

Yuelun Yu

中国・上海 | +852 9488 2685 | ar8327k@gmail.com

github.com/ar8327 | LinkedIn/Yuelun Yu

ビザサポートが必要 | CEFR C1, IELTS 7.5 | JLPT N2, 語彙：A, 文法：A

概要

5年のエンジニアリング経験を持つシニアソフトウェアエンジニア。技術軸はAI Agent 基盤、評価/記憶システム、Coding Agent のプロダクト化。ByteDance にて LLM Gateway Prompt Cache、AIGO 評価リプレイ/記憶システム、TRAE CLI 2.0 に主導またはコアメンバーとして関わり、ゲートウェイ、Agent Harness、ツールチェーン、リリース/配布基盤まで担当。LLM 能力を高可用なバックエンドと回帰可能な開発プロセスに落とし込むことを得意とし、主に Go・Python を使用。

技術スキル

- **言語:** Go、Python、Shell、SQL
- **AI/Agent:** Coding Agent、Agent Loop/Harness、Prompt Cache、Tool Use / Function Calling、Tool Call Replay、Agent 評価 & オブザーバビリティ、マルチ Agent 協調、RAG& 記憶管理
- **バックエンド基盤:** LLM Gateway、MySQL、Redis、Elasticsearch、MongoDB、Kafka、RocketMQ、gRPC、Docker
- **フレームワーク & ツール:** Gin、Claude Agent SDK、OpenAI API、Feishu/Lark API

職務経験

シニアソフトウェアエンジニア - TRAE CLI 2.0 / AI Native 開発

2025 年 12 月 – 現在

ByteDance (中国・杭州)

- TRAE CLI 2.0 のコア開発に参加。Codex のエンジニアリング基盤上にマルチモデル CLI Coding Agent を構築し、Agent Loop、Compact コンテキスト管理、Plugin システム、LLM Transport、リリース/配布基盤などの主要モジュールを担当。
- AI Native 開発フローを設計・実装。リリース/ロールバック、ユーザー問い合わせ対応、Bug 原因分析、競合/上流追跡、機能実装、自動検証（テスト/スクリーンショット/録画）を Agent に委譲し、反復作業を削減。
- Agent 効果の自動改善ループを推進。オンラインフィードバックを回帰可能な case に変換し、評価リプレイで Agent Loop、Prompt、ツールチェーンの問題を特定して改善に反映。

シニアソフトウェアエンジニア - AIGO Agent 評価 & 記憶システム

2025 年 6 月 – 2025 年 11 月

ByteDance (中国・杭州)

- Argos の SRE Agent である AIGO 向けに評価リプレイ基盤をゼロから構築。過去の Tool Call 結果を記録し、回帰時に当時のメトリクス/ログ/APM データを返すことで、観測データの期限切れによる再現不能問題を解決。
- 二系統の評価体系を設計。LLM-as-judge でオンライン case を高速に一貫性評価し、人工 golden データセットで長期的な確定回帰を行う、オンライン case から評価・精密ラベル付け・リプレイ回帰までの閉ループを構築。
- ファイルシステムベースの Agent 記憶システムを設計。MEMORY.md ルートと topic ファイルで長期記憶を管理し、impression 書き込み、自動 merge/split、miss 駆動の検索最適化を実現。AIGO および後続の AI Native 開発の基盤となった。
- マルチ Agent 編成による Go LSP 自動構築を探索。計画、コード生成、ビルド/テスト、フィードバック反復の Agent Loop を設計し、複雑なソフトウェア開発自動化の可能性と限界を検証。

シニアソフトウェアエンジニア - Trae LLM Gateway / Prompt Cache

2025 年 2 月 – 2025 年 6 月

ByteDance (中国・杭州)

- Trae LLM Gateway Prompt Cache をゼロから主導。OpenAI、Gemini、Anthropic、国内モデルなど異種 Provider 上に統一されたキャッシュ意味論を抽象化し、クライアントに安定したプロトコルを提供。
- クライアントと協調して Prompt 構造を最適化し、system prompt の安定化と Context の追記型設計を徹底。session 粘着と TTL 内優先スケジューリングにより Prompt Cache ヒット率を 90%+ まで向上させ、コスト、first-token latency、TPM で約 90% 規模の改善を実現。
- ヒット率、コスト、レイテンシ、token 課金の監視ダッシュボードを構築。cwd が system prompt

に混入してヒット率を下げる問題や Gemini thinking token 課金異常を検出し、修正と返金につなげた。

シニアソフトウェアエンジニア - APM メタデータシステム

2024 年 4 月 – 2025 年 6 月

ByteDance (中国・杭州)

- 200 万サービスをサポートするメタデータ管理システムを独立設計・実装。QPS 3000+、SLA 99.9% 以上。
- SQL 類似 DSL クエリエンジンを実装し、フィールド依存関係を自動解析、20 以上の外部データソースアクセスを統一化。多層キャッシュアーキテクチャ (Redis + MongoDB) とインテリジェント降格戦略を設計し、外部サービス障害のゼロ影響を実現。

シニアソフトウェアエンジニア - IT 資産 & CMDB データ基盤

2020 年 11 月 – 2024 年 4 月

ByteDance (中国・北京/杭州)

- 百万規模構成項目 CMDB プラットフォーム (IT 資産管理システムの基盤データインフラ) の再構築を主導。500 以上の内部システムに対応。Python/Django から Go マイクロサービスへ移行し、スループット 6 倍 (500→3000 QPS)、P99 レイテンシ 50 倍改善 (10s→200ms) を実現。
- MySQL + Elasticsearch デュアルストレージエンジン + Canal 準リアルタイム同期パイプラインを構築。DSL クエリ言語で基盤ストレージ差異を抽象化。カナリアリリース戦略を採用し、3 ヶ月の再構築プロジェクトを無障害で稼働。
- 上層の IT 資産管理システム (百万規模サーバースペア部品管理) を設計。日平均数百台の障害修理をサポートし、修理時間を 30% 短縮。分散予約在庫メカニズムと多層ハードウェア互換性ルールエンジンを構築。

個人プロジェクト

セルフホスト型マルチユーザー AI Agent システム (オープンソースプロジェクトの深度改造)

github.com/ar8327

- オープンソース AI Agent フレームワークを深度改造し、Feishu/Telegram/QQ/Web マルチチャネル接続をサポート。Claude Agent SDK に基づく Agent 実行エンジンを構築し、Docker コンテナ分離とホストプロセスの双方モードで稼働。
- Memory Agent 記憶管理システムを独自設計。per-user 独立記憶サブプロセス、自動セッションアーカイブ、印象抽出、知識整理、インデックス自己修復、セマンティック検索、多層記憶構造を実装し、LLM Agent の長期記憶欠損問題を解決。

特許

- 画像中のデータ自動抽出方法、発明特許、特許番号：CN201910972334.8、共著
- デュアルチャネル畳み込みニューラルネットワークによるノイズ不正テキスト識別方法、特許番号：CN109670041A、共著

受賞歴

- 国家奨学金 (2018 年、2020 年)
- 教育部サービスアウトソーシングイノベーションコンテスト全国二等賞 (2020 年)
- インテル組込みシステム設計コンテスト全国二等賞 (2018 年)

学歴

杭州電子科技大学

情報システム学学士、GPA: 4.18 / 5.0

2017 年 9 月 – 2021 年 6 月

中国・杭州