Ameer Hamza Khan

🔗 LinkedIn | LeetCode | GitHub | Kaggle

📞 +91-7987918845
✉️ameerhamzakhan1305@gmail.com

SUMMARY

B.Tech IT student with expertise in high-scale AI systems, LLM applications, and Full-Stack development. Solved **1000+ problems on LeetCode (Top 6% globally)**. Proven track record in optimizing RAG pipelines, synthetic data engineering, and fine-tuning domain-specific models for production-grade reliability.

Technical Skills

Programming: Python, C/C++, JavaScript (Node.js, React.js), SQL (PostgreSQL).
AI/ML & Generative AI: LLMs, RAG, LangGraph, AI Agents, Fine-tuning (Unsloth, QLoRA), Vector DBs (Qdrant, ChromaDB)
Tools & Infrastructure: Docker, AWS (ECS, S3), Git, Postman, GGUF Quantization, FastAPI, Streamlit.
Core Computer Science: Data Structures & Algorithms (DSA), Object-Oriented Programming (OOP), Operating Systems, DBMS, Computer Networks.

WORK EXPERIENCE

AI Engineer Intern | Vaquill AI (Dec 2025 – Present):

- **High-Scale RAG:** Engineered a legal retrieval system managing **19M+ documents**, implementing a scale-first architecture for Supreme and High Court records.
- **Deep Reasoning Tier:** Developed "Deep Search" via multi-hop retrieval, achieving a **91% Recall@10** (11% improvement over baselines) and reducing query latency by 24%.
- **Synthetic Data Engineering:** Architected a high-fidelity pipeline using **DeepSeek APIs** to generate 19k+ complex legal reasoning pairs, leveraging Chain-of-Thought (CoT) to bridge knowledge gaps in Indian statutes.
- **LLM Fine-Tuning:** Fine-tuned **Llama 3.1 8B** via Unsloth/QLoRA on **A100 GPUs**; outperformed **GPT-4o and o3-mini** in specialized BNS/BNSS/BSA legal benchmarks.
- **Reliability:** Built a critique layer (CRAG) with self-reflection loops, enforcing strict grounding to retrieved statutes to eliminate hallucinations in legal consultations.

***IDEAS-TIH (ISI Kolkata)** | Summer Intern | **May 2025 – Jul 2025** Live Demo | GitHub*

- **Dukaan Sahaayak:** Developed a smart inventory system using **LangGraph-powered AI agents** for automated billing and natural language-to-SQL conversion.
- **Process Optimization:** Integrated Gemini OCR for handwritten extraction, reducing manual billing work by **70%**.
- **Local Inference:** Deployed Ollama LLMs for offline query processing to ensure data privacy and system availability.

PROJECTS

***Proactive-AIoT-Assistant - Context-Aware Smart Environment** – (Python, LLMs, FastAPI, ChromaDB, AIoT) | GitHub*

- Designed a **SENSE-THINK-ACTION** pipeline interpreting sensor data (Fit, Maps, Calendar) to predict user needs and automate smart environment responses.
- Developed a hybrid reasoning model combining deterministic decision graphs with LLM intent extraction.

***Sadaf Bot – Conversational AI Assistant with Vision & Long-Term Memory** – (Python, LLMs, ChromaDB, Ollama, STT/TTS) | GitHub*

- Built a speech-enabled AI assistant with vision capabilities and modular long-term vector memory.
- Implemented **multi-threaded async processing** for real-time concurrent STT, TTS, and memory updates.

Achievements & Certifications

- **LeetCode**: Solved **1000+** problems, Max Rating **1835** (Top **6%)**, achieved a personal best rank of **1114**.
- **Codeforces**: Best Contest Rank **772** (in Codeforces Round 1062, Div. 4)
- **Certification**: 32-hour advanced training in Deep Learning, Generative AI, and Business Intelligence.
- **JEE Mains:** Ranked **Top 2%** among **1M+ candidates**; qualified **JEE Advanced (2022)**.

EDUCATION

Indian Institute of Engineering Science and Technology (IIEST), Shibpur | 2022 – 2026

Bachelor of Technology in Information Technology | **CGPA: 8.24**