Aniket Majety

Atlanta, GA   aniketmaj21@gmail.com   +1 404 457 2790

amajety1

Education

Indian Institute of Technology, Madras *December 2023*
Bachelor of Technology in Chemical Engineering

Georgia State University *December 2025*
Bachelor of Science in Computer Science

- **GPA:** 3.7/4.0
- **Achievements:** President's List (Spring 2024, Summer 2024), Dean's List (Fall 2024)
- **Coursework:** Data Structures, Software Development, Machine Learning, Big Data Programming

Skills

Languages: Python, JavaScript, Java, SQL, HTML, CSS, C, C++
Technologies: Flask, React.js, Node.js, MongoDB, AWS (S3, Elastic Beanstalk, CloudFront), PyTorch, Pandas, Oracle, Unix, Git, Jira, Confluence, Docker, Kubernetes
Certifications: AWS Certified Cloud Practitioner (CCP)
Concepts: Object-Oriented Programming (OOP), SOLID Principles, Software Development Life Cycle (SDLC), Debugging, Problem Solving, RESTful APIs, Full-Stack Development, Cloud Computing, Data Analysis, Database Management, System Design, Performance Optimization, CI/CD, Agile

Experience

AI Engineer Intern *Atlanta, GA*
Inpharmd *Nov 2024 – Present*

- Developed a low-latency AI-driven clinical search engine by integrating TF-IDF ranking, reducing search latency by 80% while maintaining similar ranking accuracy to previous Sentence Transformer models.
- Fine-tuned a biomedical LLM using BioMedBERT, improving contextual relevance by 25% in clinical query responses.
- Optimized LLM inference for real-time medical queries, reducing API token usage by 30% through query truncation and compression techniques.
- Replaced Pinecone vector DB with PGVector in PostgreSQL, reducing monthly infrastructure costs by $20,000 while maintaining efficient vector search capabilities.
- Implemented multi-threaded parallel processing for AI queries, improving response speed by 60%.

Projects

Fullstack Developer — AI-Powered Medical Literature Assistant *Deployed Website Link*

- Engineered an AI-driven query expansion module using GPT-based models, increasing search coverage by 45% and improving retrieval of relevant medical literature.
- Optimized real-time LLM-based summarization pipeline, reducing response generation time by 35% while preserving high clinical accuracy.
- Implemented AI-powered content filtering and duplicate detection, reducing redundant search results by 50% and enhancing search precision.
- Developed a scalable AI inference pipeline using multi-threaded processing, reducing average API response time from 2.5s to 1.2s under high query loads.

Fullstack Developer — Study Match *GitHub*

- Built a system to match students with study partners by comparing courses, schedules, and majors, using a custom similarity algorithm, improving match accuracy by 50%.
- Added secure login and account protection by hashing passwords with bcrypt and implementing JWT-based authentication, reducing unauthorized access attempts by 90%.
- Created a real-time chat system using WebSockets and Redis pub/sub, enabling horizontal scaling across multiple servers to support 10,000+ concurrent users.
- Improved database performance by indexing frequent queries, denormalizing user-course relationships, and batching read/write operations, reducing response time from 1.8s to 500ms.
- Integrated Redis caching to store recent study match recommendations, reducing repeated database lookups by 60% and speeding up response times.