

Yuelun Yu

Shanghai, China | +852 9488 2685 | ar8327k@gmail.com
github.com/ar8327 | LinkedIn/Yuelun Yu

Visa Sponsorship Required | CEFR C1, IELTS 7.5 | JLPT N2, Vocabulary: A, Grammar: A

Summary

Senior Software Engineer with 5 years of experience across AI Agent infrastructure, evaluation/memory systems, and Coding Agent productization. At ByteDance, led or played a core role in LLM Gateway Prompt Cache, AIGO evaluation replay and memory, and TRAE CLI 2.0, owning key modules from gateway services and Agent Harnesses to toolchains and release/distribution systems. Strong in turning LLM capabilities into highly available backend systems and regression-friendly engineering workflows, mainly with Go and Python.

Technical Skills

- Languages: Go, Python, Shell, SQL
- AI/Agent: Coding Agents, Agent Loop/Harness, Prompt Cache, Tool Use / Function Calling, Tool Call Replay, Agent evaluation & observability, multi-Agent collaboration, RAG & memory systems
- Backend Infrastructure: LLM Gateway, MySQL, Redis, Elasticsearch, MongoDB, Kafka, RocketMQ, gRPC, Docker
- Frameworks & Tools: Gin, Claude Agent SDK, OpenAI API, Feishu/Lark API

Professional Experience

- Senior Software Engineer - TRAE CLI 2.0 / AI-Native Development Dec 2025 – Present
ByteDance, Hangzhou, China
- Core contributor to TRAE CLI 2.0, a multi-model CLI Coding Agent built on the Codex engineering base; owned key modules including Agent Loop, Compact context management, plugin system, LLM Transport, and release/distribution pipelines.
 - Designed and landed AI-native engineering workflows where Agents handle release/rollback, user support, bug triage, competitor/upstream tracking, feature implementation, and self-verification with tests, screenshots, and recordings.
 - Drove an Agent effectiveness feedback loop: turning online feedback into replayable cases, using evaluation replay to locate Agent Loop, Prompt, or toolchain failures, and feeding the results back into product and harness iteration.
- Senior Software Engineer - AIGO Agent Evaluation & Memory Systems Jun 2025 – Nov 2025
ByteDance, Hangzhou, China
- Built AIGO evaluation replay from scratch for an Argos SRE Agent: recorded historical Tool Call results and replayed the original metrics, logs, and APM data during regression, solving the issue of expired observability data making old cases unreproducible.
 - Designed a dual-track evaluation system: LLM-as-judge for fast consistency checks on online cases, and manually curated golden datasets for long-term deterministic regression, forming an online case to quick judge to annotation to replay loop.
 - Designed a file-system-based Agent memory system with a MEMORY.md root and topic files, supporting impression writing, automatic merge/split, and miss-driven retrieval optimization; it became the memory foundation for AIGO and later AI-native development.
 - Explored multi-Agent orchestration for automatically building a Go LSP, designing a plan-code-build-test-feedback loop and validating both the feasibility and limits of AI-automated complex software development.
- Senior Software Engineer - Trae LLM Gateway / Prompt Cache Feb 2025 – Jun 2025
ByteDance, Hangzhou, China
- Led Trae LLM Gateway Prompt Cache from scratch, abstracting unified cache semantics across heterogeneous providers including OpenAI, Gemini, Anthropic, and domestic models, while exposing a stable protocol to clients.
 - Collaborated with clients to stabilize system prompts and make context append-only; added session affinity and TTL-aware priority scheduling, improving Prompt Cache hit rate to 90%+ and delivering roughly 90% gains in cost, first-token latency, and TPM.

- Built monitoring dashboards for hit rate, cost, latency, and token billing; caught a client change that put cwd into system prompts and hurt cache hits, and identified Gemini thinking-token billing anomalies that led to vendor correction and refunds.

Senior Software Engineer - APM Metadata System

Apr 2024 – Jun 2025

ByteDance, Hangzhou, China

- Independently designed and implemented metadata management system supporting 2M services with 3000+ QPS and 99.9%+ SLA.
- Implemented SQL-like DSL query engine with automatic field dependency resolution, unifying access to 20+ external data sources; designed multi-tier caching (Redis + MongoDB) and intelligent degradation strategy, achieving zero impact from external service failures.

Senior Software Engineer - IT Asset & CMDB Data Platform

Nov 2020 – Apr 2024

ByteDance, Beijing / Hangzhou, China

- Led the refactoring of the million-scale CMDB platform (the underlying data infrastructure for the IT asset management system), serving 500+ internal systems. Migrated from Python/Django to Go microservices architecture, boosting throughput 6x (500 to 3000 QPS) and optimizing P99 latency 50x (10s to 200ms).
- Built MySQL + Elasticsearch dual-storage engine with Canal-based near-real-time sync pipeline; developed DSL query language abstracting underlying storage differences; adopted canary release strategy, ensuring zero-incident deployment over 3-month refactor.
- Designed the upper-layer IT asset management system (million-scale server spare parts management), supporting hundreds of daily fault repairs with 30% reduction in repair time; built distributed pre-reservation inventory mechanism and multi-layer hardware compatibility rule engine.

Personal Projects

Self-hosted Multi-user AI Agent System (Deep Modification of Open-source Project)

github.com/ar8327

- Deeply modified an open-source AI Agent framework to support multi-channel access (Feishu/Telegram/QQ/Web); built Agent execution engine on Claude Agent SDK with both Docker container isolation and host process modes.
- Designed and built Memory Agent system with per-user memory subprocesses, automatic session archival, impression extraction, knowledge organization, index self-repair, semantic retrieval, and multi-layer memory structures, addressing the long-term memory gap in LLM Agents.

Patents

- Method for Automatic Extraction of Data in Images, Invention Patent, CN201910972334.8, co-authored
- Identifying Noisy Illegal Short Texts using Dual-Channel CNN, CN109670041A, co-authored

Awards

- National Student Scholarship (2018, 2020)
- Second Prize, Service Outsourcing Innovation Competition, Ministry of Education (2020)
- Second Prize, Intel Embedded System Design Contest (2018)

Education

Hangzhou Dianzi University

Sep 2017 – Jun 2021

B.Sc. in Information Systems, GPA: 4.18/5.0

Hangzhou, China