Nabaraj Subedi

subedinab | Website | nsubedi1@uwyo.edu | +1 (307) 357 6341

Education

University of Wyoming 2025–2027
Master's of Science in Computer Science
Relevant Courses: Intro to LLM, Machine Learning & Data Mining
Institute of Engineering, Tribhuwan University 2020–2024
Bachelor of Engineering–Electronics, Communication & Information Engineering

Work Experience

Graduate Research Assistant Aug 2025 – Present
University of Wyoming

- Conducted research on multimodal LLMs and RAG systems, designing experiments, evaluating model performance, and optimizing cost–efficiency, accuracy, and groundedness for large-scale enterprise chatbot applications.
- Developed and maintained data/model evaluation pipelines, integrating cloud tools, clustering methods, and validation workflows to support fine-tuning, benchmarking, and deployment of state-level (WYDOT) conversational AI systems.

Junior Machine Learning Engineer July 2024 – Aug 2024
PalmMind Technology

- Built scalable RAG chatbots for EV dealers and insurance clients using OpenAI models, LangChain, and LangGraph, leveraging orchestrator-worker workflows, routing, and parallelization for efficiency.
- Enhanced data retrieval and system performance by implementing web scraping, OCR (Tesseract), and Redis-based caching for faster, context-aware conversational AI.

Teaching Assistant Sep 2024 – July 2025
Pashchimanchal Campus, Tribhuwan University

- Taught C programming,Object Oriented Analysis and Design,Operating System,and Data Mining.
- Designed lab activities, guided semester-end projects, and provided detailed feedback to enhance students' practical programming and problem-solving skills.

Projects

WYDOT Multimodal Chatbot
Research Project, University of Wyoming

- Developed and deployed a multimodal RAG chatbot (images/audio/video/PDFs) using fine-tuned Gemini 2.5 Flash/Pro, with enhancements like multi-hop reasoning, self-validation, and Workspace integration.
- Processed 1,656+ WYDOT engineering documents, extracting structured/unstructured data using LlamaParse, Unstructured, PyMuPDF, and Qwen2.5 captioning, and performed clustering to map document similarity and versioning.
- Built and validated a large, high-quality QnA dataset and used it to fine-tune multimodal models for improved precision, groundedness, and conversational efficiency.

Citi biker history data analysis
Personal Project

- Performed large-scale data analysis on 35M+ Citi Bike trips using BigQuery and Python, completing full data engineering steps including ingestion, cleaning, feature engineering, and anomaly detection, resulting in a high-quality dataset ready for modeling.
- Built advanced analytics and machine learning pipelines (Random Forest, clustering, KDE, Sankey, community detection) to uncover user behavior patterns, weather impacts, rebalancing needs, and network dynamics—providing actionable insights for urban mobility optimization.

Banking Chatbot
Company Project

- Researched LLM embedding strategies, cost optimization techniques, and token-efficient prompting, evaluating multiple embedding models and architectures to improve semantic accuracy while reducing inference cost for large-scale deployments.
- Developed agentic tool-calling OpenAI-based chatbots for banking and e-commerce use cases, enabling automated workflows such as file retrieval, SQL querying, and task execution through structured tool-calling pipelines.

Nepali Image Captioning
Generating Coherent Paragraph-Length Descriptions Using Transformer
Undergraduate Major Project

- Developed a deep learning system for generating paragraph-length Nepali captions using Transformer architecture with Inception V3 feature extraction, trained on a curated dataset (20,350 pairs from Stanford Paragraph dataset translated/corrected to Nepali + 800 cultural heritage images).
- Compared Transformer (BLEU-1: 0.23, BLEU-2: 0.35, BLEU-3: 0.53, BLEU-4: 0.59) against LSTM with ResNet152, optimizing hyperparameters (batch size 32, learning rate 0.01, 8 attention heads, dropout 0.2) for Nepali's complex grammar.

- Preprocessed data with tokenization, vectorization (Keras TextVectorizer), and vocabulary building (14,022 unique words); integrated into a full-stack app (React frontend, Flask/Node backend); published in Journal of Soft Computing Paradigm (March 2024).

Skills

Languages	Python, C, C++, Java, SQL
AI/ML	Pytorch,Pandas, NumPy, Seaborn, TensorFlow, PyTorch, Scikit-learn, Keras, LangChain, LangGraph, HuggingFace transformer, CNNs, RNNs, Transformers, Linear/Logistic Regression, Clustering (K-means, DBSCAN)
Tools	Googel Cloud, ARCC Computing cluster
LLMs	Opensource LLM and VLM,Gemini Finetuning
Soft Skills	Communication, Team Collaboration, Time Management

Publications

Subedi, N. et al. (Jan. 2024a). "Drowsiness and Crash Detection Mobile Application for Vehicle's Safety". In: *Journal of IoT in Social, Mobile, Analytics, and Cloud* 6.1, pp. 54–66. URL: https://doi.org/10.36548/jismac.2024.1.005.

– (Jan. 2024b). "Nepali Image Captioning: Generating Coherent Paragraph-Length Descriptions Using Transformer". In: *Journal of Soft Computing Paradigm* 6.1, pp. 70–84. URL: https://doi.org/10.36548/jscp.2024.1.006.