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.

Umme Athiya

📍 Chicago, IL | ✉️ ummeathiya2023@gmail.com | 📞 (312) 684-9667

🔗 **LinkedIn** | 🖥️ **Portfolio** | 💻 **GitHub**

April 15, 2025.

Hiring Manager

Subject: Application for Senior AI/ML Engineer Position

Dear Hiring Manager,

I'm excited to apply for the Senior AI/ML Engineer role with **4+ years of experience** designing and deploying production-grade AI systems, I specialize in **scalable recommendation engines, LLM optimization, and real-time computer vision**—skills that align closely with your team's mission to build transformative AI solutions that bridge research and real-world impact, develop cutting-edge AI systems that solve complex business challenges and create intelligent systems that push the boundaries of what's possible with machine learning.

What I Bring to Your Team:

1. Proven AI Leadership

- Led cross-functional teams to deliver **high-impact AI products**, such as a multi-modal recommender (CLIP+BERT) that boosted engagement by **28%** and an enterprise RAG system handling **10M+ queries/day** at 92% accuracy.
- Pioneered **cost-efficient LLM deployments**, slashing inference costs by **40%** via TensorRT-LLM while maintaining **<500ms latency**—critical for real-time applications.

2. Full-Cycle ML Expertise

- Scaled **computer vision systems** to process **15M+ daily inferences** with **99.97% uptime**, leveraging Kubernetes and MLOps best practices.
- Automated **CI/CD pipelines** for ML models, reducing deployment time by **40%** and production incidents by **65%** through proactive drift detection.

3. Innovation in Generative AI

- Developed a **privacy-preserving LLM platform** (FHE + QLoRA) that enabled secure AI for **200+ enterprise users** without compromising performance.
- Fine-tuned **70B-parameter LLMs** with **8x cost reduction**, demonstrating expertise in large-scale model optimization.

Beyond Technical Execution:
What excites me most is bridging the gap between cutting-edge research and business impact. For example:

- At IBM, I **translated 3 academic papers on GNNs** into production features that generated **$800K in incremental revenue** within 6 months.
- Led an **AI ethics taskforce** that developed model cards and bias mitigation strategies now used company-wide, reducing fairness-related escalations by **65%.**
- Spearheaded **cost-benefit analyses** for AI projects, helping leadership prioritize a 70B LLM fine-tuning initiative that achieved **4x ROI.**

These experiences taught me that exceptional AI engineering requires:

- **Business acumen** to align models with organizational goals
- **Stakeholder management** to educate non-technical teams
- **Systems thinking** to anticipate downstream impacts

Why I'm Excited About This Opportunity

I thrive in environments that push the boundaries of AI while solving real-world problems. My experience spans **research, engineering, and deployment**, making me uniquely equipped to contribute to your team's goals.

I'd welcome the chance to discuss how my skills can support your organization's vision. Please feel free to schedule time at your convenience via email or phone. Thank you for your consideration—I look forward to the possibility of collaborating.

Best regards,
Umme Athiya