Ishaan Aggrawal

Shamli, Uttar Pradesh, India
+91 9258895224 ishaanaggrawal101@gmail.com LinkedIn GitHub

Professional Summary

Innovative **Full Stack & Generative AI Engineer** with hands-on experience building production-grade GenAI systems, RAG pipelines, and LLM evaluation microservices. Strong foundation in **backend engineering, machine learning fundamentals, and scalable system design**. Experienced in deploying low-latency, cost-efficient AI applications with safety guardrails.

Education

Indian Institute of Information Technology (IIIT), Sonepat **2024 – 2028**
B.Tech in Computer Science and Engineering — GPA: 9.3/10 *Sonipat, Haryana*

Indian Institute of Technology (IIT), Madras **Ongoing**
B.Sc in Data Science and Applications — GPA: 8.5/10 *Online*

Technical Skills

GenAI & LLM Systems: RAG Architectures, Long-Context Grounding, LLM Evaluation, Prompt Engineering, Vector DBs (pgvector)
Backend & APIs: FastAPI, Flask, Node.js, Express, REST APIs, WebSockets, Docker
Frontend: React.js, Next.js, TypeScript, HTML5, CSS3
ML/Data: Scikit-learn, XGBoost, Pandas, NumPy, Feature Engineering, Model Evaluation
Databases & Tools: PostgreSQL, Supabase, MongoDB, Git, GitHub, DSA

Projects

BeyondChats – LLM Safety & Evaluation Microservice

- Built an asynchronous microservice to audit RAG chatbot responses using a tiered evaluation pipeline (cache → guardrails → 8B → 70B).
- Reduced evaluation cost by **80%** through intelligent model routing while maintaining high accuracy.
- Designed for scale using background task execution, rate limiting, and cache-first architecture.

AI Customer Support Agent – RAG System

- Developed a full-stack GenAI application that answers customer queries using document-grounded retrieval.
- Implemented semantic search using **pgvector + HNSW** and low-latency LLM inference via Groq.
- Deployed a live demo covering ingestion, retrieval, and generation pipelines.

AI Legal Document Analysis System (RAG + Guardrails)

- Built a GenAI system to analyze and compare legal documents using long-context grounding to prevent hallucinations.
- Implemented safety features including file validation, ghost PDF detection, and liability guardrails.
- Designed for regulated domains where accuracy and traceability are critical.

Customer Churn Prediction System

- Architected an end-to-end **Machine Learning pipeline** to predict customer attrition using demographic and financial data.
- Developed a robust **RESTful API using Flask** to expose the trained model for real-time inference requests.
- Integrated the AI backend with a responsive client-side interface to demonstrate seamless **model-to-app communication**.

- Deployed the full inference system on **Render**, ensuring scalable access to the prediction endpoint.

ChatterBox – Real-Time Chat Application

- Built a real-time chat application using the MERN stack and Socket.IO for low-latency messaging.
- Implemented JWT authentication, user presence tracking, and scalable WebSocket communication.

Coding & Problem Solving

LeetCode: 200+ problems, Max Rating: 1491
HackerRank: 4 in C/C++
Core: Data Structures, Graphs, Dynamic Programming

Certifications & Achievements

- CodeVeda Hackathon – 2nd Runner-up
- Hackzilla – IIIT Sonepat
- AutoML for Beginners – Udemy
- Linear Algebra & Probability Coursework