

Zachary Montague

Senior AI Platform Engineer · LLM inference, model post-training, and agent systems

Trenton, IL (St. Louis metro) · Contact through LinkedIn: [linkedin.com/in/thelinuxcowboy](https://www.linkedin.com/in/thelinuxcowboy) · github.com/zrmonty

SUMMARY

Senior AI Platform Engineer with 12+ years in cloud architecture, site reliability, and enterprise security. Builds and operates production LLM systems (model fine-tuning and post-training, multi-GPU inference serving on 2x NVIDIA H100, evaluation, and guardrails) for ~600 users in regulated financial services. Deploys high-throughput local GPU inference inside strict compliance boundaries.

EXPERIENCE

Senior AI Platform Engineer, Regulated financial services firm

Jul 2026 to Present

- Own the firm's internal inference platform, deploying and operating local GPU model serving so sensitive code and data never leave the compliance boundary.
- Post-train and fine-tune models on 2x NVIDIA H100s, and benchmark serving configurations (TP/DP scaling and 1.58-bit ternary quantization) against internal coding workloads, in progress.
- Built the serving and delivery platform behind 5+ production AI systems used by ~600 staff, standardized on LangChain/FastAPI services shipped through CI/CD pipelines with OpenTelemetry observability and MCP integrations.
- Built the evaluation and monitoring framework used across the firm's AI portfolio, measuring accuracy, reliability, token economics, and latency SLAs (TTFT/ITL) for every production model and agent.
- Set the platform's safety and access controls (guardrails, isolated Docker execution runtimes, Entra ID, network segmentation) and established vulnerability scanning after identifying an exposed staging environment.
- Advise Compliance and Information Security on tool vetting and technical risk reviews.
- Review platform submissions and set technical standards for internally built applications.
- Run the internal enablement platform for citizen developers, hardening, testing, and documenting prototypes for production.
- Mentor two engineers on the team and lead AI engineering training for 20+ staff.
- Shipped an Azure RAG service over government regulations and 100+ vendor contracts that flags noncompliant terms.
- Reduced inference costs by eliminating redundant agent runs.
- Built agent-scored prospect research, yielding hundreds of warm introduction leads.

Principal Security Consultant & Founder, Nomad Cybersecurity Group

Nov 2023 to Jun 2026 · Remote

- Delivered cloud architecture reviews, infrastructure hardening, remediation plans, and zero trust controls for a primary client.
- Earned rewards for valid DeFi security findings through Code4rena, Sherlock, CodeHawks, and ImmuneFi.

Site Reliability Engineer, Gen

May 2022 to Nov 2023 · Remote

- Built Datadog SLIs/SLOs and Python observability for 50+ services.
- Automated PagerDuty incident workflows.
- Implemented ArgoCD GitOps with hardened Kubernetes and audit trails, reducing deployment errors by 40%.

Cloud Engineer, Saint Louis University School of Medicine

Feb 2020 to May 2022 · St. Louis, MO

- Secured AWS research environments with KMS and TLS.
- Automated VPC and security group provisioning with Terraform and Ansible.
- Managed HPC resources, cloud budgets, and disaster recovery validation.

Cloud Engineer, Saint Louis University

Jul 2019 to Feb 2020 · Remote

- Built AWS/Azure infrastructure and GitLab CI pipelines with autoscaling for peak registration loads.

Earlier roles, Multiple employers

2006 to 2019

- Senior System Administrator, RnS Distribution (2017 to 2019).
- System Administrator, Stoopid Chikin (2015 to 2017).
- Technical Support & Junior Administration (2006 to 2015).

PROJECTS

Local LLM Decision Engine: Logit Scoring Ticket Router

Oct 2026 · Python, SGLang, CUDA, ROCm, Llama 3.2

- Built a local decision engine with CUDA and ROCm deployment support that routes support tickets by reading next-token logits on SGLang /v1/score (Llama 3.2).
- Designed each decision as a single prefill with a restricted softmax over label tokens and a margin policy for automatic routing or human review.
- Benchmarked three serving lanes (logit scoring, constrained decode, free generation), achieving 4.5x to 5.6x faster decision latency with zero unparseable outputs, compared with 4 to 6 failures per 30 tickets in free generation.

Multi-GPU LLM Inference Benchmarking & Post-Training Pipeline

In progress · 2x NVIDIA H100, vLLM, SGLang

- Fine-tuning and post-training open-weights models (SFT, LoRA/QLoRA) and running multi-GPU serving benchmarks on 2x NVIDIA H100.
- Comparing tensor parallelism (TP=2) with data-parallel replicas (DP=2) across BF16 and quantized checkpoints.
- Measuring TTFT, ITL, KV-cache saturation, and P99 SLA goodput under concurrent load to find the best throughput-per-dollar serving topology.

Proofwork Plugin

2026 · Codex, Claude Code

- Built an engineering plugin for Codex and Claude Code that guides agents through planning, implementation, code review, and verification workflows.
- Published it to the ChatGPT plugin marketplace.

SKILLS

Inference, Post-Training & Serving: vLLM, SGLang, CUDA, Multi-GPU Orchestration (2x NVIDIA H100, TP/DP scaling), Model Fine-Tuning & Post-Training (SFT, LoRA/QLoRA, DPO), Post-Training Quantization (AWQ, 1.58-bit ternary), Logit-Level Scoring, Inference Benchmarking (TTFT/ITL/KV-cache profiling), ROCm/HIP, LLM APIs, MCP

Platform & Cloud: Kubernetes, Docker, ArgoCD/GitOps, Terraform, Ansible, AWS, Azure, Linux, GitLab CI

AI Engineering: Python, FastAPI, LangChain, RAG, Embeddings, Vector Databases, Multi-Agent Orchestration, Function Calling/Tool Use, Agent Evaluation, Prompt Design, REST APIs, Bash

Reliability & Security: OpenTelemetry, Datadog, PagerDuty, SLIs/SLOs, Entra ID, IAM, Zero Trust, Sandboxed Execution, Vulnerability Scanning, AI Governance