

Yupeng Lu

US Permanent Resident | No sponsorship required
yplmicro@gmail.com | [linkedin.com/in/yupeng-lu-845a0b411](https://www.linkedin.com/in/yupeng-lu-845a0b411)
GitHub: github.com/BrickerP | Public work: github.com/cookiy-ai

AI Agent Engineer

SUMMARY

AI Agent Engineer and backend platform builder focused on production agent systems: MCP/tool-use contracts, public API boundaries, CLI/skill distribution, evidence-bound report synthesis, Report Voice Platform, monetized workflows, and evidence-gated releases. U.S. Permanent Resident; no sponsorship required.

PUBLIC PROOF

- [public user-research skill](#) - public AI-agent skill for end-to-end user research across Claude, Codex, Cursor, and OpenClaw-style clients.
- [public TypeScript CLI](#) / [npm CLI package](#) - TypeScript CLI for AI interviews, synthetic users, quant surveys, recruitment, playback, billing, and REST/MCP-compatible workflows.
- [MCP package](#) and [developer portal](#) - one-command skill/MCP setup, Redoc/OpenAPI docs, and public agent onboarding surface.

EXPERIENCE

Cookiy AI - AI Agent Engineer

Aug 2025 - Present | Silicon Valley HQ / Beijing Engineering Team

Voice-AI platform for user-research interviews. Built agent-native research infrastructure across MCP tools, public REST APIs, OAuth/billing, report generation, Report Voice Platform, interview/playback flows, and E2E/release gates. Evidence snapshot: **1,040 Cookiy merged PRs** spanning MCP/API, report synthesis, voice runtime, video clips, billing, and release gates.

- **Contract-first agent/API platform.** Designed and hardened Zod/OpenAPI/MCP-style tool contracts, DTO/runtime validation, pagination/status semantics, CLI payload compatibility, and public API boundaries.
- **MCP and tool registry.** Owned 41 tools across 4 MCP servers and 26 Zod/OpenAPI schema files, including study, interview, quant, billing, recruit, guide, playback, and report workflows.
- **CLI/skill distribution.** Published public skill and TypeScript CLI surfaces; migrated shell workflows to npm CLI; added MCP-first/REST-fallback behavior, token/login flows, and Cursor/Claude deep-link paths.
- **Tool-using report/study chat agents.** Re-architected scoped registries, artifact load-before-edit discipline, evidence bundles, direct-save tools, and locked-fact guards; removed 7 duplicate tools and slimmed prompts by ~2.5k LOC.
- **Evidence-bound report synthesis.** Moved research reports from fragile single-shot prompting to objective-first multi-pass synthesis with cohort-batch passes for large corpora, objective-independence and emergent-discovery passes, chunked HTML fidelity/readiness gates, report-driven auto clips (MediaConvert + GCP Transcoder fallback), screener/chat corpus QA, and editorial HTML preservation.
- **LLM production reliability.** Hardened OpenAI/Gemini execution paths under release pressure: native fetch, uncompressed SDK responses, preserved auth headers, typed client options, runtime pinning, Cloudflare/401 handling, and preview backports.
- **Streaming and long tool work.** Owned SSE/WebSocket/MCP streamable-HTTP paths with heartbeats, abort-cleanup isolation, hard timeouts, run-ID propagation, and stream-boundary regression tests.
- **Eval/release infrastructure.** Built dual-surface E2E harness (291 cases: 168 SaaS + 123 CLI/MCP), 7 L0-L6 gate profiles, preview SaaS/voice/share-link oracles, runtime evidence packets, trace correlation, p95 budgets, baseline ledger, and notification loops.
- **Realtime interview, avatar, and Report Voice Platform.** Worked across LiveKit rooms, Deepgram STT, TTS providers, Gemini/OpenAI LLM paths, and Tavus/SpatialReal avatar integration (~200 ms barge-in); shipped a dedicated report-voice delivery runtime with streaming TTS, speculative follow-up turn-taking, VAD barge-in, voice-role training, and Cookiy-account admin sessions.
- **Agent-compatible auth and billing.** Hardened compact CLI tokens, OAuth/Bearer semantics, wallet-ledger idempotency, StripeEventLog webhook audit, pay-before-reveal report billing, post-queue debit linking, and minimum-disclosure tool responses.
- **Platform/release operations.** Supported Nginx gateway/deploy paths, cross-environment short-link routing, cookie isolation, preview/backport propagation, and Homebrew installer distribution.

SMU - Software Engineer / 0-to-1 Product Owner

Feb 2024 - Jan 2026 | Los Angeles, CA

Data-platform role that expanded into a founder-sponsored 0-to-1 product initiative: after shipping core commerce/ERP infrastructure, independently built and operated [GetDateLove](#) as a production consumer product.

- Built Python AsyncIO pipelines connecting Amazon Seller Central + Shopify with ERP; processed 500k+ daily orders and cut sync latency from 15 min to 30 s.
- Shipped Redis/Kafka inventory sync, FastAPI/PostgreSQL tariff services, Elasticsearch HS Code lookup, Spark migration jobs, Docker/Jenkins CI/CD on AWS, and Sentry loops.
- Built GetDateLove end to end with FastAPI, Redis asyncio, PostgreSQL, WebSocket messaging, S3 voice notes, PayPal checkout, AWS/Nginx/CloudFront, CI/Playwright release gates, and GDPR/privacy controls.
- Reduced operational and frontend entropy by replacing blocking Redis patterns, consolidating 11 Redis clients, batching SQL, moving message storage to redis.asyncio, and extracting stable frontend state owners.

Quant Trading Systems Venture - Co-Engineer

Apr 2026 - Present

Private live-trading systems venture on AWS EC2: scanner to factor filter to IBKR bracket orders to S3 Parquet warehouse to researcher dashboard.

- Improved live execution safety with IBKR ibapi heartbeat/reconnect, account-update ordering, broker+DB startup reconcile, bracket/OCO error fixes, and deterministic flatten paths.
- Reduced scanner/data latency with event-based scanner waits, vectorized reject reasons, warm-cache fast paths, SQLite parse optimization, lazy-load startup paths, and reused timezone/schedule parsing.
- Built research/ops visibility: [live dashboard](#), Pipeline Health, Trade Cards, Factor/Calibration/Drift views, Signal Funnel, error timeline, gzip uploads, daily S3 Parquet archive, blocking mypy, AST invariant checks, and systemd/cron recovery assets.

Huawei Technologies - Software Engineer, SPU Development

Apr 2023 - Feb 2024 | Beijing

- Built C/C++ SPU/router features for NE40E traffic statistics and NetStream monitoring; fixed 50+ packet-processing, flow-log, URL parsing, HTTP header, and SmartNet resource issues.

EDUCATION

Boston University - M.S. Mathematical Finance & FinTech | GPA 3.78

Computational Methods, Stochastic Calculus, Algorithmic Trading

The Ohio State University - B.A. Physics | GPA 3.84 | Dean's List