执行每日动态更新；重写聚合 SQL、增加覆盖索引和缓存后，部分核心查询从 **10s+ 降到 500ms**，发布耗时从 **30 分钟降到 5 分钟**

项目经历

RAG-QA System

FastAPI • Vue • LangGraph • Qdrant

- **多知识库问答系统**：面向企业知识库“答案必须能追溯来源”的需求，覆盖文档上传、异步解析、统一聊天、SSE 流式回答、引用回传和检索诊断
- **三路召回与引用溯源**：在检索服务中实现结构化检索、全文检索、向量检索三路召回，结合 query rewrite、RRF 和 rerank 返回 citations、grounding_score、耗时分解和 trace_id
- **LangGraph 流程重构**：使用 LangGraph 重构问答流程，引入 checkpoint、interrupt/resume 与 step_events，减少同一问题的重复检索和重复计费

EasyCloudPan

Spring Boot • React • PostgreSQL • Redis • MinIO

- **分片上传与断点续传**：面向大文件、弱网中断和重复上传场景，设计秒传、断点续传和 SSE 进度回传，支撑 **1000+ 并发上传**、上传成功率 **>99.5%**
- **请求签名与安全边界**：设计 HMAC-SHA256 + timestamp + nonce 请求签名，并接入 JWT 双 Token、Magic Number 校验、多租户隔离和文件访问切面，收紧文件平台安全边界
- **多级缓存与监控**：引入 Caffeine/Redis/DB 多级缓存和 Prometheus/Grafana 监控，项目文档口径下 API P95 **<500ms**、数据库查询 P95 **<100ms**

IPA 方言语音转写原型

Flask • PyTorch • Transformers • torchaudio

- **IPA 方言语音转写**：面向方言语音采集后的 IPA 转写整理，开发本地 Web 原型，支持批量上传、录音、单文件复转、SSE 进度和 Excel 导出
- **音频标准化与音标拆分**：使用 transformers.pipeline、torchaudio 和 ffmpeg 标准化音频输入，并将音标拆成声母、韵母、声调，方便研究人员后续复核与整理

九州通四向穿梭车路径规划系统

Python • A* • PyQt5 • multiprocessing

- **四向穿梭车路径规划**：面向医药物流仓库中的四向穿梭车调度，基于 42×27 仓储网格建模道路、货架与障碍，完成 A* 寻路、地图编辑和多车路径展示原型
- **通行规则与冲突处理**：设计空车/载货两套通行规则和基于优先级的时间步冲突处理，并用 PyQt5 逐帧展示搜索过程、等待动作和调度结果

技能栈

- **AI 工程**：RAG、LangGraph、Agent Runtime、工具调用、工作流编排、OpenAI API、向量检索、引用溯源
- **后端与数据**：Python、FastAPI、AsyncIO、Pydantic、SQLAlchemy、Java、Spring Boot、MyBatis、PostgreSQL、MySQL、Redis、ClickHouse
- **全栈交付**：React、Next.js、TypeScript、Vite、uni-app、REST API、支付集成、管理后台、多端协同
- **工程质量**：pytest、pytest-asyncio、Vitest、Playwright、Prometheus、Grafana、健康检查、发布回滚、lint、build、E2E

教育背景

南阳理工学院 · **数据科学与大数据技术** 2021.09 - 2025.06

- GPA 3.8/4.5，Top 5%
- 校级三好学生、校级优秀学生干部、院级奖学金