VIKHYAT CHAUHAN

+1-201-616-6373 | vikhyat.chauhan@gmail.com | LinkedIn | GitHub | vikhyatchauhan.com

SKILLS

- **AI & LLM Engineering:** RAG, Agentic AI, Fine-Tuning (LoRA/QLoRA), LangChain, LangGraph, OpenAI/Anthropic APIs, hybrid search, cross-encoder reranking, embeddings, Pinecone, FAISS, LLM-as-judge, LangSmith, Evidently, OpenTelemetry, Prompt Engineering
- **ML & Infra:** PyTorch, TensorFlow, TensorRT, HuggingFace, Transformers, scikit-learn, FastAPI, GCP (Vertex AI, GKE, Cloud Run), AWS (EC2, S3, EKS), Kubernetes, Docker, CI/CD (Jenkins), MongoDB
- **Languages & Systems:** Python, C++, Java, TypeScript, SQL, Linux, Git, CUDA

PROJECTS

Professional RAG Pipeline | Professional RAG **|** ***GCP, Claude 3.5, Pinecone, FastAPI, Cross-Encoder***

- Two-Stage Retrieval: Engineered a high-precision pipeline utilizing BGE-base embeddings for vector search, followed by an MS-MARCO Cross-Encoder reranker to maximize context relevance for long-form synthesis.
- Production Guardrails: Built an automated LLM-as-judge evaluator to score response faithfulness; implemented SHA-256 query hashing for semantic caching, reducing LLM API overhead by 30%.
- Observability: Deployed as a containerized microservice on Google Cloud Run with per-query token/cost tracking, latency instrumentation by stage, and grounded generation logic to eliminate model hallucinations. Built from scratch - no LangChain or LlamaIndex.

Physiological Stress Classifier | Physiological Stress Classifier **|** ***PyTorch, Conv1D Autoencoders, SVM, Scikit-learn***

- Unsupervised Feature Extraction: Developed a deep learning pipeline using Conv1D Autoencoders to compress high-dimensional physiological data (BVP, EDA, EMG, TEMP) from wearable sensors.
- Multimodal Fusion: Engineered a classification system on the WESAD dataset, achieving 83.5% accuracy in three-class emotion detection across 15 subjects.
- Signal Processing: Implemented robust preprocessing and normalization for heterogeneous sensor sampling rates, transforming raw bio-signals into actionable features for real-time stress monitoring.

PROFESSIONAL EXPERIENCE

Virginia Polytechnic Institute and State University **2024 - Present**
Graduate AI/ML Researcher - NSF-Funded *Blacksburg, VA*

- Architected a LangGraph + Llama 3 multi-agent system for grammar-constrained SDF XML generation, automating 100+ simulation environments and eliminating 40 hours/week of manual authoring.
- Designed a deadline-aware Mixture of Experts (MoE) planner for UAVs with hardware-aware arbitration across heterogeneous TPUs, reducing mission latency and energy consumption by 13%.

GE HealthCare **2022 - 2024**
Software Engineer II *Milwaukee, WI*

- Led the technical architecture migration from a legacy C++ monolith to Kubernetes-based Java/Spring Boot services, reducing MRI application runtime by 80%.
- Developed an MRI report generation system (MedCLIP, PyTorch, TensorRT) fusing scan sequences with patient history, outperforming SOTA accuracy by 6.5% across clinical datasets.
- Spearheaded TensorRT optimization of Transformer models (CUDA/C++), increasing real-time segmentation throughput by 20% on FDA-regulated hardware.
- Introduced automated quality gates and CI/CD pipelines (Jenkins, SonarQube) that reduced release defects by 60% while maintaining zero critical patient-safety incidents.

EDUCATION

Virginia Polytechnic Institute and State University **Aug 2024 - May 2026**
M.S., Computer Engineering (Thesis Track) (GPA: 4.0)

- **Achievements:** NSF Grant Recipient

Uttarakhand Technical University **Jul 2016 - Dec 2020**
B.E., Electronics and Communication Engineering (GPA: 3.8)

- **Achievements:** First Division

AWARDS & RECOGNITION

- **Patent Holder:** Master-Slave Microcontroller Communication System for Home Automation
- **GE Impact Award:** Recognized for exceptional engineering contribution to clinical imaging systems
- **NSF Graduate Research Grant:** National Science Foundation funding for AI / autonomous systems research
- **GCP Professional ML Engineer:** Expected (2026)