

ASHISH ADHIKARI

Gurjadhara, Kathmandu, Nepal | adhsashish125@gmail.com | ashish-adhikari.com.np | +977 9846856199

AI Engineer specializing in fine-tuning open-source LLMs (LLaMA, Mistral) for Devanagari script and Nepali financial-document understanding using LoRA/QLoRA — currently leading QuickFox's migration of QuickRef, a RAG and agentic (LangGraph/ReACT) banking knowledge assistant, from a cloud LLM (Gemini) to a self-hosted fine-tuned model. Builds production AI systems for Nepal's banking and fintech sector, including an award-winning AI cheque clearing platform recognized as Nepal's ICT Best Product 2025.

CORE COMPETENCIES

LLM Fine-Tuning (LoRA/QLoRA, SFT) | RAG & Agentic Workflows | Low-Resource & Nepali NLP | Devanagari Tokenization | Computer Vision & OCR | Cloud Deployment (AWS, Docker)

TECHNICAL SKILLS

LLM & NLP: LLM Fine-Tuning (LoRA/QLoRA, Instruction Tuning/SFT), Hugging Face Transformers, PEFT, TRL, Transformer Architectures, Devanagari Tokenizer Adaptation, RAG Architecture, LangChain, LlamaIndex, LangGraph (ReACT Agents), Prompt Engineering, Gemini API

ML/DL & Computer Vision: Python, PyTorch, TensorFlow, YOLOv5, OpenCV, OCR (Tesseract), Machine Learning

Backend & APIs: Django, FastAPI, REST API Development, Django REST Framework

Databases & Vector Stores: PostgreSQL, MySQL, MongoDB, Qdrant, FAISS, SQLAlchemy, Firebase

Data Engineering: Python ETL, SQL Pipelines, Data Curation & Cleaning, API Integration, Data Transformation

Cloud & DevOps: AWS (S3, EC2, IAM, VPC), Docker, Nginx, CI/CD

AWARDS & RECOGNITION

ICT Best Product Award — Nepal 2025 | QuickFox

Awarded to Quick-CCA (AI-Powered Cheque Clearing Application) at Nepal's national ICT awards, recognizing outstanding innovation in financial technology and AI-driven banking automation.

WORK EXPERIENCE

QuickFox — Automation & AI Engineer

June 2025 – Present

QuickRef: Banking Knowledge Assistant — Gemini-to-Local LLM Migration

Tech Stack: LLaMA, Mistral, Hugging Face Transformers, PEFT, TRL, LoRA/QLoRA, Gemini API, LangChain, LlamaIndex, LangGraph, Qdrant, Python, PyTorch

- Leading the migration of QuickRef — QuickFox's RAG and agentic (LangGraph ReACT) banking/finance knowledge assistant — off the Gemini API onto a self-hosted, fine-tuned open-source LLM (LLaMA, Mistral) to eliminate per-call inference costs.
- Fine-tuning the replacement model using LoRA/QLoRA and supervised instruction tuning (SFT) via Hugging Face Transformers, PEFT, and TRL, targeted at Nepali/Devanagari financial-document Q&A.
- Curating, cleaning, and augmenting Nepali/Devanagari financial-domain training data, and adapting tokenizer vocabulary/merge rules to reduce token fragmentation on Devanagari script.
- Designing evaluation benchmarks to compare the fine-tuned local model's accuracy and response quality against the existing Gemini-based baseline.
- Preserving the existing RAG retrieval and ReACT agentic workflow (semantic chunking, Qdrant vector search, multi-turn tool use) while swapping the underlying generation model, keeping responses

explainable and source-grounded.

Quick-CCA: AI-Powered Cheque Clearing Application (ICT Best Product, Nepal 2025)

Tech Stack: Django, React, Docker, Nginx, PostgreSQL, Python, YOLOv5, OCR, Computer Vision

- Engineered the full Django backend for a production AI cheque clearing platform used by a Nepalese bank, automating signature verification and OCR to eliminate manual cheque authentication.
- Built REST APIs integrating with Finacle (Infosys core banking system) to execute live financial transactions, managing cheque workflows, inference results, and audit-compliant transaction records in real time.
- Containerized the full-stack application with Docker and Nginx, delivering a production-ready deployment serving the React frontend and all backend API routes.
- Reduced cheque validation time to near real-time by optimizing computer vision inference pipelines, significantly cutting per-cheque manual processing effort.

AI Signature Verification & OCR Engine (Devanagari & Latin Script)

Tech Stack: Python, FastAPI, OpenCV, Tesseract, YOLO Models, Custom ML Models, LLM-Assisted Training, Docker, Nginx

- Built a fully offline FastAPI-based signature verification and OCR system that extracts cheque fields (account number, amount, payee, date) in Devanagari and Latin script with high accuracy — ensuring complete data privacy with no cloud dependency.
- Accelerated custom OCR model fine-tuning by using LLMs to auto-generate structured training labels, improving extraction accuracy while keeping all production inference on-device.
- Engineered a signature matching pipeline using feature extraction and similarity scoring, with preprocessing (noise removal, skew correction, contrast enhancement) to handle real-world cheque quality variations reliably.

FREELANCE WORK

Agentic AI Financial Intelligence Platform

Tech Stack: FastAPI, React, FAISS, PostgreSQL, Python, RAG Architecture, JWT

- Designed and built an end-to-end agentic AI platform for financial operations — autonomously handling journal processing, ledger management, P&L generation, and balance validation without manual input.
- Architected a hybrid retrieval system combining PostgreSQL and FAISS vector search, enabling the AI to deliver explainable, source-grounded financial insights.
- Developed FastAPI backend services for transaction management and financial computation with automated Debit/Credit consistency validation.

EDUCATION

Bachelor of Computer Engineering — Pokhara University | 2019 – 2024

+2 in Science — Shree Gogan HSS | 2016 – 2018

ADDITIONAL INFORMATION

Languages: English, Nepali (native), Hindi

Certifications: Associate Data Engineer in SQL (Datacamp), Data Engineer in Python (Datacamp), AI Engineer (Datacamp), Machine Learning Specialization — Intro, Supervised, Advanced Algorithms, Unsupervised (Coursera, Andrew Ng)