Rohan Sharma  
📍 Bengaluru, India | 📞 +91 98765 43210  
✉ rohansharma.ai@gmail.com | 🔗 linkedin.com/in/rohansharma-ai  
💻 github.com/rohansharma-ai

PROFESSIONAL SUMMARY

AI/ML Engineer with 8+ years of experience in Generative AI, NLP, and Deep Learning. Proficient in building LLM-powered AI systems, Retrieval-Augmented Generation (RAG) pipelines, and fine-tuning transformer models. Expertise in AI-driven search engines, chatbots, and finance AI solutions using Hugging Face, OpenAI, LangChain, and Vector Search. Strong background in scalable AI architectures on AWS/GCP/Azure.

Key Strengths: Generative AI, LLM Fine-Tuning, Prompt Engineering, AI Search, RAG Pipelines, Transformer Models, Multimodal AI, MLOps, Cloud AI Deployment.

TECHNICAL SKILLS

* AI & LLMs: GPT-4, LLaMA, Mistral, Falcon, OpenAI APIs, Claude, Gemini, T5/BART
* Vector Search & RAG: FAISS, ChromaDB, Weaviate, Pinecone, Milvus
* NLP & Deep Learning: Transformers, BERT, LSTM, CNNs, Attention Mechanisms
* Cloud & MLOps: AWS (SageMaker, Bedrock), GCP Vertex AI, Azure AI, Kubernetes, Docker
* Big Data & Streaming: Apache Kafka, Flink, Spark, Databricks, NoSQL (MongoDB, ElasticSearch)
* Development & Frameworks: Python, PyTorch, TensorFlow, LangChain, FastAPI, Streamlit

WORK EXPERIENCE

Senior Generative AI Engineer

XYZ AI Solutions Pvt. Ltd., Bengaluru (2021 – Present)

* Built a streaming LLM chatbot with RAG, integrating FAISS + LangChain + OpenAI APIs, improving response accuracy by 30%.
* Developed an AI-powered search engine with multimodal RAG, increasing search relevance by 40% using T5/BART fine-tuned models.
* Designed a real-time fraud detection system leveraging Confluent Kafka + Databricks + MongoDB, reducing false positives by 25%.
* Fine-tuned BART/T5 for document summarization, optimizing latency and reducing summary length by 50% while maintaining context.
* Deployed scalable AI models on AWS using SageMaker, Lambda, and Step Functions, reducing costs by 20% with efficient auto-scaling.

Tech Stack: Python, PyTorch, TensorFlow, OpenAI, LangChain, FAISS, Pinecone, Kafka, MongoDB, AWS Bedrock, SageMaker

AI/ML Engineer

ABC Tech Pvt. Ltd., Hyderabad (2017 – 2021)

* Developed an AI Personal Assistant using LLM + RAG + Tools, integrating real-time context-aware retrieval for enterprise users.
* Optimized an E-commerce Customer Behavior Model with Kafka + NiFi + Spark, improving recommendation accuracy by 35%.
* Built a Social Media Sentiment Analysis system using Apache Flink + Elasticsearch, processing 500K+ tweets per day in real-time.
* Implemented a Portfolio Optimization AI for Algo Trading, leveraging quant models with LSTMs to enhance trading returns by 20%.

Tech Stack: Python, Hugging Face, PyTorch, Kafka, Spark, FastAPI, Power BI, Elasticsearch

KEY PROJECTS

1️⃣ Streaming LLM Chatbot with RAG

* Problem: Needed an AI chatbot with contextual memory and dynamic retrieval.
* Solution: Implemented LangChain + FAISS + OpenAI APIs to build a streaming AI chatbot with real-time knowledge retrieval.
* Impact: Improved chatbot accuracy by 30%, reducing hallucination cases.

2️⃣ AI-Powered Search Engine with Multi-Modal RAG

* Problem: Poor document search accuracy across unstructured datasets.
* Solution: Integrated T5/BART for NLP retrieval + vector search (FAISS/Pinecone).
* Impact: Enhanced search accuracy by 40% and reduced retrieval time.

3️⃣ Generative AI for Cybersecurity - AI-Based Threat Detection

* Problem: Need for an intelligent system to detect AI-driven cyber threats.
* Solution: Built an LLM-powered cybersecurity model with real-time threat detection.
* Impact: Improved anomaly detection by 35%, reducing cyberattack response time.

EDUCATION

* 🎓 M.Tech in Artificial Intelligence | IIT Hyderabad | 2017
* 🎓 B.Tech in Computer Science & Engineering | NIT Trichy | 2015