

SUMMARY

Software Engineer building **production GenAI systems** end to end: **QLoRA fine-tuning** and **vLLM** serving of open-weight LLMs, **knowledge-graph RAG**, and **LangGraph** agentic pipelines exposed as **MCP** services on AWS, GCP, and SAP BTP. Shipped systems cutting inference cost **10x** and p95 latency **90s to under 11s** for **5,000+ internal users**.

EDUCATION

Binghamton University, State University of New York

Master of Science in Computer Science

Aug 2023 – Dec 2025

Binghamton, NY

School of Engineering, MIT ADT

Bachelor of Technology in Computer Science

Aug 2019 – May 2023

Pune, India

EXPERIENCE

SAP Labs US

Software Developer

Jan 2026 – Present

Palo Alto, CA

- Replaced a **3-hop GPT** pipeline with two **QLoRA** adapters on **DeepSeek-R1 14B**: **98%** accuracy at **10x** lower cost.
- Cut p95 latency from **90s to under 11s** via single-task adapter isolation and **vLLM PagedAttention**; **3x** throughput.
- Closed the retraining loop with a **HANA** accept/reject store feeding in-training **ORPO** preference tuning, **S3**-versioned and GitOps-shipped (**Docker**, **ArgoCD**, **KServe**, **SAP BTP**) across **10+** model versions.
- Architected a **knowledge-graph RAG** pipeline exposing it as a production **MCP** server (SSE, JSON-RPC 2.0, PKCE, OpenTelemetry), cutting query time **92%** (**1hr to 5min**) for **5,000+** content developers globally.
- Collapsed a **200+** artifact cascade across 2 systems of record into 2 **FastMCP/SSE** calls: **1.5 days to 5 min**.
- Parallelized writes over **10** entity-locked threads; converted a multi-minute blocking push to a polling job.
- Built a **5-hop OAuth** principal-propagation chain with session, user, and tool-level audits on agent writes.

Michigan Health Information Network

Software Engineering Intern

Mar 2025 – Jan 2026

Remote, USA

- Mapped physician notes to EHR schemas via **Vertex AI RAG** with federated **FHIR SPARQL** resolution: **4 hrs to 10 min**.
- Lifted recall@20 from **61% to 88%** over **48K** records via free-text-only embeddings, demographics as metadata filters, and a **Neo4j** pre-filter pruning **87%** of vector search space.
- Lifted diagnostic recall from **79% to 91%** via majority voting over parallel **MedLM** runs; **FastAPI/React** event stream.

PROJECTS

Agentic Tool Orchestrator | *FastAPI, OpenAI, Python*

- Built **agentic** orchestrator on the **OpenAI API** routing queries across **191 tools** in **7 services** via a two-stage hierarchical router, cutting per-call schema payload to **30** schemas, context **40%** (37k to 22k), instruction-following: **50% to 100%**.
- Engineered a fail-open loop with parallel execution, identical-failure circuit breakers, budget-bounded turns, and clarify-before-mutate gating, capping any misroute at one round trip, validated by a **24-test** end-state suite.

GraphGuard – Resilient Agent & Governance Framework | *LangGraph, Pydantic, Python, OpenAI*

- Created a **LangGraph** multi-agent pipeline with 5-classifier consensus and conflict escalation, precision: **84% to 96%**.
- Cut malformed LLM outputs **90%** via **Pydantic** schema validation and self-healing retry loops at every graph node.

Industry-Specific Layoff Tracker | *Python, Flask, MongoDB, Faktory, NLTK, REST APIs*

- Ingested and processed **208K+ Reddit & 4chan posts/month** into a real-time sentiment analysis pipeline.
- Built **Python/Flask** APIs with sentiment/toxicity scoring (**98% accuracy**) and Plotly dashboard visualizations.

Tuberculosis Detection via Transfer Learning | *TensorFlow, ResNet-50, Python*

- Trained ResNet-50 on TBX11K with CLAHE preprocessing and augmentation, achieving **92% accuracy**; validated via Grad-CAM, F1, and AUC. Co-authored an **IEEE-published** paper synthesizing **60+** studies into a medical AI pipeline.

TECHNICAL SKILLS

Languages Python, TypeScript / JavaScript, SQL, SPARQL

LLM & Agents LLM fine-tuning (QLoRA / PEFT / TRL), ORPO / DPO, Hugging Face, PyTorch, vLLM, LangGraph, RAG, MCP, Knowledge Graphs, Evals (RAGAS / LangSmith)

Backend & Data FastAPI, Node, React, PostgreSQL, SAP HANA, Neo4j, Pinecone, MongoDB, Pandas / NumPy

Cloud & Infra GCP (Vertex AI, BigQuery, Cloud Run, Pub/Sub, Vector Search), AWS, Docker, Kubernetes, KServe, ArgoCD, Terraform, OpenTelemetry, Linux, Git