VIKHYAT CHAUHAN

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

SUMMARY

Production AI engineer who ships — TensorRT inference pipelines on FDA-regulated hardware, ROS 2 simulation stacks, and LangGraph agentic systems, all with measurable results. I close the gap between research and deployment: I'll take your hardest AI/ML bottleneck from prototype to production-grade, quantified, and compliant.

SKILLS

- **AI & LLM Engineering**: RAG (two-stage retrieval, BGE embeddings, cross-encoder reranking), LangGraph, LangChain, LlamaIndex, LLM-as-judge, agentic workflows, semantic caching, OpenAI APIs, Llama 3, Qwen, FAISS
- **ML & Inference**: PyTorch, TensorFlow, TensorRT (ONNX export, FP16 precision tuning, custom CUDA/C++ plugins, layer fusion), HuggingFace Transformers, scikit-learn, MedCLIP, Conv1D Autoencoders
- **Cloud & Infrastructure**: GCP (Cloud Run, GKE, Vertex AI), AWS (EC2, Lambda, S3, EKS), Kubernetes, Docker, FastAPI, Spring Boot, CI/CD (Jenkins, SonarQube), MongoDB, RabbitMQ, Kafka, MQTT, Microservices, REST
- **Languages & Systems**: Python, C++, TypeScript, SQL, CUDA, Linux, Git

PROFESSIONAL EXPERIENCE

Virginia Polytechnic Institute and State University **2024 - 2026**
Graduate AI/ML Researcher *Blacksburg, VA*

- Designed and experimentally validated a brain-inspired UAV navigation system achieving 34.6% reduction in compute energy and elapsed time over baseline ($p < 0.0001$, n=1,000 runs, 3 real-world-mapped environments) - NSF-funded
- 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.

GE HealthCare **2022 - 2024**
Software Engineer II *Bengaluru, IN*

- Led the technical architecture migration from a legacy C++ monolith to Kubernetes-based Java/Spring Boot services, reducing MRI application runtime by 80%, by designing and implementing image normalization and segmentation algorithms in C++.
- Profiled and optimized a Swin Transformer segmentation model end-to-end - ONNX export, FP16 precision tuning, custom CUDA/C++ plugin authoring for unsupported ops, and layer fusion via TensorRT builder API - achieving 20% throughput uplift on FDA-regulated hardware enabling real-time intraoperative use.
- Developed a multimodal MRI report generation system (MedCLIP, PyTorch, TensorRT) fusing scan sequences with patient history, achieving quantifiable impact with a 6.5% improvement in diagnostic classification F1 over baselines.
- Introduced automated quality gates and CI/CD pipelines (Jenkins, SonarQube) for production deployment and optimization, reducing release defects by 60% while maintaining 0 patient-safety regressions across 12 quarterly releases.

TNM Electronics **2020 - 2022**
Software Engineer I *Bengaluru, IN*

- Owned the full platform from day one; architected an AWS-backed microservice backend (EC2, Lambda, RabbitMQ, MongoDB, MQTT broker), scaling from 0 to 1000+ devices while sustaining 99.9% uptime.
- Conceived and patented a master-slave microcontroller communication system for home automation, cutting final product BOM cost 80% and directly enabling first design-win customers.
- Reduced field bug reports 98% vs. prior manual integration method by designing a structured hardware-software validation workflow as the company's sole engineer.

PROJECTS

Professional RAG Pipeline

- Built a two-stage retrieval pipeline (BGE-base embeddings → MS-MARCO Cross-Encoder reranker) evaluated with Hit@K and MRR metrics; LLM-as-judge scoring showed consistent faithfulness gains over single-stage retrieval across 50+ test queries - implemented from scratch without LangChain.
- Implemented SHA-256 query hashing for semantic caching, reducing LLM API overhead by 30%; added per-query token/cost tracking for full inference observability.
- Deployed as a containerized microservice on Google Cloud Run with latency instrumentation by stage and grounded generation logic to eliminate model hallucinations.

Web Automation Agent

- Built a multi-step agentic browser automation system using LangGraph orchestration and Browser Use for DOM interaction, powered by a locally-quantized Qwen model on RTX 5060 Ti - no cloud API dependency.

- Designed novel agentic workflows with stateful task graphs, retry logic, conditional branching, and form-filling across multi-page web workflows.
- Achieved ~60s end-to-end task completion with sub-300ms first-token latency on local inference.

Physiological Stress Classifier

- Developed a Conv1D Autoencoder pipeline to compress high-dimensional physiological data (BVP, EDA, EMG, TEMP) from wearable sensors into compact unsupervised representations.
- Engineered a multimodal classification system on the WESAD dataset, achieving 83.5% accuracy in three-class emotion detection across 15 subjects.
- Implemented preprocessing and normalization for heterogeneous sensor sampling rates, transforming raw bio-signals into features for real-time stress monitoring.

CANavigator — Brain-Inspired Drone Navigation Framework

- Architected a modular ROS 2 + Gazebo Harmonic simulation stack with procedural city-grid and Perlin-noise NFZ generators, deterministic seeded RNG for bit-exact replay, and a physics-accurate DJI FlyCart 30 dynamics model (2nd-order actuator latency, aerodynamic drag, Ornstein-Uhlenbeck wind gusts).
- Implemented a dual energy model tracking CPU computational cost alongside motion energy (208.9 J/m EPM baseline), with per-strategy CSV telemetry across 8 metrics including compute latency, zone violations, and event deadline miss rates.
- Designed a fair head-to-head benchmarking protocol — all four planners (APE1/APE2/APE3/CA) navigate identical procedurally generated arenas with shared event sequences, with only runs where all strategies reach the target counted toward the 1,000-run corpus.

EDUCATION

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

- **Achievements:** Defended M.S. Thesis on Brain Inspired Drone Navigation System using MISD Architecture 4.0 GPA
- **Coursework:** CS5424: Advanced Machine Learning, CS5465: Applications of Machine Learning, ECE5504: Computer Architecture

Uttarakhand Technical University **Jul 2016 - Dec 2020**
B.E., Electronics and Communication Engineering *Uttarakhand, IN*

- **Achievements:** Graduated with First Division

AWARDS & RECOGNITION

- **Patent Co-Inventor:** Master-Slave Microcontroller Communication System for Home Automation, TNM Electronics
- **GE Impact Award:** Recognized for independently modernizing a critical legacy MR imaging application to production microservices, directly improving clinical software delivery at scale
- **NSF-Funded Graduate Research Assistant:** Supported on NSF Award #2204780 (CCF: Small: Paradox and Brain-inspired Computer Architecture, PI: JoAnn M. Paul); contributed significantly to thesis research on brain-inspired autonomous systems
- **Honor Society of Phi Kappa Phi:** Inducted, Virginia Tech Chapter (2026); top 10% of Master's graduate students across all disciplines
- **Omicron Delta Kappa National Leadership Honor Society:** Selected for membership, Alpha Omicron Circle, Virginia Tech (2026 Cohort); recognizes achievement in academics, research, and service