Surbhi

linkedin.com/in/surbhi-1688b3183 github.com/Surbhi-sys ss2000surbhisingh@gmail.com
+91 8804401220

Summary

GenAI Engineer with **3.3+ years of experience** building enterprise-grade **LLM-powered applications**. Specialized in **Retrieval-Augmented Generation (RAG)**, vector search, prompt engineering, and end-to-end AI pipelines using cloud-based and on-premise LLMs. Strong experience in building AI chatbots, document ingestion pipelines, semantic retrieval systems, and conversational orchestration. **Immediate Joiner.**

Skills

Generative AI / LLM Engineering:
LLMs: Chat Bison, Ollama | RAG (Retrieval-Augmented Generation) design | Prompt Engineering (few-shot, chain-of-thought, multi-turn) | Semantic Search | Hallucination Mitigation | Agentic AI / Crew AI

Embeddings & Vector Databases:
Gecko Embeddings | Ollama Embeddings | ChromaDB | Vector Indexing | Query Optimization | Efficient Semantic Retrieval | Metadata Tagging | Document Chunking

AI Frameworks, Tooling & Programming:
Python | LangChain | Embedding Pipelines | Agentic AI Orchestration (Crew AI) | Monitoring / Logging for AI workflows

Experience

Quess Corp — GenAI Engineer Jan 2025 – Dec 2025 | Pune

- Designed and developed **AskHR**, an enterprise conversational AI chatbot enabling employees to access HR policies and internal documents using natural language queries.
- Built a scalable **RAG-based pipeline** to process and query enterprise documents from PDFs.
- Implemented **chunking, metadata enrichment, and vector indexing** to enhance retrieval accuracy and reduce hallucinations.
- Developed **semantic search and context-aware retrieval** using vector embeddings for accurate and compliant responses.
- Created **prompt templates and multi-turn conversational flows** aligned with enterprise communication standards.
- Ensured production reliability through **logging, evaluation, and unit testing** of AI services.

HCL Technologies — GenAI Developer Oct 2022 – Jan 2025 | Noida

- Developed a **Manufacturing Knowledge Assistant** to help operators and engineers quickly find critical information from work instructions, maintenance manuals, and safety documents.
- Built a **RAG-based AI system** with semantic search using vector embeddings for fast and accurate retrieval of operational documents.

- Implemented **context-aware guidance and step-by-step instructions** to answer technical queries related to machine operations and safety procedures.
- Designed **AI pipelines** for document ingestion, embedding generation, retrieval, and response validation across multiple structured and unstructured sources.
- Reduced errors and improved operational efficiency by **minimizing hallucinations and ensuring factual accuracy** in AI-generated responses.

Internship Experience

Wipro — Project Engineer Intern Mar 2022 – Jul 2022, Pune

- Assisted in backend development for an internal e-commerce application.
- Implemented **global search, pagination, cart and wishlist logic**.
- Developed **Excel/CSV export functionality** and supported API debugging.
- Collaborated with Angular teams to ensure smooth API integration.

DigiLocker (Ministry of IT) — Software Engineering Intern Aug 2020 – Dec 2020, Delhi

- Contributed to backend and frontend modules under the **Digital India** initiative.
- Assisted in API validation, feature enhancements, testing, and documentation.

Education

B.Tech in Computer Science Engineering
BPUT, Rourkela 2022
CGPA: 9.13

Achievements

Google Developer Student Club — Web Developer Lead (2020–2021)
Led workshops and mentored students; built strong foundations in system design and AI engineering.