

Muhammad Maaz Ur Rehman

📍 Karachi, Pakistan | 📞 +92 3472918636 | ✉️ maazr098@gmail.com | [🌐 LinkedIn](#) | [🐙 GitHub](#) | [📁 Portfolio](#)

EXPERIENCE

Founding Full-Stack AI Engineer

April 2025 – Present

COROS AI

Remote

- **System Architecture:** Designed and built a production-grade AI platform **from scratch** using a **microservices architecture**, supporting 100 concurrent users.
- **RAG Optimization:** Architected and optimized Retrieval-Augmented Generation pipelines, reducing average AI response latency from **27{32s to 6{10s** by improving semantic search and retrieval efficiency.
- **AI Evaluation & Quality:** Implemented automated **AI eval pipelines** using relevance scoring to continuously measure and improve response accuracy and grounding.
- **Hallucination Reduction:** Improved grounding and reduced hallucinations using **HyDE-based retrieval strategies** and controlled context injection.
- **Memory System:** Engineered a **semantic memory layer** with topic-isolated user spaces, semantic scoring, and time-based decay (15-day high priority, 30-day reduced priority) to prevent context pollution and cross-topic leakage.
- **Multi-Model Orchestration:** Experimented with and integrated multiple LLM providers (OpenAI, DeepSeek, Together AI, Llama-based models) to balance quality, latency, and reliability.
- **Streaming & Async Pipelines:** Implemented streaming AI responses and asynchronous RAG pipelines to improve perceived latency and user experience.
- **Cloud Cost Optimization:** Reduced GCP infrastructure costs by optimizing instance scaling and shifting to request-based resource allocation.
- **Ownership & Leadership:** Owned features end-to-end including system design, backend APIs, AI logic, and deployment; contributed to architecture decisions, code reviews, and technical direction within a 3-engineer founding team.

Full-Stack Engineer

April 2023 – May 2025

PRIV, Inc.

Remote

- **Appointment Booking System:** Maintained a high-traffic booking system, increasing successful bookings by 30%.
- **Real-time Updates:** Reduced double-bookings by 95% using real-time availability features.
- **Scalable Notifications:** Built a scalable system using WebSockets, reducing missed appointments by 40%.
- **Performance Optimization:** Improved system performance by 20% through optimized queries and caching.
- **Mobile-First UI:** Created a responsive UI with React, boosting mobile bookings by 50%.

PROJECTS

NextBook AI | *Next.js, React, Flask, Llama, Pinecone, RAG* [\[Live Demo\]](#)

- Led the design and development of an AI-powered book recommendation system for Abul Kalam University Library. Built a **retrieval-augmented LLM pipeline** using library data and verified sources to improve recommendation accuracy.

NOTABLE ACHIEVEMENTS

Top Performer, CSIT Department: First Place for a semester-long OOP project in C++. Led a team to develop an innovative application demonstrating problem-solving, leadership, and collaboration. [\[View Details\]](#)

Winner, Data Insights Hackathon: Built an AI-powered Library Recommendation System applying ML algorithms for personalized recommendations, recognized for innovation and impact. [\[View Details\]](#)

TECHNICAL SKILLS

Languages: HTML5, CSS3, JavaScript (ES6+), TypeScript, Python, SQL

Frameworks: React, Next.js, Node.js, Express.js, Flask, LangChain, LangGraph

AI/ML Engineering: Retrieval-Augmented Generation (RAG), Agentic Workflows, LLM Orchestration, AI Infra

Databases: PostgreSQL, MySQL, MongoDB, Pinecone (Vector DB)

Cloud & Infrastructure: Microsoft Azure, Google Cloud Platform (GCP), Docker, Vercel, CI/CD

Tools and Technologies: Git, TailwindCSS, REST APIs, FastAPI, OpenAI API

Development Practices: Agile Methodology, AI-Driven Development (AIDD) Test-Driven Development (TDD)

EDUCATION

NED University of Engineering and Technology

Karachi, Pakistan

Bachelors in Computer Science, Major in Artificial Intelligence and Machine Learning

2023 – 2027