

SUMIT ARYAL

officialsumitaryal@gmail.com | aryalsumit.com.np

 LinkedIn |  GitHub |  Hugging Face |  Google Scholar

Kathmandu, Bagmati - 44600, Nepal

OBJECTIVE

Machine learning practitioner and open-source contributor with expertise in NLP and retrieval-augmented generation pipelines. Experienced in end-to-end research, and implementation of retrieval-augmented generation (RAG) pipelines and document-centric NLP workflows. Passionate about collaborating on research, and bridging theoretical advances with real-world applications.

EDUCATION

• Pulchowk Campus, IOE, Tribhuvan University

November 2019 - April 2024

Bachelors in Computer Engineering

Lalitpur, Nepal

- Relevant Courses: Human Language Technologies, Data Mining, Artificial Intelligence, Big Data Technologies, Calculus, Linear Algebra, Probability and Statistics, Digital Signal Analysis & Processing, Object-Oriented Programming, Discrete Mathematics, Data Structures & Algorithms.

PUBLICATIONS

C=CONFERENCE, J=JOURNAL, S=IN SUBMISSION, T=THESIS

[C.1] Sumit Aryal*, Anku Jaiswal (2024). **BERT-Based Nepali Grammatical Error Detection and Correction Leveraging a New Corpus**. Presented at *IEEE INSPECT-2024*, ABV-IIIITM, 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. [\[Link\]](#)

EXPERIENCE

• Smart Data Solutions

November 2025 - Present

Software Engineer I, Machine Learning

Eagen, MN, USA

- **Structured Data Extraction:** Developed document processing pipelines using Vision Language Models (VLMs) with custom processing logic for structured data extraction from variable-layout PDFs and scanned images; optimized high-throughput inference with vLLM for low-latency deployment
- **Customer Facing Chatbot:** Building a Retrieval-Augmented Generation (RAG) application powering a customer-facing chatbot; integrating vector retrieval, conversational memory, and LLM reasoning to provide accurate, context-aware responses for client inquiries.

• Root Level AI

January 2025 - October 2025

Machine Learning Engineer

Kathmandu, Nepal

- **Engineered an Agentic RAG system:** Designed and studied retrieval augmented generation for multi-turn support dialogs. Built a reproducible benchmark suite with synthetic and human-labeled queries, evaluated dense plus sparse hybrid retrieval with cross-encoder reranking, and conversational memory. Reported Recall@k, and MRR, and end-to-end latency.
- **Retrieval Systems at Scale:** Operated a multitenant, sharded Qdrant cluster with replication and write consistency. Studied shard key strategies, and tenant-scoped payload indexes for fast filters.

• DoorIT

April 2024 - January 2025

Machine Learning Engineer

Kathmandu, Nepal

- **LLM Integration:** Created RAG assistants using OpenAI and Gemini. Used LLMs to automate content generation tasks like summaries and descriptions, replacing manual tasks and improving operational efficiency.
- **Text Analytics Pipeline:** Developed sentiment analysis and NER systems using Hugging Face transformers.
- **Team Development:** Mentored 2 interns through the complete NLP project lifecycle, from dataset creation to BERT fine-tuning and evaluation.

SELECTED OPEN SOURCE CONTRIBUTIONS

• LlamaIndex

Qdrant Integration, Memory Component

2025

- **Qdrant Sharding:** Enabled distributed Qdrant deployments, delivered configurable sharding and replication with write consistency controls in QdrantVectorStore, added custom shard-key selection across CRUD flows. [PR #19652](#).
- **Payload field indexes:** Implemented payload_indexes support with creation logic in both sync and async flows; added fixtures and tests. [PR #19743](#).
- **Memory Component:** Added missing synchronous wrapper for put_messages in memory component to align sync and async interfaces. [PR #19746](#).

• OpenAI

GPT-OSS

2025

- **Python Agents SDK example:** Added a ready-to-run example that sets up an agent with local gpt-oss models, connects a filesystem MCP server, custom weather tool, and streams agent responses. [PR #14](#).

SKILLS

- **Core Languages:** Python, Rust, C/C++, Shell Scripting
- **ML/DL Stack:** PyTorch, HuggingFace Transformers, Scikit-learn, TensorFlow, Pandas, NumPy
- **Natural Language Processing:** LlamaIndex, LangChain/Graph, Fine-tuning, Agentic Workflow, Text Analytics
- **ML Infrastructure & MLOps:** Docker, Kubernetes (EC2/ECR), AWS (EC2, S3, Lambda, ECR), CI/CD Pipelines
- **Data Systems:** Qdrant, PostgreSQL, SQLAlchemy, DynamoDB, Redis
- **API & Deployment:** FastAPI, RESTful Design, Streamlit, Serverless Architectures
- **Specializations:** NLP, RAG, Low-Resource Language Processing, Recommendation Systems, Computer Vision
- **Soft Skills:** Communication, Problem Solving, Teamwork, Quick Learning, Mentoring

PROJECTS

- **Research Agent: Research Agent with planning, search, reflection and synthesis** 2025
Tools: Python, Pydantic, asyncio [G]
 - **Agent graph and control:** Async state machine for plan, search, reflect, decide, synthesize, runs concurrent searches with adaptive concurrency and reflection cadence, honors budgets and deadlines, early stops on confidence, cancels and drains tasks before synthesis.
 - **Retrieval and synthesis:** Tool mediated web search, extract citations from grounding metadata, resolve and normalize URLs, insert inline citation markers, deduplicate sources, synthesize the final answer with adaptive thinking budgets and answer length targets.
- **RAG Project: Retrieval pipeline with OCR ingest and hybrid vector search** 2025
Tools: FastAPI, Hugging Face embeddings, Qdrant, PostgreSQL, Gemini OCR [G]
 - **RAG architecture:** FastAPI service with Postgres and Qdrant, OCR text extraction via Gemini, structured chunking, and dual dense plus sparse embeddings, indexed in Qdrant with filterable metadata.
 - **Retrieval and answer flow:** hybrid search that fuses dense plus sparse results using Reciprocal Rank Fusion with tunable top-k and alpha, returns grounded answers with cited snippets.
- **Other Projects:** June 2021 - Present
Tools: Python, C++, PyGame, SFML, Tkinter
 - **HTML Parser using LLM:** Cleans HTML and extracts e-commerce attributes using LLM independent of the e-commerce website, outputs CSS selectors and XPath as JSON. [G]
 - **Travel Recommendation Project:** Developed a recommender system that generates personalized itineraries based on user behavior, preferences and budget. [G]
 - **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. [G]

HONORS AND AWARDS

- **Best Project Award** December 2024
Pulchowk Campus [G]
 - Awarded for developing a Nepali Grammar Correction System.

PROFESSIONAL MEMBERSHIPS

- **Nepal Engineering Council**, Membership ID: 87426 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, Large Language Models