CODY LEE

Los Angeles, CA · Open to relocation
cody.lee.cl1@gmail.com · (626) 838-4155
GitHub: github.com/clee12111 · Portfolio: [your-domain.com]

Building end-to-end LLM pipelines — retrieval, agent loops, evaluation — and forensically documenting where they break.

EDUCATION

University of California, Riverside · B.S. Computer Science — Expected Jun 2026

Relevant coursework: Natural Language Processing, Artificial Intelligence, Data Science Ethics, Probability & Statistics, Operating Systems

PROJECTS

Polymarket Trading Bot Autopsy — Apr–Jun 2026

Published: 2-page reflective summary → github.com/clee12111/polymarket-autopsy/blob/main/reflective_2page.pdf

- Documented three measurement bug classes in a self-built trading pipeline that inflated paper trading results **~135×** and made the strategy actively misleading: **the bots my paper trading flagged as best performed worst live**.
- Built **3-layer LLM classification pipeline** (Haiku → Sonnet → Opus) classifying **2,409 wallets** into 12 strategy archetypes; **180 paper bots, 417,008 simulated trades**. Tiered-model design: **$30–40 total cost vs ~$150 if run on Opus alone**.
- 45-day solo project across Polymarket, Binance, and on-chain data; pipeline reconciled millions of cross-source transactions.
- Built and operated live execution environment with wallet auth, kill switches, and trade logging; deployed real capital across 6 architectures over 35 hours of live execution.
- Published 15-page technical autopsy and 2-page reflective summary covering the bugs, the methodology, and the infrastructure decisions worth keeping.

Aether — LLM Workflow Engine — Apr 2026

- Built hybrid RAG pipeline over financial compliance documents — **BM25 + all-MiniLM-L6-v2 dense embeddings** in ChromaDB, fused via **Reciprocal Rank Fusion**, then reranked with a **flashrank cross-encoder (ms-marco-MiniLM-L-12-v2)**; planner/executor/critic agent loop with audit trails.
- Retrieval precision: **96% on the eval suite**, perfectly reproducible across 5 runs; end-to-end correctness 87%; average cost **$0.65 per run** via tiered model routing (Opus plans, Haiku critiques).
- Engine/domain separation: retrieval and agent layers are generic; domain knowledge enters through configuration, so the architecture generalizes beyond financial documents.

ChainTax — Crypto Tax Engine — Feb–Mar 2026

- Built cross-source data ingestion and reconciliation pipeline: **540K+ events in a single pipeline run** across Hyperliquid API and two on-chain APIs (Alchemy, DeBank), with bridge detection so the same dollar isn't double-counted across chains.
- FIFO lot matching with SpecID retrospective comparison; three competing IRS funding treatments (Section 163(d), basis-adjustment, Section 165(c)(2)) produce three Form 8949 variants per pipeline run, with Section 1092 offsetting-position detection across spot, perp, and correlated pairs.

vLLM Retrieval Forensics — In Progress, 2026

- Building hybrid retrieval system over the vLLM codebase, with a stage-by-stage forensic study of where LLM-based code retrieval fails. Extends the polymarket autopsy methodology to AI infrastructure systems.
- Publishing pipeline-stage forensic case studies as work proceeds.

TECHNICAL SKILLS

Languages	Python, TypeScript, SQL, Bash
LLM / RAG	Hybrid retrieval (dense + sparse), chunking, embedding, reranking, vector databases (ChromaDB), prompt engineering, LLM evaluation, agent orchestration, Anthropic/OpenAI SDKs
Data	Cross-source ingestion and reconciliation, pandas, DuckDB, SQLite, async websocket pipelines
Infrastructure	Git, REST APIs, Streamlit, production deployment to VPS infrastructure
Methodology	Forensic measurement, failure-mode analysis, statistical evaluation, technical documentation