Suyog Dev Khanal

whysuyog@gmail.com | +977 9761694441 | Imadol, Nepal | Linkedin | Personal Site | Github

SUMMARY

Aspiring Applied GenAI Engineer focused on learning by building AI projects, with a growing emphasis on **Generative AI (GenAI)** applications. Interested in working with LLMs end-to-end—prompting, tool/function calling, RAG-style knowledge workflows, and lightweight evaluation—while keeping solutions practical and user-focused. Developing skills across Python, APIs, automation, and integrating GenAI into real projects.

EDUCATION

Vedas College, Tribhuwan University, *BCA* 12/2022 – 12/2026 | Kumaripati, Nepal

Gems Institute of Higher Education, *+2 Science* 06/2020 – 06/2022 | Dhapakhel, Nepal

PROJECTS

Streaming LLM Chat Agent (LangChain, Groq, FastAPI) Jan, 2026

- Built a streaming chat agent using **LangChain** and **Groq GPT-OSS-120B** for real-time LLM responses.
- Implemented **Server-Sent Events (SSE)** to stream model outputs incrementally.
- Added **LangGraph memory checkpointing** for thread-based conversation state.
- Exposed the agent via a **FastAPI API endpoint** and deployed it.

Semantic Similarity Tool (Gemini Embeddings) Jan, 2026

- Implemented a semantic similarity tool using **Google Gemini embeddings (gemini-embedding-001)**.
- Computed **cosine similarity with NumPy** to measure similarity between text inputs.
- Structured outputs to return **vector embeddings and similarity scores**.

Tokenization Analysis Toolkit (tiktoken, WordPiece) Jan, 2026

- Built a toolkit to analyze **LLM tokenization using OpenAI tiktoken (BPE)** and **BERT WordPiece**.
- Extracted **token IDs, token texts, and token counts** for input sequences.
- Used **HuggingFace Transformers (BertTokenizerFast)** to demonstrate WordPiece tokenization.

Mini Search Engine (Keyword, TF-IDF, BM25 search) On the way..

- Built a document ingestion and text normalization pipeline covering lowercasing, punctuation/digit removal, tokenization, and stopword filtering.
- Implemented a keyword search retriever that matches normalized query tokens against preprocessed document tokens.
- Next: implement TF-IDF and BM25 ranking algorithms as more sophisticated retrieval methods.

SKILLS

Python, Langchain, FastAPI, Numpy, LLM APIs, Git, Ollama

INTERESTS

Philosophy, History, Strategy, Vedic Astrology, Chess, Instruments, Creation, Storytelling