Gowtham Sarveswaran

gowtham0992@gmail.com | 720-209-2305 | Denver, CO
linkedin.com/in/gsarveswaran | github.com/gowtham0992

PROFESSIONAL SUMMARY

I build cloud platforms that scale themselves, and I build the eval infrastructure that keeps the models running on top of them honest. 10 years owning production infrastructure for high-traffic distributed systems, turning messy manual operations into automation that means I don't get paged at 3am. I lead Cloud Foundations work across AWS multi-account orgs, EKS/Kubernetes, Terraform/Terragrunt IaC, networking, IAM, observability, and incident response, with Python automation throughout. On nights and weekends I ship open-source LLM eval and agent infrastructure, the same class of tooling that gates model and prompt changes before they reach production.

TECHNICAL SKILLS

Languages / Eval Python (primary automation + tooling), Bash. Prompt regression testing, LLM eval pipelines, agent-first infrastructure, MCP servers.

AWS / EKS AWS Organizations, multi-account boundaries, IAM and permission management, VPC/networking, DNS/Cloudflare, multi-region AWS, EKS, ECS, Lambda, CloudWatch, RDS Aurora, Karpenter autoscaling, Cilium networking, Linux, Docker, EC2/VMs.

Identity / IT SSO, AWS Cognito, Microsoft Entra ID, SCIM provisioning, Tailscale mesh VPN. Org-wide access management, onboarding/offboarding automation.

IaC / GitOps Terraform, Terragrunt, Ansible, Helm, ArgoCD, GitLab CI/CD, GitHub Actions. Module design, state management, plan-on-PR and apply-on-merge delivery.

Reliability / Obs Prometheus, Grafana, Datadog, Splunk, ELK. Alerting/logging foundations, SLO/SLA management, incident response, on-call ownership, backups/restores, self-healing automation.

GPU / FinOps 8xH100 cluster management and training throughput, GPU scheduling. Cloud cost optimization, rightsizing, spot strategies, cost reporting.

PROFESSIONAL EXPERIENCE

Public Cloud Engineer | Spectrum *Mar 2025 – Present*

- Lead cloud foundation initiatives across AWS, Kubernetes, observability, automation, and cost optimization. Guide an informal team of 6. I pick up whatever needs doing, from architecture decisions to writing Terraform modules to reviewing PRs.
- Designed and shipped production observability using Prometheus, Grafana, Datadog, and Splunk. Cut mean-time-to-detection on failures across distributed systems by getting alert quality right instead of just adding more alerts.
- Improved our EKS operating posture end-to-end. Autoscaling with Karpenter, node pressure reduction, safer rollout patterns, and deployment reliability. The clusters handle traffic spikes now without anyone losing sleep.
- Reduced cloud spend by 45% through automated rightsizing, spot instance strategies, and cost visibility dashboards that let engineering teams make their own informed decisions.
- Built automation-first workflows in Python that eliminated repetitive toil from deploys, health checks, and infrastructure maintenance. The goal is always to make the system handle it, not a person.
- Own identity and access management across the org. SSO integrations with AWS Cognito and Microsoft Entra ID, SCIM-based user provisioning and deprovisioning, and access reviews that keep onboarding and offboarding clean across AWS, GitHub, and SaaS.
- Own on-call for cloud infrastructure. Every page becomes a post-incident automation task, not just a resolved ticket.

Cloud Systems Engineer IV | Spectrum *Aug 2021 – Feb 2025*

- Scaled Kubernetes and Docker clusters through high-traffic spikes, deployment failures, and Linux-level reliability issues. Debugged node pressure, networking, and rollout problems across thousands of pods.
- Designed the Terraform and Ansible delivery patterns that became the standard across multi-cloud environments. Reduced manual provisioning to near-zero with repeatable, auditable modules.
- Drove ArgoCD-based GitOps adoption for deployment orchestration. Plan-on-PR, apply-on-merge, with drift detection and automated rollbacks. Took deploys from nerve-wracking to boring.
- Built proactive monitoring and health-check automation that caught failure modes before customers noticed. Reduced incident frequency quarter-over-quarter.
- Owned infrastructure projects end-to-end. Requirements, implementation, rollout, and operational handoff across teams. I was the engineer and the project manager, which meant things actually shipped.
- Stood up Tailscale mesh networking for secure access to internal services and infrastructure across environments, cutting reliance on brittle VPN setups.

- Built deploy safety, backup/restore, and capacity planning for stateful workloads on RDS Aurora. The kind of work that makes 3am pages go away permanently.

Cloud Systems Engineer II & III | Spectrum*Jul 2016 – Aug 2021*

- Wrote Python and Bash automation for deployment orchestration, operational workflows, and CI/CD pipelines using Helm and GitLab. If we were doing it manually more than twice, I automated it.
- Ran containerized workloads on ECS before the team standardized on EKS, and wrote Lambda functions for event-driven automation and operational tasks.
- Built monitoring and alerting on CloudWatch to catch failure modes early across AWS services.
- Created reusable infrastructure practices and AWS training that helped other teams improve security, reliability, and delivery consistency.
- Evaluated emerging platform tooling through hands-on proofs of concept, translating messy operational needs into practical patterns the team could adopt.

SELECTED BUILDER PROJECTS

Redline Prompt Regression Detection Platform github.com/gowtham0992/redline

- Built an open-source CI framework that generates eval suites from prompt logs, replays changed prompts, and gates model/prompt changes before they hit production. The same problem frontier eval tooling exists to solve.
- Ships as a PyPI package, GitHub Action, and MCP server on the official MCP Registry. Built for agent-first infrastructure where AI agents are part of the stack.

OpenAI Parameter Golf Challenge Accepted Non-Record Submissions github.com/openai/parameter-golf

- Optimized training throughput on 8xH100 GPU clusters, working through resource contention, scheduling pressure, and hardware constraints at scale.

Jawbreaker Hugging Face Build Small Hackathon github.com/gowtham0992/jawbreaker

- Fine-tuned MiniCPM-1B with a custom LoRA for scam detection. 632-case eval suite with zero dangerous misses. Built for a friend's grandmother who'd been hit by scam messages.
- Ships as a live Gradio Space, published model, and open dataset. No external LLM APIs, fully local inference.

Link Production Implementation of Karpathy's LLM Wiki github.com/gowtham0992/link

- Built and released a production MCP server for local knowledge-graph memory, with durable storage, validation workflows, and registry distribution.

Picochat Honest SLM Training Factory github.com/gowtham0992/picochat

- Built an end-to-end SLM training factory with preflight and release gates, contamination checks, and 8xH100 runbooks. I wanted to understand the full training pipeline, so I built one.

CERTIFICATIONS, EDUCATION & RECOGNITION

Certifications: AWS SysOps Certified; Certified Kubernetes Administrator (CKA).

Education: MS, Computer Engineering, University of Colorado Denver; MBA, Global Leadership, Colorado Technical University.

Recognition: Spectrum Technical Leadership Development Program (2023); Spectrum Superstar Award (2019); Spectrum Hackathon (2026).