Tarush Singh

Cincinnati, OH | singh.tarus@northeastern.edu | +1 (513) 857-8501
LinkedIn: linkedin.com/in/tarush-singh-246144113 | GitHub: github.com/TarushS-1996 | Portfolio: tarushs.streamlit.app

SUMMARY

AI/ML Engineer with 4+ years of experience designing and deploying production-ready machine learning, distributed systems, and NLP solutions across e-commerce, mobility, and cloud. Strong expertise in Java, Python, RESTful APIs, and microservices architecture with hands-on work in transformers, attention mechanisms, and embedding-based architectures. Skilled at leading cross-functional teams, building scalable AI/enterprise pipelines, and deploying solutions in cloud-native environments (AWS, GCP, Kubernetes).

EXPERIENCE

Senior Machine Learning Engineer
FinOpsly Jun 2025 - Present, Cincinnati, OH

- Architected distributed AI agent pipelines using LangChain, FastAPI, and REST APIs for Snowflake data workflows of 200,000+ rows, increasing throughput by 3x and surfacing cost anomalies valued at $200K+ annually.
- Engineered scalable memory management (FAISS), improving retrieval precision to 95%+ and enabling session histories with 50+ queries per user.
- Implemented ML-driven preprocessing on retrieved data, reducing token usage by 65% while maintaining complete coverage and cutting latency by 50%.
- Optimized AI evaluation (A/B testing, bias/variance checks), increasing accuracy by 35% for domain-specific questions.
- Mentored two engineers in MLOps and scalable design, reducing onboarding time by 30% and raising team code coverage from 72% to 90%.

AI Engineer – Graph-Aware LLM/LMM Systems
Modlee AI Jul 2024 – Apr 2025, Remote

- Designed graph-based RAG agents with Neo4j and LangChain, boosting retrieval accuracy by 25% and increasing client ticket throughput by 20%.
- Refined pipelines with feedback-driven chunk re-ranking, cutting hallucinations by 20% and improving client-facing output quality.
- Integrated surrogate ML models for retrieved data filtering, improving relevancy by 15% while reducing compute overhead.
- Collaborated with cross-functional teams, accelerating project delivery by 30% and raising adoption rates above 85%.

NLP Development Engineer
HappSales Pvt Ltd Sep 2019 - Apr 2022, India

- Deployed RASA/BERT-based NER agents for 20+ enterprise clients, improving input accuracy by 15% and reducing manual data entry by 30%.
- Automated deployment of 10+ microservices using AWS Lambda and Docker, cutting release cycles by 40% and ensuring downtime under 1 hour/month.
- Built continuous monitoring dashboards, reducing bug resolution time by 25% and improving monthly performance reviews.

PROJECTS

Dominex – Drug Lifecycle Management Application (Java)

- Developed a full-stack application in Java/Swing with SQLite backend and JavaMail API to track pharmaceutical drug development from discovery to FDA approval.
- Designed system architecture with class and sequence diagrams, ensuring clear modularity and event-driven data flow.
- Built reporting workflows enabling researchers to track stage progress and generate compliance-ready documentation. [GitHub]

Fingen Insights Analytics Platform

- Scaled a multi-agent RAG/analytics system (LangChain, LlamaIndex, OpenAI, vLLM, Snowflake), integrating tabular, text, audio, and image data.
- Reduced query latency by 60% and delivered automated dashboards adopted by 5+ departments. [GitHub]

Audio Sentiment Analysis (HPPC)

- Built OpenCV-based model detecting stress/anxiety with 98% accuracy, optimized via Weights&Biases sweeps.
- Reduced training time by 60% with automated hyperparameter tuning. [GitHub]

AutoClass: Agentic Method Controller via MCP

- Devised a docstring-introspection framework for automated mapping of natural-language queries to class methods.
- Built a dependency-aware execution engine, achieving 3x faster task completion. [GitHub]

Model Evaluation & Visualization Toolkit

- Created interactive dashboards (Tableau, SHAP, BLEU, ROUGE) for explainability, cutting debugging time by 35%.
- Supported evaluation of 20+ model deployments. [Medium]

EDUCATION

M.S. in Information Systems
Northeastern University Boston, MA Apr 2024

- Relevant Coursework: Neural Networks Architecture, Parallel ML, Data Engineering, Java Enterprise Development

B.Tech in Electronics and Communication Engineering
SRM Institute of Science and Technology Chennai, India Apr 2018

- Relevant Coursework: Soft Computing, Intro to Robotics, Data Engineering

CERTIFICATIONS

Deep Learning Specialization by Andrew Ng
Coursera

SKILLS

Java (Swing, REST APIs, multithreading), Python (Pandas, NumPy, scikit-learn), SQL (Postgres, MySQL), R, Bash

Spring Boot, Kafka, PyTorch, TensorFlow, Hugging Face, LangChain, LlamaIndex, vLLM, OpenAI, BERT, GPT, RLHF, surrogate modeling, NLP, microservices, event-driven systems

AWS (SageMaker, Lambda, EMR, EKS), GCP (Vertex AI), Docker, Kubernetes, Ray, Airflow, MLFlow, Databricks

Tableau, Dash, PowerBI, matplotlib, seaborn, business intelligence dashboards