Aman Singhal

linkedin.com/in/amansinghalml | +1 (551)-344-6364 | github.com/AmanSinghal927| aman.singhal@nyu.edu

EDUCATION

New York University New York, NY
Master of Science, Computer Science Sep 2022 - May 2024
Natural Language Processing (NLP), Large Language Models, Artificial Intelligence, Algorithms, Data Structures, Optimization

Delhi Technological University Delhi, India
Bachelor of Technology, Electronics and Communication Engineering Aug 2014 - May 2018
Computer Vision, Machine Learning, Mathematics (Calculus, Linear Algebra, Probability, Statistics), Signal Processing

EXPERIENCE

Research Scientist, Together.ai., NY, USA Sep 2024 – Present

- Engineered datasets, pre-training & fine-tuning recipe; trained LLMs for spec-decoding to boost 8-40% TPS across regimes
- Profiled vLLM & TensorRT R1 multi-token prediction on Blackwell B200s; enabled in-house inference engine deployment

Data Science Intern, Chegg Inc., CA, USA May 2023 – Aug 2023

- Architected RLHF post-training pipeline for edtech LLM, boosted alignment to drive customer satisfaction score to 1.2x
- Spearheaded hackathon team, built Retrieval Augmented Generation (RAG) study group recommendation, winning runnerup

Data Scientist, UnitedHealth, Delhi, India | *Awards* July 2021 – July 2022

- Pioneered information extraction using multimodal data to medical & legal; automated workflows for pharmacy benefits
- Directed 3-person team, delivering cosine-based search results, T5 summaries & BERT named entities for 10k daily queries

Research Assistant, IIIT Hyderabad (Prof. Vinay Namboodiri & Prof. C V Jawahar) | *Papers* | *Code* Aug 2020 – Mar 2021

- Co-authored computationally inexpensive transfer learning research for low-resource domain adaptation & neural machine translation. Achieved state-of-the-art +18.0 BLEU over existing deep learning algorithms
- Open-sourced semi-supervised synthetically generated dataset for language translation, advancing Generative AI research

Data Scientist, TransOrg Analytics, Gurugram, India | *Papers* | *Code* June 2018 – Aug 2020

- Automated loan underwriting, unsupervised fraud prevention and economics models for insurance risk at American Express
- Deployed resource allocation regression forests using exploratory analytics & big data mining for the Montana state govt.

SKILLS

Deep learning: NLP (OpenAI, HuggingFace, spaCy), AI/ML (PyTorch, Axolotl, VLLM, SGLang, Megatron-LM), WandB
Machine Learning: Numpy, Pandas, Jupyter, Scikit-Learn, data vizualization (Tableau, Matplotlib), Statistical Learning (SciPy)
Languages: Python, C++, SQL/ MySQL, R, scala, apache spark, Hive, latex, ui/ux (javascript), business intelligence (tableau)
Software: Kubernetes, Cloud (aws, azure), Docker, Git, PineCone, Kubeflow, deepspeed, TF-Serving, Apache, ETL (hadoop)

PROJECTS

Pretraining & Finetuning Pipeline, Together AI: Optimized draft model GPU utilization and delivered leaderboard inference throughput for high-frequency targets (Llama 70B/405B) on Artificial Analysis

- Pioneered layer pruning algorithm for speculator models & reduced depth by 50% while matching baseline acceptance rate
- Optimized pretraining data mix; scaled from 3T to 1T tokens through synthetic data curation reducing training time by 33%
- Developed production traffic evaluation & ablated hyperparameters; created recipe also applied to reasoning models (Qwen)

Training Framework Optimization, Together AI: Architected features to enable memory-efficient in-house finetuning engine

- Integrated sequence parallelism & flex attention; delivered long-context model training for leading agentic code editor client
- Engineered activation streaming, enabling distillation from large LLMs (R1, Kimi K2) without disk storage bottlenecks

Reinforcement learning from Human Feedback, Chegg Inc: Enhanced science-benchmark correctness on 8B models by 10-20% using in-house preference data, Proximal Policy Optimization (PPO) and reward hacking mitigation

- Reduced training iteration time by leveraging parameter efficient finetuning LoRA, FlashAttention, and DeepSpeed
- Architected mixture-of-agents automated evaluation framework; trained Flan-T5/DeBERTa reward models to 75% accuracy

Semantic Search, UnitedHealth: Led UI, Python APIs for question-answering web app, querying 1B documents with 9s p50

- Elevated result relevance via topic modeling through embeddings, PCA dimensionality reduction and k-means clustering
- Engineered TensorFlow Serving RNN summarization models and HuggingFace BERT for Named Entity Recognition
- Deployed Azure OCR & YOLO object detection backend; orchestrated ETL pipelines via NoSQL database with Docker

Audio Analytics, IIT Delhi (code): Led signal processing and multi-modal feature engineering for non-profit social platform

- Automated xgboost, CNN & ResNet speech classification; developed AWS pipelines for real-time content moderation
- Accelerated time-cost saving with hyperparameter tuning, data augmentation, feature engineering; scaled to 1M households

PUBLICATIONS & OPEN-SOURCE CONTRIBUTIONS

- **Published** Exploring Pairwise NMT for Indian Languages research at the ICON conference, ACL Anthology link
- **Submitted** “Kitty: Accurate and Efficient KV Cache Quantization with Dynamic Channel-wise Precision Boost”, MLSys’26
- Developed leaderboard neural deep learning transformer models at the **workshop** for language translation (WAT’20 - link)
- Directed cross functional team & led requirement gathering, communicating business intelligence to leadership at TransOrg
- Teaching assistant for graduate-level applied mathematics & artificial intelligence at Computer Science Department, NYU