SUMIT ARYAL

Machine Learning Engineer

+977-9840254499 | sumitaryal310@hotmailcom | aryalsumit.com.np

LinkedIn | GitHub | Hugging Face | Google Scholar

Kathmandu, Bagmati - 44600, Nepal

OBJECTIVE

I am a Machine Learning Engineer who turns ideas into AI apps that people use. I have created chatbots that talk like a friend and recommendation systems that feel like they know you. I enjoy the entire process, from writing code in Python to Dockerizing, deploying on Kubernetes, and setting up CI/CD so that it just runs. What really gets me excited is working together with product, design, and operations to solve real problems as a team.

EXPERIENCE

- **Root Level AI [🌐]** *January 2025 - Present*
  Machine Learning Engineer Kathmandu, Nepal
  - Architected and deployed a Retrieval Augmented Generation conversational AI system following Agentic AI patterns.
  - Optimized knowledge retrieval pipelines by implementing hybrid search (sparse + dense embeddings) and context-aware reranking, reducing hallucinations in LLM responses.
  - Research, design and implementation of a two-tower recommendation system, replacing the traditional recommendation system, to increase user engagement.

- **DooriIT [🌐]** *April 2024 - January 2025*
  Machine Learning Engineer Kathmandu, Nepal
  - Developed conversational agents using LLamaIndex, integrating OpenAI and Google Gemini APIs.
  - Implemented CI/CD-driven RAG pipelines with FastAPI, Docker, and Kubernetes
  - Mentored two interns on data preparation, model training, and evaluation for sentiment analysis and NER.

- **Fusemachines [🌐]** *January 2023 - August 2023*
  AI Fellow Kathmandu, Nepal
  - Completed intensive ML/DL curriculum, focusing on techniques like regression models, neural networks, and transformer-based architectures such as BERT.
  - Built sentiment-analysis and text-classification models
  - Applied image processing techniques, including detection and segmentation.

- **Bajra Technologies [🌐]** *September 2022 - December 2022*
  QA Trainee Kathmandu, Nepal
  - Developed and executed test cases for web applications, identifying and reporting over 50 bugs, enhancing product reliability.
  - Automated end-to-end testing using Cypress, reducing manual testing efforts by 15%.
  - Conducted API and load testing using Postman and JMeter.
  - Collaborated with developers to ensure bug fixes, contributing to a project delivery on time.

EDUCATION

- **Pulchowk Campus, IOE, Tribhuwan University** *November 2019 - April 2024*
  Bachelors in Computer Engineering Lalitpur, Kathmandu
  - Relevant Courses: Human Language Technologies, Data Mining, Artificial Intelligence, Advanced Calculas, Linear Algebra, Probability and Statistics, Digital Signal Analysis and Processing, Object Oriented Programming, Big Data Technologies, Discrete Mathematics, Data Structure and Algorithms.

SKILLS

- **Programming & Scripting:** Python, C/C++, Shell Scripting
- **ML/DL Frameworks:** PyTorch, HuggingFace Transformers, Scikit-learn, TensorFlow
- **Applied ML Frameworks:** Llama Index, LangChain, LangGraph, Qdrant Vector DB, OpenAI API, Gemini API
- **Web Development & APIs:** FastAPI, RESTful Design, Streamlit
- **Databases& Data:** Postgres, SQLAlchemy, Pandas, Numpy
- **Cloud & DevOps:** AWS, Docker, Git, Kubernetes, CI/CD Pipelines
- **Specializations:** Natural Language Processing, Deep Learning, Large Language Models, Computer Vision
- **Soft Skills:** Communication, Problem Solving, Teamwork, Quick Learning, Mentoring

Projects

- **Nepali GEC: Developed a system for grammar checking and correction in Nepali language.** *June 2023 - March 2024*
  Tools: BERT, Python, Hugging Face, Pytorch, Flask [📄]
  - Developed a pipeline utilizing BERT for Nepali Grammar Correction.
  - Developed Nepali Grammar Detection model, achieving 91.15% accuracy
  - Curated a large parallel corpus for the task of Nepali Grammar Correction.
  - Created a system that takes Nepali text as input, checks its grammar, and suggests corrections.
- **Chat with Multiple PDFs: Developed a system for querying multiple PDFs using RAG** *April 2024*
  Tools: Langchain, Hugging Face, Python [GitHub]
  - Developed a system where users can upload PDFs and ask questions about the content.
  - Implemented Langchain and Hugging Face Transformers for document segmentation, vector conversion, and FAISS for vector database.
  - Semantic search was used to retrieve ranked results, which are utilized by an LLM to generate answers.
- **HTML parser using LLM: Developed an API for extracting e-commerce attributes from HTML content** *May 2024*
  Tools: FastAPI, Python, Hugging Face, Docker [GitHub]
  - Built a FastAPI-based API using Hugging Face's 'meta-llama/Meta-Llama-3-8B-Instruct' to parse e-commerce HTML, extract attributes validate content, and return CSS selectors/XPaths as JSON.
  - Clean HTML by removing extra tags and optimize the LLM context window to save tokens on meaningful content.
  - Containerized the service with Docker.
- **Recommendation System: A personalized travel itinerary web-app for travelers to Nepal** *January 2023 - May 2023*
  Tools: React, Django, Python, Flask, REST APIs, Pandas [GitHub]
  - Developed a recommender system that generates personalized itineraries based on user preferences and budget.
  - Implemented collaborative filtering to enhance the accuracy of recommendations by analyzing user behavior and preferences.
- **Other Projects:** *June 2021 - Present*
  Tools: Python, C++, PyGame, SFML, Tkinter
  - 8 Puzzle Visualizer: Implemented and visualized different algorithms, such as A^*, BFS, DFS, IDDFS and Greedy to solve the 8-puzzle problem using Python and Tkinter. [GitHub]
  - Bachiyo Game: Mario-like platformer game with various levels and sound effects using C++ and SFML. [GitHub]
  - Image Compression: Compressed images using Huffman Tree Algorithm in C++ [GitHub]
  - Stadium Modeling: Modeled a stadium using Python, Pygame and Blender. [GitHub]

Publications

C=Conference, J=Journal, S=In Submission, T=Thesis

[S.1] Sumit Aryal*, Anku Jaiswal (2024). **BERT-Based Nepali Grammatical Error Detection and Correction Leveraging a New Corpus**. Presented at *IEEE INSPECT-2024*, ABV-IIITM, Gwalior, India, December 07-08, 2024. [Link]

[T.1] Sumit Aryal, et.al. (2024). **Nepali Grammar Correction**. Undergraduate Thesis, Pulchowk Engineering Campus, Institute of Engineering, Tribhuvan University. [Link]

Honors and Awards

- **Best Project Award** *December 2024*
  Pulchowk Campus [🌐]
- **Best Paper Award** *December 2024*
  IEEE INSPECT-2024, ABV-IIITM [🌐]

Professional Memberships

- **Nepal Engineering Council** *October 2024 - Present*

Certifications

- **DeepLearning.AI:** Deep Learning Specialization *December 2023*
- **DeepLearning.AI:** Generative AI with Large Language Models *October 2023*
- **Fusemachines Inc:** Microdegree in Deep Learning *August 2023*
- **Samsung:** Coding and Programming *May 2023*
- **Fusemachines Inc:** Microdegree in Machine Learning *May 2023*
- **DeepLearning.AI:** Machine Learning Specialization *November 2022*

Additional Information

Languages: Nepali (Native/Fluent), English (Fluent/Professional Proficiency), Hindi (Conversational Proficiency)
Interests: Machine Learning, Natural Language Processing, Low-Resource Language Processing, Large Language Models, Artificial Intelligence