Koshish Shrestha

9863336676 koshish62@gmail.com koshishshrestha.com.np linkedin.com/koshishshrestha

About Me

Research-driven AI/ML Engineer specializing in NLP and generative AI, focused on building and evaluating LLM-based systems with strong awareness of model behavior and deployment considerations.

Education

Kathmandu Engineering College **2019 – 2024**

Bachelor of Computer Engineering *Kalimati, Kathmandu*

(Aggregate Percentage: 81.27%)

Skills

Programming & Data: Python, SQL, NumPy, Pandas, Scikit-learn, DVC
Machine Learning Techniques: Exploratory Data Analysis (EDA), Feature Engineering, Predictive Modeling, Model Evaluation, Time Series Forecasting
Deep Learning & Generative AI: PyTorch, TensorFlow, Transformers, BERT, Large Language Models (LLMs), Prompt Engineering, Retrieval-Augmented Generation (RAG), Agentic Workflows, LangChain
MLOps & Deployment: FastAPI, Docker, MLflow, LangFuse, Airflow, Model Monitoring, CI/CD Fundamentals, RabbitMQ
Tools: AWS (Foundational), Git, GitHub, BitBucket, Linux, Jira
Soft Skills: Leadership, Team Collaboration, Problem Solving, Self-Learner

Experience

Machine Learning Engineer II @ GritFeat Solutions **August 2024 – Present**

- Owned end-to-end development of 3+ LLM-based NLP systems, from prompt engineering and RAG design to deployment and monitoring using Python, Transformers, LangChain, FastAPI .
- Designed and optimized RAG-powered conversational agents, improving response relevance by 30% and reducing inference cost by 25% using LLMs, embeddings, retrieval pipelines .
- Migrated synchronous REST APIs to an event-driven queuing architecture, increasing system robustness and throughput by 40% using RabbitMQ, async workers, Docker.
- Integrated GPT-style LLMs into 3+ production applications, handling multi-turn dialogue, context management, and failure scenarios using FastAPI, prompt engineering, agentic workflows .
- Built end-to-end time series forecasting pipelines (SARIMA, ARIMA, DeepAR) for business-critical applications, orchestrated with Airflow and tracked using MLflow, enabling accurate, data-driven decision-making.
- Built LLM evaluation and monitoring pipelines across 10+ model iterations, tracking hallucinations and regressions using MLflow, LangFuse, custom evaluation metrics .

Associate Machine Learning Engineer @ GritFeat Solutions **July 2024 – July 2025**

- Developed production-grade conversational AI systems using LLMs, RAG pipelines, and agentic workflows for real-time applications.
- Implemented advanced prompt engineering strategies to handle complex user intents, ambiguity, and multi-step reasoning, improving response accuracy.
- Built model deployment pipelines with FastAPI and Docker, ensuring robust inference reliability and consistent latency across production workloads.
- Designed logging and evaluation frameworks to monitor response accuracy, latency, and user satisfaction across multiple models using MLflow, LangFuse.
- Researched emerging LLM techniques and assisted in rapid prototyping of new architectures and workflows.

Machine Learning Fellow @ GritFeat Solutions **April 2024 – July 2024**

- Gained hands-on experience across the full machine learning lifecycle, including data preprocessing, model training, evaluation, and deployment.

Publications

Comparison Analysis of Nepali News Classifier

- Authored and presented the paper on OKRP conference and published in the KEC Journal. Paper Link

Personal Projects

KaanunSathi **RAG, RAPTOR, LangChain, LLM**

- Designed and implemented an advanced RAG system for Nepali legal documents, optimized for accuracy, traceability, and response latency.
- Applied hierarchical retrieval (RAPTOR) and embedding optimization to handle long-context legal queries efficiently.
- Built evaluation workflows to measure answer relevance and failure modes across retrieval and generation stages.

Nepali News Classifier **Python, NLP, BERT, Transformers**

- Performed comparative analysis of ML and deep learning models, optimizing classification accuracy through feature engineering and preprocessing.
- Implemented data preprocessing, visualization, and analysis techniques to enhance model performance.
- Built the classifier in PyTorch using the NepBERT model from Hugging Face.

LipiStudio **Python, FastAPI, OpenAI Whisper**

- Built a real-time transcription and captioning system using Whisper, focusing on low-latency STT and human-readable output for production usage.

Activities

Student Lead @ PALS **September 2023 – May 2024**

- Enriched my skills in leadership, teamwork, collaborations and work ethic under guidance of IIT Alumni Champions.
- Attended IIT Madras for special programs and events, demonstrating a commitment to expanding knowledge and networking opportunities.
- Collaborated with IIT alumni at PALS to promote responsible technology innovation among South Asian student networks.

President @ KEC IT Club **2021 – 2024**

- Lead team of I.T Club in all sectors.
- Represented the club in different coding competitions, hackathons etc. organized by other colleges and organizations.

College Representative @ Engineers Vlogs

- Played a key role as a member of the organizing team for Orbit Engineering Expo 2022 and handled entire robotics section during the expo.

Certifications and Course Reviewed

Data Scientist Associate - DataCamp [Link]

AWS Cloud Quest: Generative AI Practitioner - Training Badge [Link]

AWS Cloud Quest: Cloud Practitioner - Training Badge [Link]

Resident Student Workshop: Think, Create, Engineer - IIT, Madras **December 2023**

- Achieved first place in the implementation of a 'StormWater Drain Blockage Detection System' as part of a mini project. Led a team to design and execute the project, demonstrating problem-solving skills and technical competence.
- Engaged in a dynamic workshop at IIT Madras focused on fostering creative thinking and engineering skills. Explored System Thinking Sustainability, Design Thinking and various concepts.

HEx Genius Hackathon **July 2023**

- Developed a web application that facilitates direct communication between local bodies and government entities.
- Implemented an issue resolution system where unresolved issues are automatically escalated to higher authorities. Also incorporated an upvote system and other features to enhance user engagement and feedback.
- Utilized the MERN stack for the project.