

AAYUSH MAHARJAN

Machine Learning Engineer

Kathmandu, Nepal | +977 9808084972 | mhrzn.aayush@gmail.com | [LinkedIn](#)

SUMMARY

Machine learning engineer with 2+ years building and deploying ML systems in production for a US healthcare company, after 3 years as a software engineer. I work across the full pipeline, from scoping the problem with stakeholders to model design, training, and deployment through CI/CD into live Java and Python services. My document classifiers run at 98%+ accuracy in production. I also have research experience evaluating LLM and RAG systems, and I regularly run internal training sessions on ML topics for other engineers. Currently doing an MSc in IT with an AI specialization.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer, Smart Data Solutions (SDS)

May 2024 – Present

Eagan, MN, USA (Remote, based in Kathmandu)

- Own the document classification models for healthcare claims and attachments (EOBs, EOPs, UB-04s, CMS-1500s, ERAs, itemized statements, workers' comp). These run in production through the QuickClaim portal at 98%+ accuracy.
- Built image based classification pipelines using CNNs, U-Net and Transformer encoders to route documents into the right categories, replacing the older rule based approach.
- Built OCR and NLP pipelines that pull structured fields out of scanned medical documents for downstream claim automation. A lot of this work is figuring out how to turn messy real world documents into something a model can actually learn from.
- Set up and maintain the deployment side: Jenkins and GitLab CI/CD for model releases, plus an internal workflow for packaging and publishing Python environments so other teams can reuse them.
- Integrated Python model inference into our Java services using Jep, including handling CUDA memory inside JVM processes.
- Run internal training sessions on CNNs, NLP, Transformers and generative AI for the wider engineering team.

Software Engineer, Smart Data Solutions (SDS)

Jul 2021 – Apr 2024

Eagan, MN, USA (Remote, based in Kathmandu)

- Built and maintained Java services for healthcare claim processing, working with onshore PMs to turn vague client requirements into shipped features.
- Worked on Coupon Cabin, a high traffic browser extension, mostly on reliability and performance.
- Did code reviews across the team and mentored newer engineers on debugging and design patterns.

Data Engineer Intern, Fusemachines

May 2021 – Jul 2021

Kathmandu, Nepal

- Worked with senior data engineers on pipeline design and deployment using Hadoop, Spark and Airflow.

Machine Learning Engineer Intern, Treeleaf.ai

Nov 2020 – Feb 2021

Lalitpur, Nepal

- Structured training program covering classical ML, deep learning and computer vision. Implemented object detection, face detection and face recognition models.

PROJECTS & RESEARCH

RAG evaluation on multi-hop QA (MSc research)

- Compared four setups on HotpotQA: a plain LLM, standard RAG, RAG with chain of thought prompting, and RAG with DeepSeek-R1-Distill as the generator.
- Wrote the full evaluation pipeline myself, covering Answer EM/F1, Supporting-Fact EM/F1 and Joint EM/F1, including answer extraction and stripping reasoning traces from R1-Distill outputs.
- Ran everything on T4 GPUs using 4-bit NF4 quantization (BitsAndBytes) and FAISS for per question retrieval to fit 7B models in limited memory.

Lumo SFA (in development)

- A pharma sales force automation backend I am building with Django, Django Ninja and PostgreSQL. Has JWT auth, role based access control across six roles, and an adapter layer for plugging in LLM features later. Using uv, pytest and ruff.

EDUCATION

MSc in Information Technology (AI specialization), Islington College

Expected 2027

Kathmandu, Nepal. Coursework: Deep Learning, NLP, Computer Vision, Generative AI, applied RAG research

Bachelor of Information Management (BIM), National College of Computer Studies

2016 – 2020

Kathmandu, Nepal. CGPA 3.36 / 5

TECHNICAL SKILLS

ML / DL: PyTorch, TensorFlow/Keras, Hugging Face Transformers, scikit-learn, CNNs, U-Net, Transformer architectures, fine-tuning, 4-bit quantization (BitsAndBytes)

LLMs: RAG pipelines (chunking, embeddings, FAISS/vector stores), chain of thought prompting, DeepSeek-R1-Distill, Mistral, Qwen, LangChain, LLM evaluation (EM/F1, supporting fact analysis)

MLOps & Deployment: Jenkins, GitLab CI/CD, Docker, Linux, JEP (Python in Java), CUDA and GPU memory management, Python environment packaging, Kaggle/Colab

Languages: Python, Java, JavaScript, SQL, C

Data & Backend: Hadoop, Spark, Airflow, REST API design, Django, Flask, Node.js, PostgreSQL, SQL/NoSQL

Tooling: Git, Bash, uv, pytest, ruff, Selenium, n8n, SSH/WinSCP

CERTIFICATIONS

- Deep Learning Specialization, DeepLearning.AI (Coursera)
- Machine Learning: Tree-Based Models
- MongoDB: Data Modeling