Vu Nguyen

Founding MTS — environments and graders for coding agents: sealed microVM runtimes, evals that expose shortcuts, and runs you can checkpoint and resume

713-598-9156 · vuluan@gmail.com · github.com/vu1n · San Francisco (relocating)

EXPERIENCE

Independent Builder — Software Engineer 2025–present

- **Pillbox** (Rust · github.com/vu1n/pillbox) — runs coding agents in libkrun microVMs with credentials swapped in at the host boundary. Parallel candidates are graded in a separate microVM on a protected copy with tests and CI config reset and locked, so an agent can't edit its way to a pass. Paired trials showed checkpoint-gated task segmentation cutting run-to-run spread from σ 0.212 to 0.026.
- **Onoda** (Python, TypeScript) — ran coding agents on a real science workload (quantum chemistry with ORCA) and built the harness that decides whether they did it. Reference values live in an evaluator store the agent side cannot write; harness changes ship only through pre-registered gates with zero tolerance for false successes. Hill-climbing prompts and tool docs took supported claims from 39 to 54 and false successes from 15 to 0.
- Reported the negative result as one: a method reconstructed from a paper did not beat the baseline (46 vs 51 of 106 target states), with zero false successes in either arm. On the same harness, Claude Opus 5.5 placed 32 of 56 states within 0.3 eV of the reference, against 22 for DeepSeek V4 Flash and 23 for Sonnet 5.5. Billed cost ran 1.8× the agent's own estimate, so the harness meters billed cost.
- **Huddles** (TypeScript, Effect, Postgres, Cloudflare, MCP · demo reel) — the communication layer for agents. People and agents work in one room: every message, hand-off, decision and agent turn is an attributed event on one ledger, and agents are participants you address, not tabs you babysit. Each turn gets one sealed context packet; blind reviewers get none of the discussion, so they judge the work cold. Authority comes from attribution, never from text, so words in a message can't wake or empower an agent.
- Turns run on machines the member pairs in (Stratus: checkpoint, copy and resume a Claude Code, OpenCode or pi session across a microVM, a container or a Cloudflare Sandbox). Outside agents such as Nous Research's Hermes join by single-use invite over MCP without being trusted. A seeded simulator drives people, agents and nodes through the real server and checks invariants after every run.
- **Hyperstitch / Rubick** (Go, TypeScript, Python · rubick.xyz) — technical co-founder of a non-custodial copy-trading platform, live on Hyperliquid mainnet in closed beta; built backend, execution, data and infra mostly solo. The LLM classifier that turns social posts into trade signals is developed against a judge-panel eval harness; each graded item carries its evidence, and an exporter turns graded traces into SFT data. My non-technical co-founder ships frontend changes to production by directing Claude Code inside the typed contracts and CI I set up.
- **Tumbler** (TypeScript, Cloudflare · github.com/vu1n/tumbler) — secrets for agent sessions: values sealed in a Cloudflare Worker, each token pinned to the projects it may read, injected into one process and never written to disk. The agent never sees its own token.

Phantom — Senior Data Scientist Jan–Apr 2026 · Remote

- Built and owned an AI data exploration agent in production for analysts and non-engineers: retrieval over warehouse metadata, sandboxed Python analysis, an **MCP server exposing 15 tools** for Claude Code / Desktop, and HTML reports people could read without the agent.
- Evaluated the actual use case: retrieval weights tuned with Optuna + MLflow against a held-out set (recall@k, MRR), DSPy RLM schema navigation instead of dumping schemas into context, and per-turn traces of tokens, cost and tool calls. SQL AST validation, a query gateway and a tool allowlist meant the agent could not run a query nobody approved.
- When event coverage was the gap in an onboarding-funnel study, shipped the missing instrumentation in the mobile wallet repo myself and merged it with mobile engineering.

Consensys (MetaMask) — Staff Data Engineer 2022–2024 · Remote

- Led **schema-as-code** for MetaMask analytics events across product, engineering and analytics teams: canonical schemas in Git, tooling that validates changes, flags breaking ones and generates client types, enforced in CI. Adoption came from making the tooling the easy path. Malformed events stopped breaking dashboards and inflating pipeline cost.

Squarespace — Senior Software Engineer 2020–2022 · Remote

- Designed the cloud analytics schema during a warehouse migration; built ETL/ELT pipelines and reconciled every figure against legacy reports before cutover.

Earlier 2017–2020

- **Wanderjaunt** — Senior Software Engineer (2019–2020). Early-stage startup; built the end-to-end data and analytics stack.
- **Meta** — Data Engineer IC6 (2018–2019). App-responsiveness performance windows; matrix profiling for time-series anomaly detection.
- **Indeed** — Senior Data Engineer (2018). GDPR compliance across company-wide pipelines.
- **StockX** — Data Engineer Team Lead (2017–2018). Early data hire; led the data team through marketplace scale-up.

PROJECTS

Open source contributions *github.com/vu1n/hermes-huddles*

Upstream fixes to agent tools I run every day: NousResearch/hermes-agent (graceful fallback when a provider's vision model isn't entitled, blocked-task notifications for unattended runs, a Cloudflare Browser Rendering crawl backend), plus DefiLlama and GSV.

EDUCATION

- **University of Texas at Austin** — B.S., Computer Science
- **Georgia Institute of Technology** — Online M.S., Computer Science (in progress, 20 credit hours)

TECHNICAL

Languages	Python, TypeScript / Node.js, Go, Rust, SQL
Agents and evals	Claude / OpenAI / OpenRouter APIs; Claude Code, Codex, OpenCode, pi harnesses; MCP servers; DSPy, GEPA; held-out evals, judge panels, ablations, A/A noise runs, bootstrap intervals; small-scale LoRA SFT and GRPO
Sandboxes and infra	libkrun microVMs, Cloudflare Sandbox / Workers / D1 / R2, E2B, Docker, AWS, Pulumi, Kubernetes, OpenTelemetry
Data	PostgreSQL, DuckDB, ClickHouse, Snowflake, BigQuery, dbt, MLflow, Optuna