Lakshay Tyagi

Software Developer and Research Engineer building cutting-edge AI solutions at Amazon Q
My work revolves around building Search, Large Language Model (LLM), Computer Vision, Machine Learning, Retrieval-Augmented Generation (RAG), and Distributed Database Solutions

📧 lakshay.tyagi.nyu@gmail.com
📱 (+1) 551-344-7010
🐙 Github
🎓Google Scholar
💼LinkedIn
🌐 Personal Website

Education

New York University, Courant Institute of Mathematical Sciences — Sep 2022 - May 2024
MS in Computer Science; GPA: 3.97/4.0; Fellowship for exceptional academic performance — New York City, NY

Indian Institute of Technology, Kanpur (IITK) — Jul 2017 - May 2022
BTech, Double Major Electrical Engineering; GPA: 9.5/10.0; General Proficiency Medal — Kanpur, UP

Work Experience

Amazon Web Services (AWS) - Software Development Engineer - AI/ML — Jul 2024 - PRESENT

- Designed query and storage capacity for vector databases optimized for Retrieval Augmented Generation (RAG) purposes
- Helped launch GenAI search index in Amazon Bedrock contributing to CRUD APIs in backend (Java) and frontend (TS)
- Optimized ElasticSearch clusters for enterprise search with reduction in latency by 47% and savings of >150,000$ per year
- Revamped CloudFormation CI/CD pipelines to reduce operations overhead by 23% and improve deployment frequency

NYU Langone Institute - Research Engineer - Publication — Oct 2023 - May 2024

- Developed Large Language Model (LLM) solutions for labelling medical reports with pathologies with explanations
- Pretrained BERT-based models on a corpus of 3,400,000 medical reports with Masked Language Modelling
- Engineered Coreset-based Active Learning algorithms to select a corpus of 11,000 reports for labelling with a teacher model
- Trained LLAMA-2 using LORA with distributed GPU training in **PyTorch** to improve Macro F1 score to 0.89

Internships

Bayer, AI Team - Data Science Intern — Jun 2023 - Sep 2023

- Orchestrated time series forecasting pipelines in Databricks for predicting sales growth, integrated with SnowFlake using SQL
- Engineered features and trained XGBoost, LSTM (RNN), and graph models to reduce MAPE to 0.11 on backtests
- Developed a web application in Flask for visualizing forecasts for >100 million dollar sales of 20 product segments

Mitacs Globalink - Research Intern - Publication — Jun 2023 - May 2024

- Set up a distributed training framework in PyTorch for UNet-based federated learning models for brain tumor segmentation
- Achieved Dice Similarity Coefficient of 0.674 using variable local tuning of client parameters with learning rate scheduling

Samsung Research Institute - Software Engineering Intern — May 2020 - Jul 2020

- Implemented Autoencoder-based deep learning models in TensorFlow for denoising low light, grainy videos
- Delivered State of the Art SSIM of 0.89 and PSNR of 30.14 for video generation using perceptual loss and sensor input

Technical Skills

Programming Languages: Python, Java, C/C++, Ruby, Scala, TypeScript; **Databases:** DynamoDB, ElasticSearch, Snowflake, Postgres
Cloud Technologies: AWS, DataBricks, Kubernetes, Google Cloud Platform, Hadoop, Spark, Grafana **Software:** MATLAB, R
Libraries: PyTorch, Tensorflow, Keras, NumPy, Scikit-Learn, OpenCV, pandas, HuggingFace, Spring/Boot, cvxpy, Scikit-Learn, NLTK

Publications

- Humor@IITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness A Gupta*, **L Tyagi***, A Pal*, B Khurana*, A Modi: Publication in Association of Comp. Linguistics (ACL) : ACL-ICJNLP 2021
- Federated Learning Using Variable Local Training for Brain Tumor Segmentation: A. Tuladhar, **L. Tyagi,** R. Souza, ND. Forkert, BrainLes MICCAI 2021