Tarush Singh

+1 (513) 857 8501 - Portfolio - singh.tarus@northeastern.edu - linkedin.com/in/tarush-singh-246144113/ - github.com/TarushS-1996 - Medium

SUMMARY

AI Engineer with 3+ years of experience and hands-on work building scalable Deep Learning, Machine Learning, GenAI and LLM applications. Skilled in Python, TensorFlow, FastAPI, LangChain, and vector databases. Proven track record improving accuracy and efficiency via ML preprocessing and RAG techniques.

WORK EXPERIENCE

AI Engineer

Modlee AI LLC (Remote) *Jul 2024 – Apr 2025*

- Designed and deployed scalable, full-stack GenAI pipelines using LangChain, Milvus, and Neo4j to enable intelligent document retrieval, summarization, and structured insight extraction.
- Improved LLM output accuracy by over 25% through chunk-level tagging, prompt engineering, and dynamic context optimization strategies tailored to domain-specific data.
- Built FastAPI-powered REST agents integrated with structured databases to support real-time querying, enabling LLMs to generate domain-grounded summaries for financial analytics.
- Built an MCP-based ML preprocessing layer that auto-selected clustering, statistical analysis or PCA based on query and data, reducing token usage by 75% and hallucinations through domain-aware grounding.
- Engineered a Neo4j-driven adaptive reasoning system that mimicked RL-style feedback, dynamically adjusting LLM extraction logic based on mapped industry-specific terminology graphs—boosting precision by 20% on domain-aligned queries.

NLP Development Engineer

HappSales pvt ltd, Bengaluru, India *Sep 2019 - Apr 2022*

- Spearheaded a NER system powered by RASA, Word2Vec and BERT, enabling voice-driven CRUD operations within CRM systems, streamlining data management and reducing processing time by 30%.
- Engineered intuitive interfaces intertwining speech-to-text, resulting in a 20% increase in user engagement and a 15% reduction in user error rates.
- Integrated the backend NLP engine with AWS Lambda microservices for efficient and scalable hosting going from 100+ users for our SaaS based suite.
- Streamlined NLP integration by 40%, driving product excellence and cross-functional collaboration.
- Contributed to a 10-15% increase in user productivity by streamlining workflows and minimizing clicksDemo

TECHNICAL SKILLS

Programming Languages:	Python, C/C++, Java (JVM), HTML, CSS, SQL, Matlab, JavaScript
Cloud & Tools:	AWS (Lambda, EC2, S3), Azure (Functions), Git, Docker, GCP (Vertex AI), Postman
Libraries & Frameworks:	PyTorch, TensorFlow, Keras, Scikit-learn, Pandas, NumPy, LangChain, Flask, OpenCV, Jupyter, PySpark, Matplotlib, Plotly, FastAPI
ML Architectures:	CNN (YOLO, ResNet, Inception), Transformers (VLM, BERT, SAM), RNN (LSTM, RCNN), ML (SVM, KNN, Decision Tree, Random Forest, XGBoost), AI Agents (OpenAI, Langchain, MCP, AutoGPT, BentoML, RAY)
ML Ops & Deployment:	Model reproducibility(W&B, MLFlow), ONNX, TensorRT, CI/CD for ML pipelines, Terraform, Kubernetes

EDUCATION

Northeastern University

Boston, USA

Masters of Science in Information Systems *Completed - Apr 2024*

- Relevant Courses: Neural Network Architecture, High Performance Parallel ML, Data Science engineering

SRM Institute of Science and Technology

Chennai, India

Electronics and Communication engineering *Completed - Apr 2018*

- Relevant Courses: Intro to Robotics, Data structures and Algorithms, Soft Computing

PROJECTS

Financial Insights RAG — Feb 2024 - Mar 2024

- Led development of a RAG-based LLM system using OpenAI's models and LlamaIndex, integrated via Streamlit for intuitive UI.
- Improved multi-source data retrieval efficiency by 20%, handling PDF documents and Snowflake queries. GitHub

Multi-GPU audio sentiment analysis — Jan 2024 - Apr 2024

- Engineered a distributed training pipeline using PyTorch DDP with mixed precision (CUDA, autocast, pinned memory) → Achieved 40% training speedup and enhanced GPU utilization for emotion classification on audio datasets.
- Processed speech data into recurrence plots and trained CNNs for emotion classification (stress, happy, sad), achieving 70% accuracy.
- Deployed optimized model with TensorRT and ONNX for real-time psychological impact assessment at scale GitHub

Agricultural Policy Recommendation AI Agent — Jun 2023

- Developed AI agent pipeline integrating satellite imagery, demand/supply models, and weather reports into LLM-driven policy recommendations
- Built multi-modal surrogate model ensemble and deployed through an LLM for real-world agricultural insights GitHub

Weight&Biases Model Analysis — Jan 2024 - Feb 2024

- Integrated Weights&Biases for model lifecycle tracking and hyperparameter tuning → Boosted CV model accuracy by 10% and streamlined experiment reproducibility across multiple runs.
- Optimized computer vision model to 92% accuracy using the sweep method and provided tutorial article. Medium

Model Evaluation and Parameter Mapping — Feb 2023 - Apr 2023

- Conducted exploratory, statistical data analysis on attrition datasets, comparing multiple regression predictive modeling techniques like decision tree, naive Bayes, and XGBoost models (82% accuracy).
- Improved model predictive behavior to 94% by incorporating performance benchmarking using SHAP and authored findings. Medium