Nirajan Paudel

paudelnirajan | paudelnirajan.github.io | nirajan.paudel@colorado.edu | +1 (720) 595-8207

Education

University of Colorado Boulder — Boulder, CO
Master of Science in Computer Science — Expected May 2027
Relevant Coursework: Data Center Scale Computing, Object-Oriented Analysis & Design

Institute of Engineering, Tribhuwan University — Pokhara, Nepal
Bachelor of Engineering in Electronics, Communication & Information Engineering — 2020–2024

Work Experience

Machine Learning Intern — March 2025 – May 2025
PalmMind Technology

- Architected production-grade agentic RAG chatbots for EV dealers and insurance clients using OpenAI GPT-4, migrating workflows from LangChain to LangGraph to enable advanced tool calling and conversational AI capabilities with orchestrator-worker patterns and parallel processing.
- Implemented LLM routing strategy for query classification and intelligent workflow distribution, improving system efficiency by 20% while reducing operational costs through optimized model selection.
- Optimized end-to-end retrieval pipeline by integrating web scraping, OCR (Tesseract), and Redis caching, reducing average query latency by 40% and enhancing context relevance for complex multi-turn conversations.
- Applied best practices from Anthropic's "Building Effective Agents" and LangChain documentation to design robust agentic systems with improved reliability, error handling, and conversational memory.

Teaching Assistant — Sep 2024 – Feb 2025
Pashchimanchal Campus, Tribhuwan University

- Instructed 60+ students in C programming fundamentals and Information Systems (cloud computing, distributed systems, neural networks, data mining), designing hands-on lab exercises and providing comprehensive code reviews.
- Mentored 15+ semester-end projects, resulting in 85% of students demonstrating proficiency in algorithm design and implementation.

Projects

Music Separation as a Service (MSaaS) — Oct 2025 – Nov 2025
University Course Project

- Architected scalable microservices on Google Kubernetes Engine (GKE) for AI-driven music source separation, processing MP3 uploads into four instrumental stems (vocals, drums, bass, other) via asynchronous pipeline with horizontal pod autoscaling.
- Migrated storage from self-hosted MinIO to Google Cloud Storage (GCS), achieving 99.99% durability and eliminating single points of failure while reducing operational overhead by 60%.
- Engineered Redis-based job queue decoupling Flask REST API from Demucs ML workers, reducing initial response time to 200ms for inference tasks averaging 2-5 minutes, enabling throughput of 100+ songs/hour at peak load.
- Implemented Kubernetes resource management (6Gi RAM limits) and dynamic JavaScript frontend with real-time polling, preventing OOM pod evictions and delivering seamless UX for long-running background tasks.

AI-Powered Question Assistant (IOE-GPT) — June 2025 – July 2025
Personal Project

- Built semantic search system for IOE programming questions using LangChain, Groq LLaMA-3, and Milvus vector database, achieving 92% retrieval accuracy with sentence-transformer embeddings (all-MiniLM-L6-v2).
- Designed FastAPI microservice with Redis state management and Docker containerization, serving 50+ queries with sub-second response times.

Voice RAG Assistant — Feb 2025 – March 2025
Personal Project

- Developed audio-based Q&A system integrating OpenAI Whisper for transcription and Groq LLaMA-3.1-8B for generation, implementing RAG with optimized chunking strategies for 85% context retrieval accuracy.
- Built FastAPI backend with HuggingFace embeddings and LangChain, processing 30-minute audio files with 10 second query response times.

Nepali Image Captioning — Aug 2023 – March 2024
Undergraduate Capstone Project

- Developed Transformer-based model for paragraph-length Nepali caption generation using Inception V3 features, trained on 21,150 image-caption pairs (Stanford dataset + 800 cultural heritage images), achieving BLEU-4: 0.59, outperforming LSTM baseline by 18%.
- Optimized architecture with 8 attention heads, 0.2 dropout, and custom tokenization for 14K vocabulary, deploying full-stack application (React, Flask) published in Journal of Soft Computing Paradigm (March 2024).

Technical Skills

Languages	Python, C/C++, Java, SQL, Bash
ML/AI Frameworks	PyTorch, TensorFlow, Scikit-learn, Keras, LangChain, LangGraph, HuggingFace Transformers, OpenAI API
ML Techniques	CNNs, RNNs, Transformers, RAG, NLP, Audio Processing, Computer Vision, Embeddings (Sentence-Transformers)
Cloud & DevOps	Google Cloud Platform (GKE, GCS), Kubernetes, Docker, Redis, Horizontal Autoscaling, Microservices
Tools & Databases	FastAPI, Flask, Git, Milvus, PostgreSQL, Selenium, Tesseract OCR, Jupyter, Pandas, NumPy

Publications

Subedi, N. et al. (Jan. 2024a). "Drowsiness and Crash Detection Mobile Application for Vehicle's Safety". In: *Journal of IoT in Social, Mobile, Analytics, and Cloud* 6.1, pp. 54–66. URL: https://doi.org/10.36548/jismac.2024.1.005.

– (Jan. 2024b). "Nepali Image Captioning: Generating Coherent Paragraph-Length Descriptions Using Transformer". In: *Journal of Soft Computing Paradigm* 6.1, pp. 70–84. URL: https://doi.org/10.36548/jscp.2024.1.006.