

Shaswot Poudyal

Kathmandu, Nepal — ichbin84ssp@gmail.com — +977-9862578070 — GitHub

Education

Tribhuvan University – Thapathali Campus
Bachelor of Engineering in Computer Engineering

2022 – 2026

Technical Skills

Programming: Python, SQL

Machine Learning: PyTorch, Scikit-learn, XGBoost, Optuna, ASR, ONNX

LLM & AI Systems: LangChain, LangGraph, MCP, RAG, Vector Databases

Backend & Data: FastAPI, Pydantic, PostgreSQL, pgvector, Redis, Pandas, NumPy, MinIO, Alembic

MLOps & Observability: MLflow, OpenTelemetry, Arize Phoenix

Infrastructure: Docker, Git/GitHub, GitHub Actions

Experience

AI Engineering Intern – Alpinist Studios

May 2026 – Aug 2026

- Built LLM-powered applications and AI agents using Python, LangChain, LangGraph, MCP, Ollama, Retrieval-Augmented Generation (RAG), prompt engineering, and tool calling.
- Developed an AI HR Assistant and a finite-state-machine-based AI Interview Assistant for HR policy Q&A, workflow automation, multi-turn interview management, answer evaluation, and structured feedback generation.
- Engineered asynchronous AI services with Pydantic v2, FastAPI, Streamlit, Docker, structured logging, request validation, retries, observability, and graceful handling of LLM/API failures.
- Trained and evaluated deep learning models using PyTorch, Hugging Face Transformers, CNNs, LSTMs/GRUs, attention-based models, MLflow experiment tracking, and LoRA/QLoRA fine-tuning.
- Collaborated with engineers in an Agile workflow using Git, pull requests, code reviews, Jira, testing, documentation, and AI evaluation practices to deliver reliable AI features across multiple projects.

Conference and Publication

Rainfall Prediction in Kathmandu Using Machine Learning and Deep Learning

2026

International Conference on Engineering Technology and Infrastructure Development (ICETID 2026), January 2026

Shaswot Poudyal, Saurav Katwal, Rajad Shakya

Projects

AI HR Assistant & Automated ATS Platform (Multi-Agent Systems, Hybrid RAG, Information Retrieval, Distributed Systems)

- Architected a multi-agent system with LangGraph and FastAPI, using a supervisor to route requests across policy, leave, recruitment, clarification, and recap agents with Redis-based confirmation locks
- Built a production-grade hybrid RAG pipeline combining BM25, 768-dim pgvector embeddings, RRF fusion, and cross-encoder reranking with MinIO-based document ingestion
- Developed an automated ATS screening engine combining LLM evaluation with deterministic experience calculation, skill-taxonomy mapping, and weighted keyword matching
- Engineered reliable background processing using PostgreSQL Transactional Outbox and FOR UPDATE SKIP LOCKED; added OpenTelemetry/Arize Phoenix observability and 735+ automated tests

AI-Powered Technical Interview Assistant (Finite State Machines, LLM Evaluation, Deterministic Scoring, Multi-Provider LLM Routing)

- Architected an 8-state deterministic finite state machine (FSM) with server-authoritative timers and guarded transitions for reliable multi-turn interview sessions
- Engineered a two-stage LLM evaluation pipeline with rubric anchoring, score caps, and evidence-based scoring to reduce LLM score inflation

- Implemented multi-provider LLM routing using Ollama for local validation and OpenAI/Groq for complex evaluation, with retry and fallback mechanisms
- Built deterministic question distribution and scoring engines across 7 question types and 19 competency dimensions, with Pydantic scorecards, Plotly reports, and 100+ automated tests

Credit Risk Prediction & Decision Pipeline (Ensemble Learning, Imbalanced Classification, Feature Engineering, MLOps)

- Engineered an end-to-end credit risk prediction pipeline on 30K client records, reducing 24 raw attributes to 10 high-signal features and deriving cumulative delinquency features for improved risk modeling
- Designed a heterogeneous 4-model ensemble combining XGBoost, Random Forest, and PyTorch MLPs; probability averaging improved default recall from 0.35 to 0.630 with 0.781 AUC-ROC
- Addressed 78/22 class imbalance using Focal Loss and Optuna hyperparameter optimization, with 5-fold out-of-fold threshold calibration improving precision by 4.4% to 0.521
- Architected a production FastAPI + MLflow serving pipeline with Pydantic validation, real-time feature preprocessing, multi-model routing, experiment tracking, and a Streamlit/Plotly risk dashboard

Speech-to-Text & Language Modeling Pipeline (Automatic Speech Recognition, Multilingual NLP, RAG, Model Quantization)

- Built a dual-transformer ASR + language model system using PyTorch and NeMo
- Trained multilingual ASR (English/Nepali) with noise augmentation for real-world robustness
- Implemented RAG for context-aware outputs; reduced model size by ~70% via quantization
- Achieved 31.40% WER with an efficient ~13M-parameter multilingual ASR model

Kathmandu Rain Prediction (Time-Series Forecasting, Feature Engineering, Deep Learning)

- Developed an end-to-end rainfall forecasting pipeline using multi-year weather data
- Engineered temporal and geographic features; trained Bi-LSTM, GRU, RF, SVR, LightGBM
- Achieved best performance of ~3.79 RMSE

Document Reconstructor System (Computer Vision, OCR, Object Detection, Image Processing)

- Developed a computer vision pipeline to extract and reconstruct charts and tables from scanned documents using OCR and object detection.
- Designed image-processing and reconstruction algorithms to recover structured visual elements from document images.
- Reconstructed bar graphs and pie charts programmatically to preserve the structure and visual representation of extracted information.

Certifications and Trainings

Generative AI
Data Science
Agentic AI