Hemant Kumar Sah

✉ hsah5116@gmail.com ☎ 6294031626

https://www.linkedin.com/in/hemant-sah-0b1515204/

Objective

Aspiring **AI Engineer** with hands-on experience in **Python** and **Generative AI frameworks**. Skilled in building and deploying **LLM-powered applications** using **Hugging Face, LangChain, LangGraph, and RAG pipelines**. Passionate about exploring emerging AI technologies, solving real-world problems, and contributing to innovative projects.

Education

2020 – 2024 **B.Tech-CSE(RGPV)-72.9%**
Gwalior, India *Institute of technology and management*

- Member of Google Developer Student Club

Projects

AI-Powered Assistant (Python, OpenAI/Gemini API)

- Developed an intelligent AI assistant capable of understanding user queries and executing multi-step tasks programmatically.
- Integrated real-world tools: fetches live weather data, executes system commands, and generates application templates (e.g., HTML/CSS/JS Todo App).
- Implemented structured reasoning using START → PLAN → TOOL → OUTPUT workflow for chain-of-thought execution.
- Added rate limiting, error handling, and JSON parsing to ensure reliable API interactions.
- Demonstrates skills in Python programming, API integration, automation, and practical AI applications.

DocAI — RAG for document intelligence

- Built a custom RAG pipeline for document search and KPI extraction across structured, semi-structured, and handwritten data with integrated OCR/NLP.
- Optimized chunking, embeddings, and hybrid retrieval over multi-vector databases for low-latency, high-precision results.
- Implemented an evaluation pipeline (precision/recall, hit@k, qualitative analysis) to tune retrievers and prompts.
- Delivered grounded chatbot answers and exposed features via AzureML-managed endpoints.
- Stack: Python, LangChain/LangGraph, Transformers, Vector DBs, OpenAI/Gemini APIs, AzureML.

Skills

Languages:
Python (Pandas, NumPy), SQL

Tools / Platforms:
Git, Ollama, LangGraph, Vector Databases, OpenAI API, Gemini API

Frameworks / Libraries:
LangChain, Hugging Face, Pydantic, Streamlit, Transformers

Concepts:
Large Language Models, RAG (Retrieval-Augmented Generation), Prompt Engineering, MCP (Model Context Protocol), Tokenization, Embeddings, Attention Mechanism

Achievements

- Achieved **Grade A (134/170)** in *Business English Certificate Preliminary* Examination.
- Independently built and tested multiple AI prototypes using **LangChain**, **Gemini**, and **OpenAI APIs**