Sumeet Khatri

Software Engineer | Machine Learning | System Design

sumeetkhatri1771@gmail.com | (+91)9328050693 | Pune | LinkedIn | GitHub | Portfolio

Education

B.Tech in Information Technology 08/2022 – Present
Indus University, Ahmedabad, (Expected July 2026)

Professional Experience

AWS APAC Solutions Architecture Virtual Experience 09/2025
Forage
- Designed a scalable AWS Elastic Beanstalk hosting architecture that reduced client latency by ~40% during peak load, achieving 20% cost savings.
- Presented the proposed architecture in clear, non-technical language, ensuring client understanding of functionality and cost implications.
- Configured and orchestrated AWS servicesto simulate enterprise-scale application hosting and integration.
- **Skills: AWS, Elastic Beanstalk, RDS, Lambda, Cloud Architecture, Communication, Cost Optimization, System Design**

Skills

- **Programming Languages:** C++, Python
- **AI/ML:** Generative AI, LLM Fine-Tuning, ML Pipelines, Model Optimization
- **Backend & Systems:** Distributed Systems, Microservices, REST APIs, System Design
- **Cloud & DevOps:** AWS, Docker, GCP, Kubernetes, CI/CD, Infrastructure as Code
- **Databases:** SQL, NoSQL
- **Core CS:** Data Structures & Algorithms, OOP, OS, DBMS

Technical Projects

LexiVoice AI
Skills: Python, Transformers, CUDA, NLP, Model Optimization, Generative AI
- Engineered a custom 15M-parameter Transformer SLM with tokenization, training loop, and optimizations.
- Reduced inference latency by 45%, enabling real-time AI legal support for 10K+ enterprise users.
- Addressed scalability challenges of large legal LLMs, making deployment feasible for law firms.

AI Coder Buddy
Skills: Python, LangChain, LangGraph, Groq API, AI Agents, Automation
- Spearheaded development of an AI agent that planned and generated multi-file applications automatically.
- Decreased development cycle time by 30%, accelerating full-stack software delivery.
- Collaborated with frameworks like LangChain + LangGraph to improve automation workflows.

Credit Risk Model
Skills: Python, Scikit-learn, Streamlit, MLOps, Data Science, Model Deployment
- Developed interpretable ML models with demographic + bureau data, improving accuracy by 20%.
- Reduced evaluation time by 60% through MLOps pipelines and automated deployment.
- Enhanced compliance for NBFC clients by providing faster and transparent predictions.

Insights LM
Skills: Python, RAG, Supabase, N8N, System Design, Knowledge Management
- Engineered a secure, open-source clone of Google Notebook LM with a serverless RAG pipeline (Supabase + N8N).
- Delivered citation-backed AI responses, improving compliance and enterprise knowledge accessibility.
- Strengthened data privacy by implementing private, self-hosted infrastructure.

Activities

- Solved 250+ problems on LeetCode (Neetcode roadmap)
- Secured 2nd place in District-Level Science Olympiad (Grade 9)