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-1996.github.io

SUMMARY

AI/ML Engineer with 4+ years of experience designing and deploying production-ready machine learning, surrogate modeling, NLP, and agentic AI solutions across e-commerce, mobility, and cloud. Strong expertise in transformers, attention mechanisms, and embedding-based architectures for both structured and unstructured data. Familiar with building and adapting domain-specific language models to meet enterprise goals. Proven ability to translate business needs into explainable, production-scale AI solutions.

EXPERIENCE

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

- Developed agentic AI pipelines using LangChain and FastAPI to automate Snowflake CUR data retrieval for 200,000+ cloud records, increasing analytics throughput by 3x and surfacing cost anomalies valued at $200K+ annually.
- Configured vector-based memory modules (FAISS), improving historical retrieval precision from 80% to 95% and supporting 50+ concurrent user sessions.
- Conceived adaptive pre-processing models to reduce LLM token usage by 65%, halving latency for analytics queries and cutting operational costs by 50%.
- Enhanced evaluation workflows (A/B testing, bias/variance checks), boosting domain-specific model accuracy by 35% and reducing false positive rates by 20%.
- Guided and supported two engineers in MLOps and CI/CD best practices, reducing onboarding time by 30% and improving team code coverage from 72% to 90%.

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

- Built graph-driven RAG agents using Neo4j and LangChain to power traversal-based retrieval, increasing retrieval accuracy by 25% and improving client ticket throughput by 20%.
- Refined pipeline orchestration via feedback-driven chunk reranking, reducing LLM hallucinations by 20% and enhancing client response accuracy across 5+ deployments.
- Integrated surrogate model scoring for chunk prioritization, improving context relevance by 15% and response latency by 40%.
- Partnered with five cross-functional teams to deploy agentic ML systems, accelerating project delivery timelines by 30% and achieving 85% stakeholder adoption across pilots.

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

- Delivered RASA/BERT-powered NER and voice CRM agents for 20+ enterprise clients, improving entity recognition accuracy by 15% and saving 1,000+ manual data entry hours annually.
- Instituted deployment of 10+ NLP microservices using AWS Lambda and Docker, cutting release cycles by 40% and maintaining 99.9% service uptime.
- Devised model monitoring dashboards with real-time KPIs, achieving a 25% reduction in bug resolution time and enabling data-driven performance reviews.
- Collaborated with cross-functional business teams to integrate voice analytics with CRM workflows, increasing lead conversion efficiency by 18% and improving data capture accuracy across client deployments.

PROJECTS

Fingen Insights Analytics Platform

- Crafted and scaled a multi-agent RAG/analytics system (LangChain, LlamaIndex, OpenAI, vLLM, Snowflake), integrating tabular, text, audio, and image data for enterprise analytics and anomaly detection.
- Achieved a 60% reduction in query latency and enabled mechanized dashboard reporting adopted by 5+ departments, streamlining analytics for 30+ users. [GitHub]

Audio sentimenet analysis (HPPC)

- Engineered image-based Audio sentiment model (OpenCV) with 98% detection stress anxiety and other emotional states.
- Optimized model accuracy by 15% through Weights&Biases Sweep hyperparameter optimization and fine-tuning, eliminating manual hyperparameter tuning and reducing training time by 60%. [GitHub]

AutoClass: Agentic Method Controller via MCP

- Formulated a docstring-introspection framework that achieved 100% automated mapping of natural-language queries to relevant class methods, reducing manual orchestration overhead to zero.
- Developed a dependency-aware execution engine that pipelined method outputs into subsequent inputs, enabling seamless multi-step reasoning with 3x faster end-to-end task completion. [GitHub]

Model Evaluation & Visualization Toolkit

- Created an interactive dashboard suite (Tableau, SHAP, BLEU, ROUGE) for LLM/ML explainability, enabling feedback loops and data-driven stakeholder decisions.
- Supported model evaluation for 20+ deployments, reducing model debugging time by 35%. [Medium]

EDUCATION

M.S. in Information Systems

Northeastern University Boston, MA Apr 2024

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

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

Python (Pandas, NumPy, scikit-learn), SQL (Postgres, MySQL), R, Bash

PyTorch, TensorFlow, Hugging Face, LangChain, LlamaIndex, vLLM, OpenAI, BERT, Llama, GPT, LORA, RLHF, surrogate modeling, NLP, time series, anomaly detection, computer vision (OpenCV), SHAP, BLEU, ROUGE

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

Tableau, Dash, PowerBI, matplotlib, seaborn, business intelligence dashboards, Machine Learning (XGBoost, Clustering, SVM, Decision trees)