UZMA FATIMA

Boston, MA | (857) 351-8464 | fatima.u@northeastern.edu | LinkedIn | GitHub | Portfolio

SUMMARY

AI Engineer specializing in production-grade LLM systems, hybrid retrieval pipelines, and scalable GenAI infrastructure. Designed and deployed RAG architectures, fine-tuned transformer models using LoRA/QLoRA and built LLM platforms with evaluation and latency optimization. Presented LLM platform at Google's Cambridge office. I build AI that ships

SKILLS

- **Programming & Data:** Python, C++, Pandas, NumPy, SQL
- **ML & AI:** PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, LSTM, ARIMA
- **NLP & LLMs:** RAG, Hugging Face Transformers, FAISS, LangChain, LangGraph, MCP, Prompt Engineering, LoRA/PEFT
- **Cloud & Engineering:** AWS S3, Docker, Git, CI/CD, MLOps, FastAPI, Gradio, Firebase, PostgreSQL, MySQL
- **Visualization:** Power BI, Tableau, Plotly, Streamlit, React

PROFESSIONAL EXPERIENCE

Cerebrone.ai | *AI Intern* — Jan 2026 - Present

- Designed and optimized hybrid **retrieval-augmented generation (RAG)** pipelines leveraging dense embeddings + **FAISS vector indexing**, improving retrieval **Recall@K** by 30% and reducing query latency by 40% in production **LLM workflows**
- Built **agentic AI** systems with **LangGraph** enabling multi-step reasoning and task orchestration, reducing manual workflow by over 60% across 3+ use cases
- Integrated **MCP server** architecture for scalable tool use and contextual memory persistence, enabling modular LLM deployment across full ingestion-to-generation pipelines and cutting integration overhead for new service

Northeastern University IST | *Data & Operations Analyst* — Sep 2025 - Dec 2025

- Engineered **SQL**-based **data pipelines** to process 1,000+ package records daily, cutting manual processing time by 50% and improving operational throughput
- Developed Python **anomaly detection** logic to flag misrouted packages, reducing delivery errors within the first month
- Built **Tableau KPI dashboard** surfacing package volume, missing item trends, and daily SLAs – adopted by IST leadership for ongoing operations reporting

ContIQ LLC | *AI Software Engineering Intern* — Jun 2025 - Aug 2025

- Architected a production-ready **hybrid RAG system** combining **dense embeddings**, semantic similarity search, and GPT-4 generation; implemented **chunking optimization** and embedding evaluation, reducing irrelevant retrieval by 30%
- **Fine-tuned** transformer-based LLMs using **LoRA/PEFT** adapters, reducing parameter overhead by 90% while improving task-specific response accuracy by 20% and maintaining <1.2s inference latency
- Deployed **agentic AI workflow** automating user-specific retrieval and outreach tasks, saving 5+ hours/week of manual communication work per creator

Nazra Software Solutions | *Data Science Intern* — Mar 2024 - Jul 2024

- Built **hybrid recommendation** engine (collaborative + content-based filtering) using Python and **scikit-learn**, directly increasing click-through rates by 18% and driving measurable revenue uplift
- Automated **ML retraining pipeline** and **A/B testing** framework across large-scale datasets; applied VADER **sentiment analysis** on 5,000+ review, reducing model refresh time by over half and surfacing product insights that directly informed roadmap decisions

Nexasoft Sdn Bhd | *IT Intern – Data Analytics* — Jun 2022 - Nov 2022

- Optimized SQL **ELT pipelines** processing 500K+ behavioral records, increasing reporting accuracy by 45% and reducing data latency by 30%; built **Tableau dashboards** that drove +40% stakeholder engagement

PROJECTS

EmoAid – AI-Powered Mental Health Assistant | GitHub — May 2025 - Jun 2025

- Delivered a full-stack mental wellness app featuring a GPT-4 **RAG** chatbot **(LangChain + FAISS + clinical literature)**, guided meditation and wellness exercises, and personalized mood-based recommendations - achieving 92%+ prompt accuracy and improving simulated user wellbeing scores by 28%
- Engineered real-time mood tracking and sentiment analysis pipeline (300+ entries, Firebase) and shipped cross-platform via **Flutter + CI/CD** with <0.8s AI response latency, enabling seamless daily mental health check-ins

Custom LLM Fine-Tuning & MLOps Platform | Github — Oct 2025 - Dec 2026

- Production-grade MLOps platform with QLora fine-tuning, **Ray + Kubernetes** distributed training, and automated evaluation pipeline covering accuracy, hallucination, drift, and safety, improving accuracy by 20%
- Presented at Google's Cambridge, MA office, recognized by industry practitioners for production **MLOps** architecture

Intelligent Demand Forecasting Platform | GitHub — Apr 2025 - May 2025

- Built an ensemble **time-series forecasting** system (PyTorch LSTM, XGBoost, LightGBM, ARIMA) with 100+ engineered features (**lags, Fourier transforms**), achieving MAPE of 4.78% and projecting $250K+ in business cost savings
- Shipped production-ready Dockerized Gradio dashboard on **Hugging Face** with **AWS S3** and **CI/CD pipelines**, enabling real-time demand predictions accessible to non-technical stakeholders

EDUCATION

Northeastern University, Boston, USA — Sep 2024 - Dec 2026

M.S., Artificial Intelligence

- **Coursework:** Foundations of AI, Natural Language Processing, Information Retrieval, Algorithms, Machine Learning, MLOps

Asia Pacific University, Malaysia — Sep 2019 - Dec 2023

B.S., Computer Engineering

LEADERSHIP & EXTRACURRICULARS

Graduate Student Government | *Student Affairs Senator*

- Representing 20,000+ graduate students at Northeastern University, advocating for student well-being and policy improvements

Student Advisory Board, Khoury College of Computer Science | *Board Member*

- Advising on student engagement, and career initiatives while representing the Boston campus and collaborating with Khoury's global board