Rahul Dev Banjara

rdbanjara07@gmail.com +977-9866121812 GitHub LinkedIn Portfolio

Skills

Languages: Python, JavaScript, C, Java, SQL
Technologies & Tools: Python, FastAPI, LangChain, LangGraph, RAG, AWS, Docker, HuggingFace, MLflow, TensorFlow, PyTorch, YOLO, Git, CI/CD, Vector Databases, SQLAlchemy, Neo4j, Finetuning

Work Experience

TekBay Academy November 2025 – Present
AWS Machine Learning Trainer

- Designed the complete 3-month curriculum for AWS Certified Machine Learning Engineer (MLA-C01) preparation, including practice exams and evaluation materials.
- Created intuitive class presentations and student study resources, focusing on clarity, visual explanations, and minimal jargon.
- Designed and conducted hands-on labs and workshops, guiding students through real-world problem-solving using Amazon SageMaker and other AWS ML services.
- Supervised and mentored students on their capstone projects, providing technical guidance on problem formulation, model selection, MLOps practices, and deployment strategies.
- Actively teaching 20+ students per batch on weekdays, providing continuous mentoring and doubt resolution.

TekBay Digital June 2025 – Present
Associate AI/ML Engineer

- Worked on an internal healthcare platform, building an Agentic GraphRAG pipeline for both data ingestion and inference over medical datasets. Implemented the complete backend using FastAPI, enabling secure, conversational access to customer data for doctors and clients.
- Developed an Arabic OCR system that preserves document layout by combining layout analysis, OCR, and vision-language models to accurately reconstruct multi-column documents with correct structure and reading order.
- Built a document verification and fraud detection pipeline using AWS Textract, Bedrock, and Lambda. Created synthetic datasets; trained, evaluated, registered, and deployed models; and automated the full MLOps lifecycle using AWS SageMaker Pipelines, including post-deployment monitoring.
- Designed and optimized cost-efficient, high-performance Athena queries to extract actionable insights from terabytes of organization-wide AWS logs stored in Amazon S3.
- Worked on multiple vanilla RAG-based chatbot projects for internal and client use cases.

Fusemachines March 2024 – October 2024
AI Fellowship 2025

- Engineered an industry-grade Computer Vision and AI for Governance project, applying state-of-the-art deep learning models to solve real-world public-sector and compliance-related problems.
- Gained hands-on experience across the entire machine learning lifecycle, including data collection, preprocessing, feature engineering, model training, evaluation, deployment, and performance analysis.
- Worked on multiple NLP assignments, including sentiment analysis and text classification, using modern embedding techniques and transformer-based architectures.
- Participated in structured research paper reading sessions, analyzing recent advances in AI and discussing practical implementation strategies.
- Completed intensive training in Supervised and Unsupervised Learning, Deep Neural Networks, and Deep Generative Models, with a strong emphasis on applied problem-solving and best practices.

Projects

Recruitment Automation & Job Marketplace Platform *(Ongoing Development)* [GitHub]

- Designing a multi-agent recruitment automation system using LangGraph to orchestrate specialized agents across the hiring lifecycle, from job posting to onboarding.
- Implementing LLM-powered job description generation and semantic resume screening with compatibility scoring, emphasizing transparency, fairness, and bias reduction.
- Developing explainable candidate-ranking dashboards to visualize and interpret model-driven hiring insights for recruiters and HR teams.
- Building conversational agents for automated candidate communication, including email- and calendar-based interview scheduling and audio-based interview evaluation.
- Automating offer letter generation, digital signatures, and HR approval workflows to streamline post-selection processes.
- Extending the platform with an Agentic GraphRAG-based candidate knowledge base.

AI Software Engineer (CLI Tool) *(Ongoing Development)* [GitHub]

- Built a CLI-based AI software engineer capable of generating complete systems, including frontend, backend, database

schemas, and filesystem interactions.
- Utilized advanced agentic AI concepts such as subgraphs, guardrails, human-in-the-loop workflows, and secure filesystem operations.

Complete MLOps Fraud Detection System

- Implemented an end-to-end MLOps pipeline on Amazon SageMaker for credit card fraud detection.
- Captured inference data from real-time endpoints and stored it in Amazon S3.
- Configured EventBridge to trigger retraining every 10 days, running a SageMaker Pipeline that preprocesses data, transforms schemas, trains models, and evaluates performance.
- Conditionally registered models to the SageMaker Model Registry based on evaluation thresholds and automatically deployed the latest approved model to production endpoints.
- Set up Model Quality Monitor and Data Quality Monitor, along with model cards for governance.
- Detected data drift via CloudWatch alarms and sent alerts through SNS email notifications, signaling when retraining was required.

Nepali Document Genuineness Verification System [GitHub]

- Developed a system to verify the authenticity of Nepali government documents, including citizenship certificates, passports, and driving licenses.
- Created a large synthetic dataset using document templates and Python scripts with dynamic text placement and augmentation, addressing the scarcity of real Nepali document samples.
- Used ResNet50 for document-type classification and routed documents to specialized YOLOv8 models, each trained for localized feature detection.
- Applied EasyOCR for Nepali text extraction and PaddleOCR for English text extraction, followed by cleaning and normalization using regex and NLP preprocessing.
- Verified authenticity by matching extracted fields against a mock government database (JSON) using Levenshtein distance, requiring a minimum 90% similarity score.
- Built a user-friendly Streamlit interface for document upload and real-time verification results.
- Achieved an overall verification accuracy of 82%.

Personalized RAG Chatbot [GitHub]

- Built a personalized AI chatbot combining RAG with Llama 3.1 (8B) for context-aware, document-based responses.
- Developed an async FastAPI backend with ChromaDB for persistent vector storage and session-based memory cleared every 10 minutes for privacy.
- Created a responsive frontend using vanilla HTML, CSS, and JavaScript, with modular separation of parsing, prompting, and inference.
- Leveraged Groq LPU for low-latency inference and Streamlit for managing document embeddings and the knowledge base.
- Deployed on AWS Lightsail with Nginx as a reverse proxy for secure frontend and backend access.
- Delivered a full document upload, embedding generation, and query workflow.
- Solved the problem of deploying a personalized chatbot for a company landing page under a $4/month cost constraint.

Education

Herald College Kathmandu 2023 - 2026
B.(Hons) in Computer Science
Relevant Coursework: AI and Machine Learning, Object-Oriented Programming, Software Engineering, High Performance Computing, Human Computer Interaction, Computer System Architecture

Leadership & Community Involvement

- Leader at DevCorps AI Learners Community, Herald College Kathmandu:
  - Conducted a hands-on session on NumPy, Pandas, and Data Analysis for 70+ students across the college.
  - Organized and led a workshop on LLMs and Prompt Engineering, covering zero-shot and few-shot prompting, chain-of-thought reasoning, and core NLP concepts such as tokenization and LLM input processing.
  - Delivered an in-depth NLP to Generative AI workshop, including practical training on LangChain to build a functional chatbot.
  - Actively participated in community events such as Showcase Fusion, presenting innovative alumni projects spanning computer vision, Unity-based applications, and e-commerce systems.
  - Collaborated closely with peers to cultivate a supportive, peer-driven learning environment within the AI community.

Certifications

- **AWS Certified Machine Learning – Associate (MLA-C01)** [Verification Link]
- **Microdegree™ in Artificial Intelligence — Fusemachines** [Verification Link]