Trenton G. Milam

AI/ML Infrastructure Engineer

trent@trentmilam.dev · github.com/trentmilam · linkedin.com/in/trent-milam · Remote — US

SUMMARY

AI/ML Infrastructure Engineer who builds and operates local model-serving stacks, multi-GPU inference clusters, and agentic RAG pipelines. Designs reliability-engineering systems for GPU compute, automated testing frameworks, and inference-throughput optimization on constrained hardware — applying the verification-and-validation discipline of safety-critical aerospace systems to trustworthy, scalable ML infrastructure.

EXPERIENCE

AI Infrastructure Engineer · **Self-Directed Lab** 2024 – Present

- Built and operate a 2-node multi-GPU inference cluster (RTX 3090 + RTX 5090) with GPU fault diagnosis, wedge recovery, and thermal-aware autonomy for high availability.
- Architected a modular, offline-capable multi-agent AI platform (~32,000 LOC): OpenAI-compatible gateway, adversarial peer-review, and AI-safety controls with least-privilege tool bridges.
- Developed a tested job-intelligence platform (~15,158 LOC, 445 automated tests) with truth-checked LLM résumé generation, async multi-source ETL over 28,701 postings, and a FastAPI dashboard.
- Engineered GPU serving and reliability workflows: llama.cpp inference, KV-cache tuning, VRAM optimization, and automated recovery for hard GPU-wedge failures.

Test Engineer — RS-25 Flight Engines · **Aerojet Rocketdyne (L3Harris)** 2025 – Present

- Self-initiated AI: cut specification-research time ~85% by building a retrieval-augmented generation (RAG) system over a 100,000+ document technical corpus, expanding from a 35,000+ file initial pass.
- Self-initiated AI: built a read-only, least-privilege MCP bridge for autonomous corpus traversal — Qdrant vector DB, BAAI/bge embeddings, and a homegrown knowledge-graph RAG with CUDA-accelerated OCR ingestion.
- Perform cryogenic inspections, leak checks, and flight-readiness testing on RS-25 flight engines; develop test sequences and checkout procedures to flight-qualification and V&V standards.

Manufacturing Engineer · **Aerojet Rocketdyne (L3Harris)** Jul 2024 – Dec 2025

- Authored and executed 40+ V&V procedures, test sequences, and acceptance/checkout protocols aligned with flight-qualification standards.
- Developed Python/VBA automation of recurring documentation and data workflows, removing ~2–5 hours/week of repetitive manual effort.
- Led production capacity modeling for the entire Engine Assembly Facility (20+ units/work-centers), turning throughput and constraint data into planning-visibility dashboards.

Systems Engineer T1 · **Leidos Incorporated** May 2023 – Jul 2024

- Automated PCB acceptance testing with LabVIEW, helping cut test time ~5× across several hundred prototype boards for the Naval Surface Warfare Center.
- Authored and executed V&V procedures, work instructions, and acceptance-test protocols, establishing data-validation and QA standards.
- Performed system-level reliability and risk analyses on a ~$15M Navy underwater-acoustic (hydrophone array) program; supported high-voltage FATs and fiber-optic troubleshooting.

SKILLS

AI/ML Infrastructure: PyTorch, CUDA, llama.cpp, Ollama, model serving, multi-GPU orchestration, inference-pipeline tuning

Languages: Python, SQL, VBA, MATLAB

Systems & Ops: Linux/Windows admin, SSH tunneling, GPU diagnostics/reliability, LabVIEW, Wireshark

RAG & Agentic: Qdrant, BAAI/bge embeddings, knowledge-graph RAG, retrieval pipelines, MCP tool-use, multi-agent orchestration

Backend & Tooling: FastAPI, SQLite, pytest, async ETL, Docker, Git

EDUCATION

Mississippi State University — B.S. Aerospace Engineering (Astronautics), Minor in Mathematics May 2023

ADDITIONAL

U.S. citizen · Held a prior U.S. security clearance (currently inactive) — details on request · AIAA Member · INCOSE Associate Systems Engineering Professional