Umme Athiya

Chicago, IL | ummeathiya2023@gmail.com | +1 (312)-684-9667 | LinkedIn | Portfolio | GitHub

Senior AI/ML Engineer

4+ Years Building Enterprise-Grade AI Systems | LLM Specialist | Scalable Recommendation Engines | OpenCV

Technical Skills

- **Programming Languages**: Python, Java, C++, SQL, NoSQL, MySQL, MongoDB
- **Generative AI:** LLM Fine-Tuning (Llama 2, GPT-4), RAG, Diffusion Models, RLHF, NLP
- **Machine Learning:** Deep Learning (PyTorch/TensorFlow), GNNs, Ensemble Methods
- **MLOps:** CI/CD for ML, Model Serving (TF Serving, Triton), Monitoring (Evidently, Prometheus)
- **Big Data:** Spark, Hadoop, Feature Stores, Vector Databases (Pinecone, FAISS)
- **Cloud AI:** AWS SageMaker, GCP Vertex AI, Azure ML, Kubeflow Pipelines
- **Data Analysis & Visualization:** Pandas, NumPy, Matplotlib, Tableau

Professional Experience

DePaul University, Chicago, United States **January 2024 - PRESENT**
AI/ ML Engineer

- Multi-modal recommenders fusing CLIP (images) and BERT (text) boosted CTR by 28% through enhanced content understanding.
- LLM cost optimization via TensorRT-LLM quantization inference costs by 40%, maintaining <500ms latency - 10K+ queries.
- Real-time fraud detection with YOLOv8+Kafka analyzed 5M+ transactions/month, preventing $2M+ in annual fraud losses.

IBM, Bangalore, India **February 2021 – August 2023**
AI/ ML Engineer – Generative AI and Robotics

- Built and optimized recommendation systems using **collaborative filtering and deep learning models**, **processing 1TB+** of data.
- Led team of 4 to build **enterprise RAG system** handling 10M+ queries/day with **92% answer accuracy** (vs. 78% baseline)
- Developed **automated model drift detection** system using KS tests and Evidently AI, reducing production incidents by **65%**
- Created **Spark-optimized feature engineering** pipeline processing 5TB/day of user behavior data with **<1% resource wastage**
- Developed **NLP models for text generation and summarization**, improving model accuracy by **10%**.
- Automated model deployment pipelines using **Docker and Jenkins**, reducing deployment time by **40%**.

Technocolabs, Bangalore, India **August 2020 – December 2020**
Machine Learning Intern

- Built Spark-based k-anonymity pipelines securing 2M+ records (83% PII risk reduction, 98% data utility).
- Optimized BigQuery via query/partitioning refinements, slashing feature gen time 4hrs→90min (4.5x faster).

Education

DePaul University, Chicago, Illinois **September 2023 – June 2025**
Master of Science in Computer Science **(GPA: 3.85/4.0)**
Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing, Artificial Intelligence, Data Science.

Don Bosco Institute of Technology, Bangalore, India **August 2016 - August 2020**
Bachelors in Information Science & Engineering **(GPA: 4.0/4.0)**

Projects

Privacy-Preserving LLM Platform | Llama 2, PySyft, Homomorphic Encryption

- **Breakthrough:** Fine-tuned 70B-parameter LLM with QLoRA using just 2x A100s (8x cost reduction).
- **Security:** Enabled FHE inference with <15% accuracy drop vs plaintext.
- **Adoption:** Deployed as AWS SageMaker endpoint serving 200+ enterprise users.

Enterprise GenAI Platform | **Strategic Impact**: $2.3M annual cost reduction

- Led cross-functional team to build **LLM orchestration layer** serving 50+ models (GPT-4, Llama 2, Claude)
- Implemented **dynamic model routing** achieving 40% cost savings while maintaining 99.9% SLA

National-Scale Computer Vision Deployment | **Scale**: 15M+ daily inferences

- Deployed **YOLOv9-based inspection system** across 8 manufacturing hubs.
- Built **federated learning pipeline** complying with EU/US/China data laws.
- Achieved **99.97% uptime** using Kubernetes + Istio service mesh.

Certifications

- NVIDIA Deep Learning Institute (DLI) – Accelerated AI with TensorRT, Large Language Model Deployment with Triton, TensorRT-LLM, AWS Certified Machine Learning Specialty (In progress), Microsoft AI-900, DeepLearning.AI TensorFlowDeveloper, Standford Machine Learning from Coursera, AWS Educate ML Fundamentals and Foundations.