

AMAN SHARMA

Bengaluru, India | +91-9955306911 / +91 6204984854 | developeras1999@gmail.com | amansharma.in | [LinkedIn](#) | [leetcode](#) | [GitHub](#)

Summary

Software Engineer with 2+ years of experience contributing to the design, development, and optimization of software applications. Expertise in writing efficient and scalable code, and a history of effective collaboration with cross-functional teams on full-stack development and machine learning initiatives. Strong technical aptitude, reflected in extensive activity on competitive coding platforms.

SKILLS

- ❖ Programming Languages: Python, JavaScript, C++, SQL, JAVA
- ❖ Software Development: Full-Stack Development, Backend Development, Frontend Development, API Integration, System Design, Data Structures & Algorithms (DSA)
- ❖ Web Technologies: React.js, Next.js, FastAPI, Django, Tailwind CSS
- ❖ Cloud & Deployment: Docker, Kubernetes, Vercel, Netlify, Firebase, VPS (Hostinger), Streamlit
- ❖ Automation & CI/CD: Software Deployment Platforms, Methodologies and Tools, Git
- ❖ Database & Data Processing: Pandas, NumPy, ChromaDB, PostgreSQL, MongoDB
- ❖ AI & Machine Learning: Transformers, Hugging Face, LLM Models (Llama, Mistral), VLLM, Ollama, TensorFlow, Keras, LangChain, OpenAI Embeddings, ASR, TTS, langgraph
- ❖ Monitoring & Tools: Kafka, Prometheus, Grafana, Pydub

PROFESSIONAL EXPERIENCE

Salesken.ai – Machine Learning Intern (On-site) | Bengaluru, India *March 2025 – Present*

- ❖ Assisted with designing and developing high-quality software solutions by fine-tuning the Mistralai/Mistral-7B model, increasing goal detection accuracy by 10%.
- ❖ Optimized ASR service, enhancing memory efficiency by up to 25% and improving overall processing performance by 15%.
- ❖ Integrated a noise cancellation feature into the core product, leading to a 20% improvement in audio clarity for users.
- ❖ Conducted multiple proofs-of-concept (POCs) for 10-15 LLM models, evaluating their suitability and performance to meet user needs.
- ❖ Managed LLM inference servers using cutting-edge tools like VLLM and Triton, FastAPI, supporting high-throughput inference for critical applications.

Marathon Trends Advisory Private Limited – Software Developer & Research Analyst | Mumbai, India *May 2024 – March 2025*

- ❖ Revamped backtesting algorithms, reducing processing time by 67% through implementation of memoization and caching.
- ❖ Spearheaded the design and deployment of a full-stack application for research analysts using FastAPI and React.js, leading to a 50% reduction in manual data processing.
- ❖ Built a finance-specific chatbot based on RAG models with OpenAI embeddings, ChromaDB, and LangChain, resulting in 40% better response times.
- ❖ Engineered a sentiment analyzer service that leveraged real-time scraped data to deliver actionable market insights with 90% accuracy.
- ❖ Regularly constructed a technical financial database to manage over 3,000+ stocks data, optimizing data processing by 30% using Pandas.

GS Design- Full Stack Developer | Remote , May 2023 - April 2024

- ❖ Worked on front and backend development for global e-commerce clients like Squarepeg and Custoking.
- ❖ Designed and developed frontends for 5+ e-commerce applications using React.js and Tailwind CSS.
- ❖ Integrated APIs with cross-functional teams, ensuring seamless data flow and application functionality.

EDUCATION

Maulana Abul Kalam Azad University of Technology | West Bengal, India Bachelor of Technology (B.Tech) in Computer Science and Engineering *Graduated: 2023 | CGPA- 8.9*

Relevant Coursework:- [Mathematics for Machine Learning: Multivariate Calculus\(coursera\)](#) , [A Crash Course in Data Science\(coursera\)](#), [AWS Fundamentals: Going Cloud-Native, Cybersecurity and Its Ten Domains\(coursera\)](#), [Interactivity with JavaScript\(coursera\)](#)

PROJECTS

- ❖ [Interview Bot](#): Created an AI-powered speech-to-speech interview bot using Llama (LLM), PlayAI (TTS), and Whisper (ASR), allowing users to practice interviews for selected roles and receive immediate feedback.
- ❖ [OpenChat](#): Realtime group chatting application
- ❖ [Custom Rag](#): Built a Streamlit-based RAG application using Groq LLM and ChromaDB to enable document-based Q&A over uploaded PDFs/DOCX with chunked vector search and semantic retrieval.