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 while reducing operational costs through optimized model selection between GPT-4 and lighter models.
- Optimized end-to-end retrieval pipeline by integrating web scraping, OCR (Tesseract), and Redis caching, reducing query latency 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 90+ 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

Zenco - AI-Powered Code Analysis Tool — Nov 2025 – Present
Open-Source Python Package (Published on PyPI)

- Developed production-ready CLI tool supporting 5 languages (Python, JavaScript, Java, Go, C++) with Tree-sitter AST parsing and multi-provider LLM integration (Groq/OpenAI/Anthropic/Gemini), automating code documentation and analysis workflows.
- Engineered execution priority optimization algorithm using dead code detection to skip unnecessary LLM API calls, reducing token consumption and operational costs while maintaining analysis quality.
- Refactored monolithic codebase using Strategy and Factory design patterns, reducing core module size by 70% (2,057 to 603 lines) and establishing modular architecture for extensible language support and maintainability.
- Published to PyPI with automated GitHub Actions CI/CD pipeline for multi-platform testing (Linux, macOS, Windows), linting, and versioning, implementing comprehensive error handling and mock testing capabilities for development.

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 significantly reducing operational overhead.
- Engineered Redis-based job queue decoupling Flask REST API from Demucs ML workers, enabling non-blocking API responses for inference tasks averaging 2-5 minutes and supporting concurrent processing of multiple songs.
- 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, Milvus vector database, and sentence-transformer embeddings (all-MiniLM-L6-v2), deployed as FastAPI microservice with Redis state management and Docker containerization for efficient query processing.

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, GitHub Actions, CI/CD, Microservices
Tools & Databases	FastAPI, Flask, Git, Milvus, PostgreSQL, Selenium, Tesseract OCR, Jupyter, Pandas, NumPy