

徐青云

后端 / AI 平台工程师 | 分布式系统 · LLM 应用与可靠性

中国上海 | 133-0199-8597 | qingyunxu810@gmail.com | LinkedIn

技术栈

编程语言: Java、Python、TypeScript、JavaScript、SQL

后端与分布式系统: Spring Boot、RESTful API、Kafka、Pulsar、Redis、FastAPI

AI 与检索: LangChain、RAG、LLM Agent、Embedding、多查询检索、RRF、PyTorch、TensorFlow

数据与云: PostgreSQL、pgvector、MySQL、AWS、Docker

前端与工程实践: React、Next.js、Tailwind CSS、TestNG、Pytest、Git、CI/CD、Grafana、Splunk、OpenObserve

工作经历

俺来也（上海）网络科技有限公司

2026年7月 — 至今

Java 开发工程师

中国上海

- 重构短信登录与支付流程，使移动 App 与微信小程序复用同一套后端服务；阅读平台分布式锁切面源码，发现锁键空值退化缺陷波及 40+ 调用点，改用按用户 Redis 限频规避并推动平台修复。
- 通过 Transfer Channel 抽象解耦旧提现流程，新增银行直连通道，并使用银行测试系统验证 Pulsar 异步处理与 XXL-JOB 状态对账。
- 设计并交付 App 跨端支付桥接（Create / Get / Consume / Status 四端点），以短时 Redis Ticket 寄存支付上下文、按用户限频防重，在不改动既有小程序支付时序的前提下支持 App 发起支付，通过 HTTP Smoke Test 并交接前端联调。

Walmart Global Tech

2025年5月 — 2025年10月

Java 后端开发工程师

加州 Sunnyvale

- 使用 Java / Spring Boot 开发购后订单修改 API，将业务规则转化为商品资格校验、请求响应映射与状态校验逻辑，保证跨服务的订单修改结果一致。
- 接入 Kafka 订单修改事件，将同步 API 处理与下游更新和批处理解耦；使用 TestNG 建设跨服务 E2E 测试，并结合 Grafana、Splunk、OpenObserve 追踪调用链路，排查延迟与依赖服务故障。
- 开发内部 AI 订单信息聚合工具，按用户 / 订单 ID 聚合上下游多服务数据，用 LLM 提取关键字段与调用关系并标注来源，减少手动比对复杂报文的排查时间。
- 参与 Checkout 数据源迁移期间线上订单批量失败的排查：分析失败样本、构造 E2E 与集成复现，向责任团队提供报文与复现步骤，推动定位到配置变更。

Innovation AI

2024年7月 — 2024年12月

软件工程实习生

加州 San Jose

- 使用 FastAPI 开发 AI 阅读平台后端服务，管理 Amazon S3 中的图书内容，并支持文档处理与模型对话流程。
- 使用 PostgreSQL 和 pgvector 构建跨文档 RAG 流程，从多本图书中检索相关内容，让 LLM 回答更有依据。
- 使用多查询检索与 RRF 融合结果，改善跨文档内容覆盖和排序，再用于构建 LLM 上下文。

项目经历

结构化多 Agent 审议框架

2026年3月 — 至今

- 面向高风险 Prompt 设计职责分离的多 Agent 评审系统，将事实核查、风险与反例分析、意图理解和可执行性评估交给不同角色。
- 实现 Schema 驱动的状态机，覆盖 Proposal、Review、Conflict Ledger 和 Revision 阶段，在整个评审流程中支持结构化解析与问题追踪。
- 构建路由控制器，根据评审结论、严重程度和问题类型，将未解决问题分配到定向修订、用户澄清、外部验证、继续评审或终止流程。

教育背景

东北大学 (Northeastern University)

2023年1月 — 2024年12月

计算机科学理学硕士

加州 San Jose

密歇根州立大学 (Michigan State University)

2017年9月 — 2022年12月

计算数学理学学士，副修计算机科学

密歇根州 East Lansing

专业认证

AWS Certified Solutions Architect — Associate

2025年3月 — 2028年3月