

MOHAMMED FARHAN

AI Engineer | Forward Deployed Engineer | LLM & RAG Systems | MLOps

+1 571-437-6155 | farhanay916@gmail.com | Chantilly, VA | linkedin.com/in/farhanzephyr | farruro.netlify.app

PROFESSIONAL SUMMARY

Full-stack AI engineer and technical founder who ships production LLM systems end to end — model design, RAG architecture, FastAPI backends, React front ends, and AWS/Azure cloud deployment. Founder of Studygy.io, an AI edtech platform serving 1,000+ users built on LLM fine-tuning and Retrieval-Augmented Generation. Forward-deployed mindset — owns problems from discovery to production, with MLOps discipline across CI/CD, containerized deployment, and cloud AI DevOps.

TECHNICAL SKILLS

Generative AI & LLMs: LLM Fine-Tuning, Retrieval-Augmented Generation (RAG), Prompt Engineering, Agentic AI Workflows, LangChain, LlamaIndex, OpenAI API, Azure OpenAI Service, Hugging Face

RAG Systems: Embedding Models, Semantic Search, Chunking Strategies, Hybrid Retrieval, Reranking, Query Rewriting, RAG Evaluation

Cloud, MLOps & DevOps: AWS (EC2, S3, Lambda, SageMaker, ECS, CloudWatch), Azure (Azure OpenAI, Azure Machine Learning, Azure DevOps, AKS), CI/CD Pipelines, GitHub Actions, Docker, Kubernetes, Git

ML & Deep Learning: Deep Learning, NLP, Computer Vision, Regression, Classification, Clustering, Predictive Analytics

Frameworks & Audio AI: PyTorch, TensorFlow, Keras, Scikit-learn, OpenCV, FastAPI, MusicGen, So-VITS-SVC

Backend & APIs: REST API Design, Scalable Microservices, Node.js, Audio/Video Processing Pipelines, Low-Latency Inference, Load Balancing

Web, SaaS & Languages: React.js, SaaS Architecture, Payment Integration, Python (Advanced), JavaScript, SQL

Databases & Data: Vector Databases (Pinecone, Chroma, FAISS), SQL (Oracle), Big Data Pipelines

AI Dev Workflow: Cursor, Claude Code, Antigravity, Ollama (local LLMs), Open Interpreter, agentic & prompt-driven development

PROFESSIONAL EXPERIENCE

Founder & Lead AI Engineer — Studygy.io | Dec 2025 – Present

Chantilly, VA · Live: studygy.io

- Grew Studygy.io to 1,000+ users as solo technical founder — an AI study platform that converts lectures and textbooks into Cornell Method notes (cues, notes, summaries), built and shipped end to end.
- Cut note-generation time by 95% by engineering the core RAG + LLM fine-tuning pipeline (FastAPI, Pinecone vector DB, OpenAI API), which has processed 700+ documents into structured Cornell Method notes.
- Architected a scalable AWS backend supporting hierarchical outlining, critical-reading tools, and real-time collaborative study workflows.
- Automated testing and deployment with Docker and GitHub Actions CI/CD pipelines, enabling multiple production releases per week with zero-downtime rollouts.
- Owned the full product lifecycle — model design, backend, frontend UX, go-to-market — and hosted the Studygy AI App Hackathon (80 participants).

AI Engineer — Unique Computing LLC | Jul – Dec 2025

Chicago, IL (Hybrid)

- Built an accessible learning platform on Open edX for children with disabilities — architected on AWS with LLM fine-tuning and RAG to deliver adaptive, personalized lessons.
- Processed 80+ GB of data across 3 enterprise clients by building production backend systems and REST APIs powering their Big Data analytics pipelines.
- Deployed and operated ML workloads across AWS and Azure (Azure OpenAI, Azure ML) with Azure DevOps CI/CD pipelines for automated model testing, versioning, and release.
- Integrated ML, NLP, and audio/video models into production systems, collaborating directly with PhD researchers and data scientists in a forward-deployed capacity.

AI Engineer (Internship) — Vocs AI | Mar – Jul 2025

Remote

- Engineered low-latency REST APIs integrating MusicGen and So-VITS-SVC models for real-time, production-scale lyrics-to-vocals audio synthesis.
- Sustained 100+ concurrent generation requests by load-balancing the audio-synthesis backend for high-throughput, production-scale inference.

Frontend Development Intern — Superior Code Labs | 2023

- Built responsive, cross-device web interfaces using React.js, HTML5, and CSS3.

KEY PROJECTS

- Easy Helpers — SaaS Service Booking Platform.** Independently built, launched, and monetized an on-demand home-services startup generating real revenue — payments, real-time booking, and admin dashboard on a scalable REST API + React.js stack.
- Generative AI & LLM Applications Suite.** RAG document Q&A system with long-term vector memory, hybrid retrieval, and reranking, plus a multi-turn conversational assistant — GPT-4, LangChain, FAISS; containerized with Docker, deployed via CI/CD.
- Farm-To-Table AgroTech AI.** ML supply-chain optimizer using regression and clustering to match farm supply with real-time market demand, cutting waste.

EDUCATION

Master of Science in Computer Science — Washington University of Science and Technology | Jan 2024 – Jun 2025 · GPA 3.92 / 4.0

Bachelor of Engineering in Computer Science — Visvesvaraya Technological University | Jul 2019 – May 2023 · CGPA 7.72 / 10.0

CERTIFICATIONS & RECOGNITION

AI Icon Award — Tech AI Magazine (2026) · Supervised Machine Learning — Stanford/Coursera (2025) · Software Engineer Certification — HackerRank (2025) · Generative AI for Everyone — DeepLearning.AI · Machine Learning with Python — IBM · Programming for Everybody (Python) — University of Michigan