TANMAY JOSHI
Boston, MA | joshi.tanm@northeastern.edu | (857) 333-8661 | LinkedIn | GitHub

SUMMARY

Results-driven Software Engineer specializing in AI, focused on creating innovative Generative AI solutions and scalable MLOps frameworks that accelerate business objectives and enhance user experiences. Demonstrated success in architecting and deploying end-to-end AI systems that drive substantial operational efficiencies and user engagement, leveraging Python, Java, Kubernetes, and multi-cloud environments. Committed to advancing digital transformation by rapidly mastering emerging technologies and fostering collaborative innovation within agile teams to deliver high-impact results.

EDUCATION

Northeastern University, College of Engineering, Boston, MA — September 2023 – Present
Master of Science in Software Engineering Systems — *GPA: 3.80/4.0*

Institute of Technology, Nirma University, Ahmedabad, India — July 2019 – May 2023
Bachelor of Technology in Computer Science and Engineering — *CGPA: 8.02/10.0*

SKILLS

- **ML/LLM Stack**: Transformers, RAG, PyTorch, TensorFlow, LangChain, HuggingFace, OpenAI, LoRA/PEFT
- **Languages**: Python, Java, Bash, Spark
- **MLOps**: MLflow, SageMaker, Azure ML, Docker, Kubernetes, CI/CD (GitHub Actions)
- **Data/Cloud**: AWS (SageMaker, EC2, Lambda), Azure (AI Studio, OpenAI), Spark, Hadoop, MongoDB, Relational Databases

PROFESSIONAL EXPERIENCE

The American Board of Anesthesiology — August 2024 – Present
Machine Learning Engineer Intern — *Raleigh, USA*

- Deployed production RAG system (GPT-4 class) with MLflow for CI/CD & versioning; cut exam question costs 99% ($50K→$500), impacting 5K+ employees via a critical AI backend
- Led LLM fine-tuning (Q-LoRA/PEFT) on 5TB+ domain-specific corpora for production search, boosting retrieval precision 55% (ROUGE-1 0.29→0.45); validated gains via rigorous offline/online A/B tests, enhancing user search relevance
- Architected scalable multimodal evaluation platform (ASR/Whisper & LLM/GPT-4 scoring) automating feedback for 6K+ annual oral exams; cut manual review 80%, improving operational efficiency via an AI system calibrated to human benchmarks
- Deployed high-availability, low-latency ML inference APIs on Kubernetes (AKS) with product/SRE teams, reducing API deployment latency 40%; ensured 99.95% uptime for critical services at 2K+ QPS for key features using robust CI/CD & monitoring

Department of Research Computing (Northeastern University) — June 2024 – August 2024
Software Development Research Assistant — *Boston, USA*

- Implemented a data processing pipeline using PySpark and R for statistical analysis of 170M+ SLURM records, implementing time series modelling and predictive analytics that optimized GPU allocation by 25%
- Developed and containerized AI infrastructure in Linux using Docker and Podman, integrating Llama2 and Llama3 models with distributed computing frameworks in Python and Scala, reducing deployment time by 40%

IQSpatial — January 2023 – August 2023
Software Engineer — *New York, USA*

- Architected & launched RFP recommendation engine on AWS (EKS, SageMaker Endpoints, OpenSearch for vector search, DynamoDB); drove 100% product revenue uplift & sub-second latency via semantic matching (Transformer embeddings) & content filtering
- Built & automated MLOps pipeline (Airflow, MLflow, SageMaker Pipelines) for model CI/CD & real-time data ingestion from 200+ sources (10K+ docs daily), ensuring high model freshness
- Led end-to-end development of recommendation algorithms on AWS SageMaker (two-tower models, LambdaMART for re-ranking); improved relevance 35% (NDCG@10, Recall@50 uplift) & business metrics (CTR, conversion) via A/B testing
- Engineered data pipeline on AWS (EMR/Spark, S3 data lake, Glue, Kinesis) using Python (Selenium/BeautifulSoup) for high-volume RFP extraction & processing; ensured 99.95%+ uptime & data integrity with CloudWatch, X-Ray, & Prometheus monitoring

ACADEMIC PROJECTS

Draftly – AI Based UI/UX Design Assitant

- Built an intelligent UI/UX assistant with GPT-4o, crafting advanced prompts to convert natural language into production-ready code—cutting design-to-implementation time by 80% and enabling non-technical users to contribute without coding
- Developed Draftly, a full-stack React app with Tailwind, Express.js, and PostgreSQL that transforms text prompts into responsive UI code (HTML/CSS,React, Tailwind) with 95% accuracy and real-time previews
- Engineered a context-aware AI agent recognizing design patterns, accessibility needs, and UX principles; added responsive previews and intelligentcaching, reducing API costs by 40% and ensuring <2s response times

SmartMed: AI-Powered Medical Note Summarization for Faster Diagnoses

- Streamlined the diagnostic process by reducing the time physicians spend manually reviewing lengthy patient notes.
- Designed an automated system using HuggingFace to fine-tune Llama2-7B for extracting 25+ medical keywords per note. Leveraged advanced techniques like 4-bit quantization, PEFT and prompt engineering to enhance inference performance by 8%
- Simplified physician workflows with keyword-based summaries, reducing manual review time. Improved ROUGE scores by 0.4, precision for longer inputs by 4x, and delivered a scalable, user-friendly Streamlit UI

DayCare – Management System

- Developed a comprehensive DayCare Management System using Java and MVC architecture, managing over 30 students per classroom with automated age-group categorization supporting up to 3 teachers per classroom
- Designed and implemented a user-friendly GUI interface using Java Swing, featuring 4+ interactive dashboards (Student, Teacher, Summary, and Vaccine Management) with real-time data updates and automated error handling for improved administrative efficiency
- Utilized object-oriented programming principles and design patterns (Factory Pattern, Singleton) to create a modular codebase, resulting in a 40% reduction in code duplication and enabling seamless integration of new features like vaccine tracking and student performance monitoring