Rahul Sharma

+919111205441 ◇ rahulsharma383842@gmail.com ◇ LinkedIn

SUMMARY

Generative AI & Python Full Stack Developer with 4 years of experience building scalable, AI-driven applications using Python (FastAPI, Flask), React.js, and modern databases like PostgreSQL, MongoDB.
Specialized in developing Retrieval-Augmented Generation (RAG) systems, LLM-based chatbots, and knowledge-assistant platforms using LangChain, Hugging Face, and OpenAI APIs.
Adept at integrating vector databases (FAISS, Pinecone, Chroma) for semantic retrieval, designing end-to-end AI pipelines, and deploying production-ready apps on AWS with CI/CD automation.

EXPERIENCE

Generative AI & Full Stack Developer Dec '24 — Present
Zonforce Pvt Limited Ahmedabad, India

- Designed and implemented **end-to-end RAG pipelines** using **LangChain**, **Hugging Face**, and **OpenAI APIs**.
- Integrated **multiple LLM providers (Groq, Gemini, OpenAI, Claude)** for optimal performance and cost efficiency.
- Deployed **embedding and inference pipelines** on **AWS SageMaker** and **AWS Lambda**, orchestrated with **Step Functions** for automated retraining and scaling.
- Collaborated on **LLM cost optimization** by batching inference calls and caching embeddings to reduce redundant vector computations.
- Engineered REST APIs for admin dashboard modules including user management,leveraging PostgreSQL and MongoDB for efficient data handling.
- Built **FastAPI-based SSE streaming endpoints** for real-time conversational response delivery in React frontends.
- Developed a **React-based dashboard** allowing users to upload data, fine-tune bot personality, view chat analytics, and manage billing.

Generative Ai and Full Stack Developer Nov '24 — Dec '24
Calidig Solution DEWAS (Remote)

- Streamlined data aggregation and processing workflows to ensure accurate and timely reporting of performance metrics.
- Collaborated with cross-functional teams to design and implement sentiment analysis models, aligning insights with business goals and customer experience requirements.

Full Stack Developer May '24 — Nov '24
Clevdoc Platform - Bank NotiFicataion Sytem Gurugram, India

- Designed and developed an enterprise-grade **Document Portal** for secure file uploads, versioning, and contextual search using **Python (FastAPI, Flask)** and **React.js**.
- Implemented a **Retrieval-Augmented Generation (RAG) architecture** for intelligent document querying, integrating **LangChain**, **Hugging Face embeddings**, and **ChromaDB** for semantic search.
- Developed a **custom retriever pipeline** with adaptive chunking and embedding optimizations to improve context accuracy and reduce latency.
- Integrated **Pydantic-based structured response parser** and evaluation framework for consistent LLM outputs.
- Added **Token Counter & Cost Analysis** to track token usage across embedding and generation stages, enabling cost visualization via Streamlit dashboards.

Full Stack Engineer Oct '21 — May '24
Mphasis Pune, India

- Developed and optimized RESTful APIs using Python and FastApi to handle high-volume data delivery and scalable content access, ensuring performance and reliability across the e-learning platform.
- Worked on designing API documentation using tools like Swagger for better developer onboarding and seamless integration with client applications.
- Built dependency-based interceptors in FastAPI to validate incoming requests and enforce business rules before hitting core endpoints.
- Designed schema changes to support new business requirements.

SKILLS

Programming Languages Python, JavaScript, React
Databases MYSQL, PostgreSQL, MongoDB
Cloud Services AWS, SNS, AWS Lambda, DynamoDB, Amazon EC2, Gcp, Azure
Frameworks FastAPI, Flask
GenAI LangChain, Hugging Face, OpenAI, Vector Databases, RAG
Languages English

EDUCATION

Bachelors in Mechanical Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya (GPA: 6.62) Indore, India