Adrian Hayden

AI / ML ENGINEER | INFERENCE OPTIMIZATION, CUDA & GPU INFRASTRUCTURE
adrian@littlehayden.net · Austin, TX · adrianhayden.com · github.com/ahayden04 · linkedin.com/in/adrian-j-hayden

Summary

ML infrastructure engineer focused on LLM inference and training systems. Raised single-GPU SGLang serving throughput 36% with custom CUDA kernels on a self-built Kubernetes GPU cluster; fine-tunes open-weight models and builds agent harnesses and evals. Previously owned delivery of AI tools and data workflows for a ~100-member client program at HyperTrend Labs, and built real-time Solana data pipelines at Lunar Capital.

Experience

Software Engineer & Technical Instructor · HyperTrend Labs (Contract)
Nov 2025 – Nov 2026 · Remote

- Owned client delivery of AI tools and data workflows for a ~100-member program, from requirements through documentation and handoff.
- Designed a 52-week technical curriculum and taught a weekly workshop to 25–50 clients and staff.
- Led design and build of a Go automation daemon for exchange and blockchain APIs: SQLite persistence, deterministic scheduling, live execution, health/NTP checks, pause/resume safeguards.
- Shipped static binaries and Docker images for AWS VPS deployment with fund-control keys separated from limited execution keys; deployed by every client and ran ~3 months with no significant incidents.
- Packaged a ~20 GB compressed research dataset (Python, Parquet, DuckDB) with provenance and validation.

Independent ML Research & Infrastructure · Self-directed
Jan 2024 – Present · Austin, TX

- Raised SGLang serving throughput 36% for Qwen3.8-27B (NVFP4) on one RTX 5090 via custom kernels (FP8 GEMM dispatch, fused FP4 activation, NVFP4 autotuning) and a quantized speculative draft; sustained 145 tok/s (+59%) at 216K context.
- Cut NVFP4 GEMM latency 30% (72.9 → 51.1 µs); extended serving to 256K context at 131.5 tok/s decode and 103 aggregate tok/s across 4 concurrent requests (up from 74).
- Fine-tuned a 7B decompilation model, trained on sets built by synthetic-data pipelines, that emits byte-exact functions on the first attempt 60%+ of the time; now training a sub-1B model from scratch.
- Built and operate a Kubernetes GPU cluster scheduling across heterogeneous nodes with pooled CPU, memory and storage; provisioned with Ansible, monitored with Grafana, and running custom CI/CD runners.

Financial Research Intern · Lunar Capital (Web3 venture studio)
Jun 2024 – Aug 2024 · Remote

- Built Solana wallet-cluster analysis on real-time Geyser streams (Yellowstone gRPC via QuickNode).
- Wrote trading-automation and wallet-key tooling and supported financial-model research on a four-person team.

Projects

Wispr Agent CLI and Evaluation Harness
- Built a Go CLI and OMP skill giving coding agents bounded, read-only access to Wispr Flow meeting transcripts: deterministic ID resolution, structured JSON/error envelopes, pagination, untrusted-content boundaries.
- Built a 10-case cross-model eval harness scoring correctness, tool efficiency, latency, token use and cost; both tested models scored 10/10.

Dealership Inventory Database · Hoover Chrysler Dodge Jeep RAM
Mar 2021 – Nov 2021 · Summerville, SC
- Integrated the third-party vAuto API for inventory tracking.
- Designed and implemented a SQL database managing ~$2.7M in inventory with an average turnover of 350 units per month.
- Eliminated paperwork errors, such as mistyped VINs, by populating the database from the vAuto API.

Education

Western Governors University
B.S., Software Engineering · Sep 2026

Trident Technical College
A.S. · 2022

Certifications

AWS Certified Cloud Practitioner · Amazon Web Services · Dec 2023

Skills

ML & Inference: SGLang (custom kernels), vLLM, CUDA, PyTorch, Hugging Face, NCCL, Nsight, fine-tuning, synthetic data generation, model evaluation, agent harness engineering
Infrastructure: Kubernetes, Docker, Linux, NVIDIA GPU clusters, AWS, Ansible, CI/CD, Grafana, Git, Networking
Languages: Python, Go, C++, Rust, SQL, Bash
Data: Parquet, DuckDB, SQLite, exchange & blockchain APIs

Languages

English