

Gowtham Gopalakrishnan

gowthamgopalakrishnan126@gmail.com | +1 (520) 479-6876 | Santa Clara, CA
github.com/Gowtham-18 | linkedin.com/in/gowthamgopalakrishnan | gowthamg.info

SUMMARY

Staff Software Engineer specializing in cloud-native ML infrastructure and AI security, with 2+ years building production systems across model serving, document intelligence, and enterprise automation. Currently building a multi-region AI Model Serving Platform on GCP/Kubernetes at Palo Alto Networks — inference services, CI/CD, testing, and observability. Strong in Python, GCP (GKE, Cloud Run, GCS), GitLab CI/CD, and Prometheus/Grafana/Loki monitoring.

EXPERIENCE

Staff Software Engineer – AI Runtime Security

March 2026 – Present

Palo Alto Networks

Santa Clara, CA

- Develop and extend a unified pre/post-processor **SDK framework** powering **10+ security detection microservices** (prompt injection, toxic content, code detection, topic guardrails) on a GCP/GKE Model Serving Platform across **4 global regions**, reducing new service development time by **70%** via standardized patterns for Triton, vLLM, and Vertex AI integration.
- Designed a **pluggable GitLab CI/CD test pipeline** using **Cloud Run Jobs** to reach VPC-internal detection services without public endpoints, enabling **config-only onboarding** of new detection services and persisting results to **GCS**.
- Contribute to a high-performance inference pipeline (GPU-accelerated Triton, vLLM, PyO3 Rust tokenizers) delivering **sub-100ms p95 latency** with mTLS security and Redis caching for production security-content analysis.
- Operate multi-region deployment infrastructure across dev/staging/prod using **Helm, Docker, and GitLab automation**, with pre-commit security gates (Bandit SAST, Ruff) contributing to a **99.9% deployment success rate**.
- Own service observability and incident response with **Prometheus, Grafana, and Loki** dashboards, alerting based on-call workflows for production ML services.

Machine Learning Engineer Intern

September 2025 – December 2025

OneData Software Solutions

Fort Mill, SC

- Built a source-agnostic knowledge-base ingestion and vector indexing pipeline that transformed private GitHub and OneDrive documents into searchable embeddings, **enabling LLM-based retrieval over proprietary data**.
- Implemented an end-to-end workflow (ingest → extract → chunk → embed → index) using Python, OpenAI embeddings, and PostgreSQL + pgvector, **reducing manual document lookup time by ~80%**.
- Designed a modular ingestion architecture with deterministic chunking, metadata enrichment, and batched vector inserts, improving semantic retrieval accuracy and enabling scalable indexing of thousands of document chunks.

Associate Software Engineer

July 2022 – April 2023

Antworks

Chennai, India

- Developed features for the company's **Intelligent Document Processing platform** in Python, improving extraction accuracy from **70% to 90%** and reducing processing time by **25%**.
- Built event-driven document processing flows with **Kafka** and authored complex **SQL** queries powering classification, entity recognition, and reporting workloads, improving throughput by **30%**.

Cognitive Software Development Intern

February 2022 – June 2022

Antworks

Chennai, India

- Built and configured client automation workflows on **QueenBot**, the company's RPA/automation platform, translating business requirements into production workflow bots.
- Partnered with the support team to debug and resolve client workflow issues, writing **SQL** queries for data validation, investigation, and reporting.

EDUCATION

University of Arizona, Tucson – MS in Data Science

January 2024 – December 2025

Sri Krishna College of Engineering and Technology, India – BTech in Information Technology

September 2018 – April 2022

SKILLS

Languages – Python, SQL, Bash

ML Serving & AI Infrastructure – Triton Inference Server, vLLM, Vertex AI, PyTorch, LLM inference & guardrails, RAG

Cloud (GCP) – GKE, Cloud Run, Cloud Storage (GCS), BigQuery, Artifact Registry, IAM & service accounts

Containers & Deployment – Kubernetes, Docker, Helm, multi-region deployment

CI/CD, Testing & Automation – GitLab CI/CD, pytest, Docker-based e2e & smoke testing, pre-commit hooks, go-task

Observability & Reliability – Prometheus, Grafana, Loki, structured logging, on-call/incident response

Backend & Data – FastAPI, gRPC, Pydantic, Kafka, Redis, PostgreSQL/pgvector, MySQL

ACHIEVEMENTS

Karnataka State Police Hackathon 2023 – Winner | DevBhoomi Cyber Hackathon 2022 – Winner | AfroTech AI Hackathon 2024 (USA) – Runner-up | Darkathon 2022 (Narcotics Control Bureau) – Runner-up