Tarush Singh

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

SUMMARY

Experienced Machine Learning Engineer with 3+ years of expertise in NLP, Computer Vision, Deep Learning, and ML/DL Optimization. Proven track record in delivering innovative solutions and driving productivity through streamlined workflows.

EDUCATION

Northeastern University **Boston, USA**
Masters of Science in Information Systems *Sep 2022 - 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 *Aug 2014 - Apr 2018*

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

TECHNICAL SKILLS

Programming Languages:	Python, C/C++, Java, HTML, SQL, Matlab
Cloud & Tools:	AWS (Lambda, EC2, S3), Azure (Functions), Git, Docker, GCP (GCF), Postman
Libraries & Frameworks:	Tensorflow, PyTorch, Scikit-learn, Pandas, NumPy, OpenCV, Flask, Jupyter Notebook, W&B
ML Architectures:	CNN (YOLO, ResNet, Inception), Transformers (Vision Transformers, BERT), RNN (LSTM, RCNN), ML (SVM, KNN, Decision Tree, XGBoost)

WORK EXPERIENCE

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

- Developed multi-stage AI agent pipeline for retrieval-augmented generation (RAG), utilizing a custom tagging system to enrich context windows for more relevant and accurate LLM outputs. Improved extraction performance from 65–70% to 86–95% for context-matched queries.
- Streamlined the response generation structure to 100% using prompt templates resulting in consistent response structure for RAG
- Engineered autonomous agents that dynamically queried the Cube API to determine, formulate, and execute REST API calls based on user prompts and data requirements.
- Designed an ML-driven statistical analysis pipeline to summarize large-scale tabular API responses, reducing LLM context overload and improving response quality.
- Integrated clustering, cyclic encoding, and temporal context techniques to refine agent decision-making and boost response relevance over iterative prompts.
- Implemented a knowledge graph (KG)-based reasoning layer into AI agents, allowing them to utilize linked contextual entities for deeper insight and long-term memory.

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 for efficient and scalable hosting going from 100+ users to as required.
- 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

PROJECTS

Financial Insights RAG **Feb 2024 - Mar 2024**

- Led development of RAG LLM system, leveraging OpenAI's LLM and LlamaIndex tool with streamlit for intuitive UI.
- Achieved 20% increase in data retrieval efficiency, handling diverse data sources (PDF or Snowflake DB). GitHub

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

- Engineered multi-GPU training pipeline with PyTorch's DDP, boosting speed by 40% and emotion detection accuracy to 70%
- Deployed optimized model with TensorRT and ONNX for real-time psychological impact assessment at scale GitHub

Weight&Biases Model Analysis **Jan 2024 - Feb 2024**

- Explored Weight&Biases for deep learning, resulting in a 10% increase in model accuracy for smile detection in images.
- Optimized CNN model to 92% accuracy using the sweep method and provided tutorial article. Medium

Model Evaluation and Parameter Mapping **Feb 2023 - Apr 2023**

- Conducted exploratory data analysis on attrition datasets, comparing decision tree, naive Bayes, and XGBoost models (82% accuracy).
- Improved model predictive behavior to 94% accuracy using SHAP and authored findings. Medium