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 — May 2027

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 agentic RAG chatbots for EV dealers and insurance clients using OpenAI GPT-4, migrating from LangChain to LangGraph to enable advanced tool calling and parallel processing with orchestrator-worker patterns.
- Implemented LLM routing strategy (GPT-4 for complex queries, GPT-3.5-turbo for classification), reducing API costs while maintaining quality through prompt engineering and Redis caching for 500+ customer inquiries.
- Optimized retrieval pipeline integrating web scraping, Tesseract OCR, and context management, improving accuracy for insurance policy and vehicle specification queries.

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

- Instructed 90+ students in C programming and Information Systems, designing lab exercises covering cloud computing, distributed systems, neural networks, and data mining.
- Mentored 15+ semester projects, achieving 85% student proficiency in algorithm design and implementation.

Projects

Serverless vs Kubernetes NLP Inference Benchmark — Oct 2025 – Dec 2025
University Course Project

- Architected comparative cloud infrastructure deploying DistilBERT sentiment analysis on AWS Lambda (serverless) and EKS (Kubernetes), evaluating latency, cost, and scalability under variable load conditions with automated Locust testing.
- Engineered full-stack deployment pipeline using Terraform IaC for VPC, EKS cluster, Lambda functions, and API Gateway, with GitHub Actions CI/CD pushing containerized services to ECR and orchestrating multi-environment deployments.
- Built Streamlit dashboard with FastAPI backend for real-time parallel inference comparison, implementing exponential backoff to handle Lambda cold starts (60s model loading) and integrating LLM-powered performance analysis for SRE insights.
- Optimized container orchestration with Kubernetes HPA (Horizontal Pod Autoscaler) and LoadBalancer services, managing resource constraints (6Gi RAM limits) while maintaining service reliability across 2-replica deployments.

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

- 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.
- Built official VS Code extension with automatic CLI installation, diff-view previews, and cross-platform compatibility, enabling in-editor code refactoring, docstring generation, and type hint insertion with visual change confirmation.
- 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.
- Published to PyPI and VS Code Marketplace with automated GitHub Actions CI/CD pipeline for pytest-based multi-platform testing (Linux, macOS, Windows), implementing comprehensive error handling and mock testing capabilities.

Tensor AI - Context-Aware Academic Assistant — Nov 2025 – Present
Personal Project

- Architected production-grade RAG pipeline using FastAPI and Pinecone vector database, processing 10+ engineering programs with multimodal ingestion through LlamaParse for legacy PDFs containing technical diagrams and tables.
- Implemented agentic query routing with intent classification and cross-encoder reranking (ms-marco-MiniLM-L-6-v2), achieving 95% retrieval relevance through dynamic top_k scaling (3-50 chunks) based on query complexity.
- Optimized system performance with 7-layer Redis caching strategy for embeddings (7-day TTL) and LLM classifications (24-hour TTL), reducing API costs by 70% and achieving sub-200ms initial response time using Server-Sent Events streaming.
- Built React frontend with real-time SSE streaming and automated semester detection based on Nepali calendar (B.S.) for personalized, context-aware academic support without manual user configuration.

Technical Skills

Languages	Python, C/C++, Java, SQL, Bash, JavaScript/TypeScript
ML/AI	PyTorch, TensorFlow, LangChain, LangGraph, HuggingFace, OpenAI/Anthropic APIs, RAG, Transformers
Cloud & DevOps	AWS (Lambda, EKS, EC2, ECR), GCP (GKE, GCS), Kubernetes, Docker, Terraform, GitHub Actions
Tools & Databases	FastAPI, Flask, Redis, Pinecone, PostgreSQL, Tree-sitter, Git, Pandas, NumPy