

Rohan Waghmare

Palo Alto, CA | ✉ waghmarerohan30@gmail.com | ☎ 607-245-6001 | 🌐 rohanwaghmare.com | 📄 /rohanwaghmare

SUMMARY

Software Developer with 1+ years of experience building **production GenAI systems** from **agentic pipelines** and **RAG architectures** to **LLM fine-tuning** and **MLOps** on GCP, AWS, and SAP BTP. **Open to relocation; available immediately.**

WORK EXPERIENCE

SAP

Software Developer iXP (AI)

Jan 2026 – Present

Palo Alto, CA

- Fine-tuned **DeepSeek-14B** with **dual QLoRA adapters** on **SAP BTP** via KServe + FastAPI; **98% accuracy, p95 < 11s**.
- Replaced a multi-hop **GPT-4** pipeline at **10x lower cost** via **MLOps** loop (Docker, ArgoCD, S3) and dynamic **max_new_tokens** re-inference; cut pipeline time from **1.5 min to under 30s**.
- Built a **RAG pipeline** over a Knowledge Graph with confidence-scored retrieval; cut query-to-docs time by **92%** (1hr+ to under 5 mins) for **5,000+ content developers**. Exposed the system as a **MCP server** (SSE, JSON-RPC 2.0).
- Engineered a multi-step **agentic** pipeline (LangGraph / GPT-4) on **BTP** for document extraction and backend execution.
- Extended to a production **MCP server** with **PKCE + SAP IDP** auth, audit logging, and tool-call agent observability.

Michigan Health Information Network

Software Engineer

Mar 2025 – Jan 2026

Binghamton, NY (Remote)

- Built a Python data pipeline processing **500K+ records** across **10+ orgs** on AWS (ECS Fargate, S3, EventBridge, DynamoDB) with Docker/Terraform and cross-month aggregation.
- Cut processing latency **40%** via S3 Select API with conditional ScanRange and **multipart parallelization** (ThreadPoolExecutor, **20MB chunks**) across multi-GB files.

PROJECTS

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 full medical AI pipeline.

EDUCATION

Binghamton University, State University of New York

Master of Science in Computer Science

Aug 2023 – Dec 2025

School of Engineering, MIT ADT

Bachelor of Technology in Computer Science

Aug 2019 – May 2023

TECHNICAL SKILLS

- **Languages:** Python, JavaScript (React, Node, Express), C, SQL
- **AI/ML:** LLM Fine-tuning (QLoRA), HuggingFace, MCP, LangGraph, Agentic, TensorFlow, Scikit-learn, RAG, OpenCV
- **Cloud:** GCP (Vertex AI, BigQuery, Cloud Run, Pub/Sub, Vector Search), AWS, SAP BTP/AI Core, AWS CCP
- **Backend & APIs:** FastAPI, Flask, Django, REST APIs, MCP (SSE, JSON-RPC 2.0), SPARQL, RDF
- **Databases:** PostgreSQL, SAP HANA, MongoDB, Vector DB (Pinecone)
- **Tools:** Git, Linux/Unix, Docker, Terraform, Postman/Bruno, Pandas, NumPy, Grafana, Agile/SCRUM