

SIMRAN KC

Huntsville, AL | 256-936-0442 | simran.kc.official@gmail.com | [GitHub](#) | [simrankc.com.np](#) | [LinkedIn](#)

SUMMARY

ML engineer with **3+ years** building and serving production ML systems, from fine-tuning and quantization to **low-latency inference APIs**. Passionate about **GPU performance** and systems-level optimization.

TECHNICAL SKILLS

Languages: Python, C++, CUDA, SQL

GPU & Performance: CUDA kernels, kernel profiling (Nsight Compute/Systems), roofline analysis, flash attention, quantization (NF4, BitsAndBytes), inference optimization

ML: PyTorch, Hugging Face Transformers, LoRA/QLoRA, Sentence Transformers, conformal prediction

Serving & Infra: FastAPI, Redis, Docker, AWS (ECS, SQS, S3, Lambda), Qdrant, OpenSearch, Ollama, CI/CD, Linux

EXPERIENCE

Graduate Research Assistant | NASA IMPACT

Huntsville, AL | Aug 2024 - Present

- Owned end-to-end design and deployment of a scalable **multi-model inference service** (Python, FastAPI) serving the team's ML models behind a single endpoint; added Redis-based distributed task queuing, cutting API response latency by **40%**
- Cut LLM guardrails cost **~80%** by running **IBM Granite Guardian** locally on Ollama as a first-pass filter, escalating only flagged content to a heavier dynamic Risk Agent
- Fine-tuned classifiers (Indus, ModernBERT, BERT, FastText) on scientific corpora, hitting **98% precision** on multi-label tagging and **96% recall** on indexing; used conformal prediction ($\alpha = 0.05$) to drop low-confidence predictions
- Built the **orchestration layer** for NASA's Deep Research Agent: LLM workflow orchestration, **MCP tool servers** with FastMCP, and a benchmark harness for A/B testing agent configurations on synthetic datasets

Machine Learning Engineer | MutualArt

Remote (London, UK) | Sep 2023 - Jun 2024

- Shipped automated artwork description generation end to end: fine-tuned **IDEFICS-2** (8B vision-language model) with LoRA, **4-bit NF4 quantization**, and flash attention (**~75% less VRAM**), served from GPU workers consuming AWS SQS events with BLIP validation at **99.5% uptime**
- Built real-time artwork recommendations: multimodal embedding pipeline (424-dim fused features) on AWS ECS processing **80,000+ artworks weekly**, with **Qdrant** ANN retrieval blended with OpenSearch via Reciprocal Rank Fusion and a serving layer that scales independently of embedding compute
- Built a parallel image-processing pipeline for **duplicate detection** (two-stage ORB + SSIM matching, Union-Find grouping), improving customer conversion by **200%**

Associate Software Engineer | Sireto Technology

Lalitpur, Nepal | Sep 2022 - Sep 2023

- Led development of KuberIDE, a microservices-based web IDE (FastAPI, ReactJS) serving **500+ developers**; built OAuth 2.0 auth and a WebSocket proxy for language-server IntelliSense, increasing platform adoption by **300%**
- Set up **GitHub Actions CI/CD**, wrote Pytest suites (**80% coverage**), and Dockerized services with multi-stage builds

PROJECTS

CUDA Kernels Suite (WIP) (C++, CUDA, Nsight Compute/Systems)

- From-scratch GPU kernels (vector add, tiled matmul, reduction, softmax, attention) profiled with **Nsight Compute/Systems**, benchmarked vs **cuBLAS** and **PyTorch** with CUDA-event timing (warmup, median-of-N)

Production Lifecycle Management for LLM Classifiers (MS Thesis) (Python, PyTorch, HuggingFace, EvidentlyAI)

- MLOps framework for LLM classifiers in production: **drift detection** via statistical tests, **active learning** for sample selection, and incremental fine-tuning pipelines

EDUCATION

Master of Science in Computer Science | The University of Alabama in Huntsville

Expected Aug 2026 | GPA: 4.0/4.0 | Coursework: Advanced Operating Systems, Distributed Systems, Cloud Computing, Data Structures & Algorithms

Bachelor of Computer Engineering | Tribhuvan University, IOE Pulchowk Campus

2017 - 2022 | GPA: 3.7/4.0 | Full Undergraduate Scholarship (Top 1%)

AWARDS & CERTIFICATIONS

- 2nd Place, NASA IMPACT X-athon 2024:** NASA-sponsored innovation competition
- AWS Generative AI and AI Agents with Amazon Bedrock, Coursera (Feb 2026)
- CodePath Fellow (2024); Microsoft Learn Student Ambassador (2019-2021, selected from **10,000+ APAC applicants**)