Yubraj Sigdel

+977-9848219555 yubrajsigdel1@gmail.com Yubraj Sigdel yubraaj11 syubraj

Summary

AI Engineer with hands-on expertise in Machine Learning, Generative AI, RAG, and Computer Vision, focused on building scalable, production-grade AI systems end to end. Experienced in designing recommendation engines, predictive intelligence pipelines, semantic retrieval platforms, and content moderation systems backed by distributed architectures (Kafka, gRPC, Redis). Passionate about turning research into reliable, low-latency applications through strong system design and MLOps practices.

Experience

Veel Inc. **Apr 2025 – Present**

AI Engineer & Team Lead *On-site*

- Mentored **4 interns** through a structured AI engineering roadmap, accelerating their progression from core ML/LLM fundamentals to production-grade applied AI development.
- Led end-to-end architecture and deployment of multiple user-facing AI systems, owning the full lifecycle from research and system design to scalable microservice deployment across the platform.
- Architected a low-latency **recommendation engine** using **Kafka, gRPC, Redis, and GraphQL** with fault-tolerant task processing pipelines, leveraging Redis-backed status tracking and automatic retry mechanisms to ensure robustness against partial failures at scale.
- Built an event-sourced **ETL and feature engineering pipeline** continuously ingesting and sanitizing behavioral, engagement, and metadata signals across **500K+ records**, feeding cleaned features into downstream ranking and scoring models in near real time.
- Designed a multi-signal **content moderation and review system** combining AI-driven policy assessment, brand/product tagging, and authenticity scoring to surface high-quality UGC while filtering misaligned content against enterprise trust standards.
- Engineered a **predictive pricing intelligence system** with a version-controlled MLOps pipeline leveraging market signals, entity metadata, and historical performance data to generate data-driven pricing suggestions, enabling reproducible deployment and continuous model retraining.
- Built a **semantic video retrieval platform** enabling segment-level, time-stamped discovery using multimodal embeddings and conducted R&D on graph-based relational retrieval architectures for next-generation content intelligence.
- Explored and prototyped **AI-assisted video editing** capabilities, investigating generative and analytical approaches to accelerate creative workflows on the platform.
- Implemented **concurrent task orchestration** across distributed services using Kafka consumer groups with Redis-backed processing checkpoints, enabling idempotent retries and end-to-end status observability for long-running async workloads.
- Engineered an in-house **video transcoding and media processing pipeline** as a cost-efficient alternative to managed cloud services, optimizing platform-scale media transformation and workflow throughput.

Vertex Special Technology (P.) Ltd. **May 2024 – Apr 2025**

Junior AI Engineer / AI Trainee *On-site*

- Built GenAI-powered automated Gherkin test generation pipelines using RAG, minimizing hallucinations and improving software validation reliability.
- Automated DOM extraction, XPath generation, and Selenium workflows to reduce manual QA effort.
- Enhanced test automation frameworks through caching optimization and dynamic generation pipelines, reducing execution time by 40%.
- Contributed to **TelePlantDoctor** by building image embeddings, semantic retrieval pipelines, and robust computer vision workflows.

FruitPunch AI **Jul 2024 – Sep 2024**

Contributor and Volunteer *Remote*

- Annotated Nepali documents for OCR, NLP, and NER systems under the **AI for Women Farmers initiative**.
- Collaborated with OCR teams to strengthen Nepali text identification systems.

FuseMachines Nepal **Apr 2023 – Jul 2023**

AI/ML Intern *Hybrid*

- Built a context-based Nepali News Recommendation System from scratch, including data scraping, embedding pipelines, and deployment.
- Trained a Nepali Sentence Transformer from scratch using translated STS datasets with NepaliBERT baseline embeddings.

Education

Kathmandu Engineering College **2018 – 2023**
Bachelor's in Computer Engineering

Gorkha United Public School **2016 – 2018**
+2 (Science)

Projects

DistantFrames — Python Package for Intelligent Frame Extraction + Video Filtering GitHub

- Designed and published **DistantFrames**, a developer-focused Python package for **intelligent frame extraction, scene-aware sampling, and automated video frame filtering** to simplify downstream AI and media workflows.
- Enabled efficient extraction of **high-signal frames** from video streams while reducing redundancy through configurable temporal spacing, content-aware filtering, and optimized frame selection logic.
- Built to support **computer vision pipelines, dataset generation, moderation workflows, model preprocessing, and creator media tooling** with simplified developer APIs.
- Focused on **Python package architecture, reusable tooling, developer productivity, and scalable preprocessing infrastructure** for production-ready media systems.

FFmpeg MCP — Natural Language Video Editing via MCP + FFmpeg MCP Server

- Architected and developed a production-ready **FFmpeg MCP (Model Context Protocol) Server** that transforms **plain natural language instructions into executable video editing workflows**, enabling simplified creator and developer interaction with complex media operations.
- Enabled intelligent execution of **video trimming, resizing, subtitle generation, frame extraction, compression, transcoding, format conversion, and advanced media transformations** without requiring users to manually write FFmpeg commands.
- Bridged **AI-assisted natural language interfaces** with robust FFmpeg automation pipelines, significantly reducing technical friction for media editing and workflow orchestration.
- Focused on **developer productivity, media infrastructure abstraction, creator tooling, and AI-powered workflow enablement** within the MCP ecosystem.
- Demonstrated strengths in **systems design, media engineering, automation tooling, developer experience, and production AI integration**.

Handwritten Document Conversion — Vision Transformer (ViT) + NepBERT + YOLOv8 GitHub

- Developed a scalable **handwritten Devanagari digitization system** using a two-stage architecture combining **Vision Transformer (ViT)** for visual encoding and **NepBERT** for sequence decoding.
- Integrated **YOLOv8 word-level detection** for precise localization of handwritten segments, improving recognition quality and document segmentation accuracy.
- Achieved a strong **9.08% Character Error Rate (CER)** on evaluation datasets, demonstrating robust performance for real-world handwritten text digitization.
- Built to support **document automation, script preservation, OCR innovation, and Nepali language accessibility**.

Roman English to Nepali Transliteration — mT5 + Hugging Face GitHub

- Built an end-to-end **Romanized Nepali to Devanagari transliteration platform** addressing practical communication barriers for Nepali digital users.
- Curated and transformed transliteration datasets into **production-compatible mT5 training pipelines**, ensuring scalable multilingual model adaptation.
- Fine-tuned transformer architectures to improve **language accessibility, localization, and user-friendly script transformation**.
- Published both model and dataset on **Hugging Face**, emphasizing reproducibility, accessibility, and open-source ecosystem contribution.

Nepali News Recommendation System — Sentence Transformers + Semantic Search GitHub

- Designed and deployed a **context-aware semantic recommendation engine** for Nepali news using embedding-based personalization and content similarity.
- Built complete data pipelines involving **news scraping, preprocessing, storage, and semantic indexing** for recommendation workflows.
- Implemented **Sentence Transformer embeddings** to deliver personalized article recommendations based on contextual reading behavior.
- Strengthened expertise in **recommendation systems, semantic retrieval, NLP engineering, and personalization architecture**.

Technical Skills

Languages: Python, SQL, Bash

AI / ML: LLMs, Generative AI, RAG, NLP, Computer Vision, Recommendation Systems, Semantic Search, Predictive Modeling, Fine-Tuning

Frameworks: PyTorch, Hugging Face Transformers, LangChain, LlamaIndex, FastAPI, Sentence Transformers, Scikit-learn

Backend & Systems: Microservices, Kafka, Redis, GraphQL, gRPC, REST APIs, MCP, FFmpeg

Databases: PostgreSQL, Milvus, ChromaDB, pgvector

Retrieval & Search: Embedding Pipelines, Semantic & Hybrid Search, Graph-Based Retrieval, Vector Indexing

Data & Automation: Pandas, NumPy, Selenium, BeautifulSoup, Feature Engineering, Workflow Automation

DevOps: Docker, AWS, Git/GitHub, CI/CD, Production AI Deployment

Publication

College Chatbot Using RASA, KEC Journal of Science and Engineering, 2024. DOI