Utshav Paudel

AI Engineer specializing in production LLM systems for clinical and healthcare use cases.

Kathmandu, Nepal

utshav.paudel466@gmail.com | LinkedIn: utshav-paudel | github.com/Utshav-paudel | huggingface.co/Utshav

Education

Madan Bhandari Memorial College — Kathmandu, Nepal

Bachelor of Science in Computer Science & Information Technology — *Apr. 2022 – Apr. 2026*

Experience

MedForce AI — London, UK (Remote)

AI Engineer — *Jun. 2025 – Present*

- AI engineer on a clinical-automation platform deployed in an NHS pilot. Designed the agent architecture and clinical RAG pipeline targeting clinician documentation and decision-support workflows.
- Fine-tuned domain-specific LLMs (Mistral 8×22B, Llama 3, Command R family) on curated hepatology cases for medical reasoning and structured documentation; built clinical evaluation harnesses covering factuality, citation grounding, and hallucination rate.
- Cut inference cost and latency for real-time clinical use through various caching techniques (KV cache, prompt caching, semantic response cache), request batching, and quantization; built monitoring for latency, drift, and hallucination signals.

Digital-Dandelion — London, UK (Remote)

AI Engineer — *Feb. 2024 – Jun. 2025*

- Built a clinical RAG system over EASL liver-disease guidelines using a multi-route retrieval architecture: simple queries served by embedding-based retrieval, complex queries routed to page-indexed advanced RAG. Reached 97% answer accuracy vs. 85% on a state-of-the-art baseline and cut end-to-end latency by 50%.
- Trained an EfficientNet image-classification system on GCP scoring 30,000+ dental websites for modernization; lifted accuracy from 89% to 95% via targeted data augmentation and an architecture switch from the prior baseline. Released the dental scraping and extraction datasets on Hugging Face.
- Built a multimodal ranking system for JLL evaluating hundreds of commercial real-estate offices via pairwise tournament comparisons across building exterior, interior, workspace, and floor-plan imagery to identify top performers.
- Fine-tuned an LLM on hundreds of past winning creative campaigns and creative-director rubrics for Page & Page (Novo Nordisk account); validated outputs on KPIs spanning relevance, creativity, and brand-fit.

Omdena — New York, USA (Remote)

Junior Machine Learning Engineer — *Jul. 2023 – Sep. 2023*

- Coordinated a 10–20-person distributed team across multiple timezones building scraping infrastructure that assembled a 46,000-article Nepali news corpus, the working dataset for the program's media-representation research.
- Built Nepali-language NLP models capturing local idioms to classify how women and marginalized groups are represented in Nepali news media.
- Designed a media-diversity scoring model with explicit progress metrics surfaced to research stakeholders.

Projects

MeroDaktar (AI-Powered Medical Platform) | *MedGemma 4B, Gemma-3, Voice AI, FastAPI*

- Voice-driven patient interviews in Nepali producing preliminary diagnostic reports and appointment booking.
- Fine-tuned a custom MedGemma-Nepali multimodal model (4B, Gemma-3 family) for Nepali clinical image-text reasoning; released fp16, 16-bit, and 4-bit variants on Hugging Face.
- FastAPI backend handling AI inference and health-record flow with input validation.

Llama-3-70B Extractor | *Llama 3 70B, Information Extraction*

- Fine-tuned Llama-3-70B-Instruct for structured information extraction from web content; released full-weights, adapter, and 4-bit variants.
- Trained alongside two open Hugging Face datasets (Dental_website_scraping, Dentalweb_extraction) used as the source corpus.

Technical Skills

Languages: Python, TypeScript, SQL, C++
AI/ML: LLMs, RAG, LangChain, NLP, Computer Vision, Fine-tuning
Cloud & DevOps: Google Cloud Platform (GCP), Docker, Git, MLOps
Frameworks: PyTorch, TensorFlow, Scikit-learn, Pandas, NumPy, FastAPI

Open Models & Datasets

Hugging Face: huggingface.co/Utshav. Released models include the **MedGemma-Nepali** series (medical multimodal, low-resource clinical language) and the **Llama-3-70B Extractor** (structured information extraction). Open datasets: **Dental_website_scraping** and **Dentalweb_extraction** from the Digital-Dandelion work.

Selected Writing

Your RAG is Probably Worse Than You Think: Three Ways to Find Out | *utshavpaudel.com.np/blog*