MACHINE LEARNING ENGINEER

Yamraj Khadka

OPEN TO OPPORTUNITIES

CONTACT

Location	Biratnagar, Nepal ■■
GitHub	github.com/yamrajkhadka
Org	github.com/e-wakil
HuggingFace	huggingface.co/yamraj047
LinkedIn	linkedin.com/in/yamraj-khadka

SKILLS

LLM & Fine-tuning
Mistral-7B · Fine-tuning · GGUF Quant · Instruction Tuning

RAG & Retrieval
FAISS · SentenceTransformers · LangChain · LangGraph · Cross-encoder Reranking

Deep Learning
PyTorch · TensorFlow · Keras · U-Net · CNNs · LSTM/GRU/RNN

Engineering
FastAPI · Streamlit · React Native · MLflow · Docker · PyMuPDF

EDUCATION

BE, Computer Engineering
Tribhuvan University, IOE
Purwanchal Campus · In Progress

CERTIFICATIONS

Machine Learning Specialization
DeepLearning.AI × Stanford · Dec 2024

Deep Learning Certification
Simplilearn · Jan 2025 · ID 7801489

2025 Data Fellowship
DataCamp · May 2025

13.5 GB
Model fine-tuned (Mistral-7B FP16)

0.67
Mean IoU — satellite segmentation

< 24 h
Open-source community adoption

509+
LinkedIn followers (public learning)

ML Engineer · AI Researcher · Nepal ■■

Computer Engineering undergraduate building production-grade AI systems — from raw PDFs to deployed LLMs. Authored the **Nepal Legal AI System**: a complete Mistral-7B fine-tuning pipeline serving live APIs and a React Native app, community-adopted and re-quantized into Q2–Q8 GGUF variants within 24 hours of release. Presented **satellite image segmentation research at ICRTAI 2025** (mIoU 0.67). Documents everything publicly — because building in the open is how I think and grow.

EXPERIENCE

Artificial Neural Networks Intern
May 2025 – Present
Planto AI · Remote

- Building ANN-based solutions applied to real product problems in a startup environment.

Conference Presenter — ICRTAI 2025
June 2025
1st Intl. Conference on Recent Trends in AI · Nepal

- Presented research on U-Net land cover segmentation with custom composite loss (Focal Tversky + Weighted CCE).
- Received mentorship from Prof. Dr. Sudan Jha (AI & IoT) and Asst. Prof. Pukar Karki.

Independent ML Engineer
Sep 2023 – Present
Self-employed · Biratnagar, Nepal

- Built end-to-end ML pipelines: LLM fine-tuning, RAG systems, CV models, multi-agent frameworks, and full-stack deployments.
- Published open-source models on Hugging Face; maintained 17+ days of public deep learning logs on LinkedIn (500+ followers).

PROJECTS

Nepal Legal AI System
FLAGSHIP · OPEN SOURCE
github.com/e-wakil

End-to-end legal LLM pipeline for Nepal's National Penal Code 2017: PyMuPDF extraction → hierarchical chunking with deterministic IDs (*npc2017_p1_c1_s1_sub1*) → instruction dataset → **Mistral-7B fine-tuning** → GGUF quantization (13.5 GB → 4 GB) → FastAPI backend + React Native Android app. Community-adopted within 24h; re-quantized into Q2–Q8 variants by mradermacher (llama.cpp ecosystem).
Stack: Mistral-7B · FAISS · SentenceTransformers · FastAPI · Streamlit · React Native · llama.cpp · ctransformers
■ **4 live deployments · Q2–Q8 community GGUF variants · CPU-runnable at 4 GB**

Land Cover Segmentation via U-Net
RESEARCH · ICRTAI 2025
github.com/yamrajkhadka/project_on_img_seg_using_unet_archi

Custom U-Net on **DeepGlobe** dataset (7 land classes). Composite loss: 60% Focal Tversky + 40% Weighted CCE to handle severe class imbalance. Albumentations augmentation, morphological postprocessing, and Streamlit deployment. **Mean IoU: 0.67** (Forest 0.72, Urban 0.71, Unknown 0.98).
Stack: TensorFlow · Keras · Albumentations · OpenCV · Streamlit
■ **mIoU 0.67 · Presented at ICRTAI 2025 · Feedback from Prof. Dr. Sudan Jha**

GGUF + RAG Legal Chatbot
PRODUCTION
github.com/e-wakil/gguf-with-rag

Real-time legal assistant: bi-encoder retrieval (MiniLM-L6) → FAISS IVF index → cross-encoder reranking → GGUF LLM generation via WebSocket streaming. LRU query cache (500 entries), citation-grounded responses referencing exact law sections.
Stack: FastAPI · WebSocket · FAISS IVF · cross-encoder/ms-marco-MiniLM-L-6-v2 · ctransformers
■ **Real-time streaming · Citation-grounded · LRU-cached**

HerAI — Multi-Agent System (LangGraph)
AGENTIC AI
github.com/yamrajkhadka/Agentic-AI

5-agent orchestration with LangGraph: mood detection → conditional routing → RAG memory (FAISS + SentenceTransformers) → content generation → safety validation. Full stateful graph with typed state, conditional edges, and shared memory.
Stack: LangGraph · LangChain · FAISS · Streamlit · Ollama
■ **5 specialized agents · Stateful graph · Conditional routing**

RESEARCH PUBLICATION

ICRTAI 2025 · June 28–29 · Nepal

"Land Cover Segmentation from Satellite Imagery Using U-Net with Custom Loss and Morphological Postprocessing"

Custom U-Net with composite loss function (60% Focal Tversky + 40% Weighted CCE) trained on the DeepGlobe Land Cover Classification dataset. Addresses class imbalance across 7 categories. Data augmentation via Albumentations; morphological postprocessing for production masks. **mIoU: 0.67**. Mentorship from Prof. Dr. Sudan Jha on cloud-cover robustness.