

Sulav Bhandari

Lalitpur, Nepal | cropped.sulav@gmail.com | sulavbhandari.com.np
linkedin.com/sulav-bhandari | github.com/cropped33

Machine learning engineer with over 1.5 years of experience in computer vision and NLP. Currently responsible for detection models in production and the retraining that keeps them accurate. Recent work includes retrieval-augmented generation pipelines, LoRA fine-tuning of open-weight LLMs, and a transformer written from scratch.

Technical Skills

AI/ML Concepts: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), LoRA/PEFT Fine-tuning, Multi-Agent Systems & Tool-Calling, NLP, Computer Vision

Frameworks & Retrieval: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, Scikit-learn, OpenCV, Keras, Pandas, NumPy, FAISS, sentence-transformers

Languages & Tools: Python, SQL (PostgreSQL, MySQL), C, C++, Git, REST APIs, FastAPI

Experience

AI Engineer | *Sandbox Pvt. Ltd., Pokhara, Nepal*

Jan 2025 – Present

- Built real-time computer-vision systems that read on-screen state in live game clients such as Rush Royale, League of Legends, and Tibia, training OpenCV and PyTorch detection models and retraining them when UI updates broke detection, letting runs go unattended for 8+ hours and replacing roughly 6 hours of manual play per client each day.
- Built an internal RAG assistant over game specs and internal docs that returns cited answers in seconds, cutting the time the team spent searching documentation by around 60%.
- Built a multi-agent system where separate LLM agents plan and carry out steps with tool-calling, running multi-step jobs end to end and cutting the manual steps per job by about 70%.

Projects

RAG-Based Document Q&A Assistant | *Python, LangChain, FAISS, Hugging Face, FastAPI*

- Built a retrieval-augmented generation pipeline that answers questions from a chosen set of documents rather than from model training data, which cuts down hallucinated answers.
- Split documents into chunks, embedded them with sentence-transformers into a FAISS index, and used LangChain to handle retrieval, prompt building, and generation with a Hugging Face model.
- Served it behind a FastAPI REST API that cites the source chunks behind every answer and picks up newly uploaded documents without a restart.
- Live demo:** qnadocs.up.railway.app **Code:** github.com/cropped33/rag-document-qa

Nepali Language LLM on a Native Dataset | *PyTorch, Custom BPE Tokenizer, Transformers*

- Wrote a Byte Pair Encoding (BPE) tokenizer for Nepali text and a decoder-only transformer from scratch, including the architecture, training loop, and sampling code.
- Evaluated on a held-out split with validation loss, perplexity, and BLEU, using them to settle vocabulary size and hyperparameters over several training runs.

Education

Tribhuvan University, IOE – Purwanchal Campus

Bachelor's in Computer Engineering

Dharan, Nepal

May 2021 – Aug 2025

Volunteer Work

Association of Computer Engineering Students (ACES) | *Secretary*

Mar 2023 – Mar 2024

- Organised ACES TechFest 6.0, covering hackathons, competitive programming, and datathon events.
- Vice Coordinator for the hackathon track, working with teams from across Nepal.

Awards

- Global Nominee, NASA Space Apps Hackathon
- Winner, ICT Quest 2.0 (Best Futuristic Implementation)
- Winner, 5-Minute GitHub Hackathon (GitHub Campus Expert session)
- Runner-up, ACES Hackathon Competition