

Sumit Lasiwa

Dekocha, Bhaktapur, Nepal

+977 9811459601 — lasiwasumit@gmail.com

linkedin.com/in/sumit-lasiwa-30ab86346 — github.com/Sumitlasiwa — sumitlasiwa.com.np

Professional Summary Computer Engineering graduate with strong interest in Artificial Intelligence and Machine Learning. Experienced in developing AI applications, fine-tuning transformer models, and building scalable backend systems. Interested in applied AI research, low-resource NLP, and production-ready ML solutions.

Education B.E. in Computer Engineering 2022–2026 Khwopa College of Engineering, Libali, Bhaktapur

Experience

- **AI Backend Engineering Intern (Remote Cohort Internship)** July, 2026–Present
Developed a FastAPI-based CRUD application following layered architecture principles with separate route, service, and repository layers. Migrated application storage from in-memory data structures to SQLite and PostgreSQL databases. Containerized FastAPI and PostgreSQL services using Docker for consistent deployment and development workflows.

Technical Skills

- **Programming:** Python, C, C++
- **Machine Learning:** Deep Learning, NLP, Computer Vision, Transfer Learning, Fine-tuning
- **Frameworks:** PyTorch, Hugging Face Transformers, Scikit-learn
- **Data Tools:** NumPy, Pandas
- **Deployment:** FastAPI, Streamlit, Docker
- **AI Systems:** Retrieval Augmented Generation (RAG), LangChain, Vector Databases
- **Tools:** Git, GitHub, Weights Biases, SQL, Redis, MongoDB, PostgreSQL

Projects

- **RAG-Based Conversational AI with Interview Scheduling System** Developed a Retrieval-Augmented Generation (RAG) chatbot with an integrated interview booking system, enabling users to query documents and schedule interviews through a unified conversational interface. Built an end-to-end pipeline for document ingestion, chunking, embedding, and semantic retrieval using LangChain, Pinecone, and MongoDB. Improved response relevance and reduced hallucinations by implementing similarity thresholding and context-aware retrieval. Integrated Redis for session-based memory, enabling multi-turn conversations with persistent context. Deployed a scalable FastAPI backend on Render with support for Hugging Face and Gemini LLMs. **Tech:** Python, FastAPI, LangChain, Pinecone, MongoDB, Redis, Hugging Face, Gemini API
- **Nepali Grammar Error Correction using Transformer Models** Fine-tuned a multilingual T5 model for Nepali grammatical error correction using a synthetic parallel corpus and additional LLM-generated corrections. Fine-tuned a BERT reranker to further improve performance. Conducted experiments to optimize model quality and deployed the system using FastAPI. **Tech:** PyTorch, Hugging Face Transformers, FastAPI
- **Nepali Image Captioning with Attention Mechanism** Built an encoder-decoder image captioning model using EfficientNet-B0 as the encoder and an LSTM with attention as the decoder. Trained the model on a Nepali caption dataset derived from Flickr30k. **Tech:** PyTorch, Attention Mechanism, Computer Vision
- **RAG-Based YouTube Question Answering System** Developed a Retrieval-Augmented Generation system that allows users to query information from specific YouTube videos. Implemented document retrieval using vector embeddings and answer generation using LLM APIs. **Tech:** LangChain, ChromaDB, Vector Databases, LLM APIs

Certification Data Science Training – DeepMind Creation Pvt. Ltd

Languages English, Nepali, Newari, Hindi

Interests Artificial Intelligence Research, Machine Learning Systems, Low-Resource NLP