

俞月伦

上海, 中国 | +86 180 1707 8930 | ar8327k@gmail.com

GitHub/ar8327 | LinkedIn/俞月伦

英语水平: CEFR C1, 雅思 7.5 | 日语水平: JLPT N2, 词汇参考级别: A, 语法参考级别: A

个人简介

拥有五年工程经验的高级软件工程师, 主线是 AI Agent 基础设施、评测/记忆工程体系与 Coding Agent 产品化。在字节跳动先后主导或核心参与 LLM Gateway Prompt Cache、AIGO 评测回放与记忆系统、TRAE CLI 2.0 等工作, 负责从网关、Agent Harness、工具链到发布/分发的关键模块。擅长把 LLM 能力落到高可用后端系统和可回归工程流程中, 主要使用 Go、Python。

技术能力

- **编程语言:** Go、Python、Shell、SQL
- **AI/Agent:** Coding Agent、Agent Loop/Harness、Prompt Cache、Tool Use / Function Calling、Tool Call Replay、Agent 评测与可观测性、多 Agent 协作、RAG 与记忆管理
- **后端基础设施:** LLM Gateway、MySQL、Redis、Elasticsearch、MongoDB、Kafka、RocketMQ、gRPC、Docker
- **框架与工具:** Gin、Claude Agent SDK、OpenAI API、飞书/Lark API

工作经历

高级软件工程师 - TRAE CLI 2.0 / AI Native 开发

2025 年 12 月 - 至今

字节跳动, 杭州

- 核心参与 TRAE CLI 2.0 研发: 基于 Codex 工程基座构建多模型 CLI Coding Agent, 负责 Agent Loop、Compact 上下文压缩、插件系统、LLM Transport、发布/分发系统等关键模块。
- 设计并落地 AI Native 研发流程, 将发版/回滚、用户反馈答疑、Bug 归因、竞品与上游跟踪、需求实现与自验收(测试/截图/录屏)等环节交给 Agent 承接, 减少重复人工投入。
- 推进 Agent 效果自迭代闭环: 将线上反馈转化为可回归 case, 结合评测回放定位 Agent Loop、Prompt 或工具链问题, 驱动功能与效果持续调优。

高级软件工程师 - AIGO Agent 评测与记忆系统

2025 年 6 月 - 2025 年 11 月

字节跳动, 杭州

- 为 Argos SRE 智能体 AIGO 从 0 建设评测回放体系: 录制历史 Tool Call 结果, 回测时返回当时的指标/日志/APM 数据, 解决观测数据过期导致历史 case 无法复现的问题。
- 设计双轨评测体系: LLM-as-judge 用于线上 case 快速一致性评估, 人工 golden 数据集用于长期确定性回归, 形成线上 case、快评、精标、回放回归的闭环。
- 设计基于文件系统的 Agent 记忆系统: 以 MEMORY.md 根节点和 topic 文件组织长期记忆, 支持 impression 写入、自动 merge/split、miss 驱动的召回优化, 成为 AIGO 记忆能力及后续 AI Native 开发的基础。
- 探索多 Agent 编排自动构建 Go LSP: 设计规划、生成、编译测试、反馈迭代的 Agent Loop, 验证复杂软件自动开发的可行性和边界。

高级软件工程师 - Trae LLM Gateway / Prompt Cache

2025 年 2 月 - 2025 年 6 月

字节跳动, 杭州

- 从 0 主导建设 Trae LLM Gateway Prompt Cache: 在 OpenAI、Gemini、Anthropic 及国产模型等异构厂商上抽象统一缓存语义, 对客户端提供稳定协议并完成底层适配。
- 与客户端协同优化 Prompt 结构, 保证 system prompt 稳定、Context 追加式增长; 基于用户 session 做缓存粘性与 TTL 内优先调度, 将缓存命中率提升至 90%+, 带来成本、首 token 延

迟和 TPM 约 90% 量级收益。

- 建立命中率、成本、延迟、token 计费监控看板，灰度阶段发现 cwd 写入 system prompt 导致命中率下降并及时止损；通过数据分析发现 Gemini thinking tokens 计费异常，推动厂商修正并追回退款。

高级软件工程师 - APM 元数据系统

2024 年 4 月 - 2025 年 6 月

字节跳动，杭州

- 独立设计并实现支撑 200 万服务的元数据管理系统，QPS 3000+，SLA > 99.9%。
- 实现类 SQL 的 DSL 查询引擎，自动解析字段依赖关系，统一 20+ 个外部数据源的访问；设计多级缓存架构（Redis + MongoDB）和智能降级策略，外部服务故障零影响。

高级软件工程师 - IT 资产与 CMDB 数据平台

2020 年 11 月 - 2024 年 4 月

字节跳动，北京 / 杭州

- 主导百万级配置项 CMDB 平台重构（IT 资产管理系统的底层数据基建），服务 500+ 内部系统。将 Python/Django 重构为 Go 微服务架构，吞吐量提升 6 倍（500→3000 QPS），P99 延迟优化 50 倍（10s→200ms）。
- 构建 MySQL + Elasticsearch 双存储引擎 + Canal 准实时同步管道；开发 DSL 查询语言屏蔽底层存储差异；实施灰度发布策略，确保 3 个月重构零故障上线。
- 设计上层 IT 资产管理系统（百万级服务器备件管理），支撑日均数百台故障处理，维修时长缩短 30%。

个人项目

自托管多用户 AI Agent 系统（基于开源项目深度改造）

github.com/ar8327

- 基于开源 AI Agent 框架进行深度改造，支持飞书/Telegram/QQ/Web 多渠道接入，基于 Claude Agent SDK 实现 Agent 执行引擎，支持 Docker 容器隔离与宿主机双模式运行。
- 自研 Memory Agent 记忆管理系统：设计 per-user 独立记忆子进程，实现自动会话归档、印象提取、知识整理、索引自修复、语义检索和多层记忆结构，解决 LLM Agent 长期记忆缺失问题。

专利

- 一种图像中数据自动提取的方法，发明专利，专利号：CN201910972334.8，合作作者
- 基于双通道卷积神经网络识别噪声非法文本的方法，专利号：CN109670041A，合作作者

荣誉奖项

- 国家奖学金（2018、2020 年）
- 教育部服务外包创新大赛全国二等奖（2020 年）
- 英特尔嵌入式系统设计大赛全国二等奖（2018 年）

教育背景

杭州电子科技大学

2017 年 9 月 - 2021 年 6 月

信息系统本科，GPA: 4.18 / 5.0

中国 · 杭州