Nafis Neehal

nafisneehal95@gmail.com | (518) 805-8633 | Github | LinkedIn | Website

Professional Summary

Machine Learning Engineer specializing and interested in end-to-end ML/LLM system development and MLOps, with proven experience in building scalable AI solutions for real-world business problems across diverse domains.

Technical Skills

LLM Expertise: RAG, Fine-tuning (SFT/PEFT), Quantization, Prompt Engineering, Benchmarking, GraphRAG
ML/DL/Causal: PyTorch, DDP, TensorFlow, Scikit-learn, DeepSpeed, AutoML, OpenCV, SHAP, EconML, DoWhy
MLOps Stack: MLflow, Docker, CI/CD, ChromaDB, Hopsworks, PySpark, W&B, AWS (SageMaker, Lambda, EC2)
LLM Frameworks: LangChain, LlamaIndex, HF Transformers, Axolotl, Unsloth, Autotrain, Comet, PromptHub
Languages & DB: Python, SQL, R, C++, Neo4j, Google Firestore (NoSQL), MySQL, SQLite
Data Visualization and Others: Streamlit, Gradio, R-Shiny, Tableau

Research & Development Experience

IBM - RPI Research Projects Feb 2022 – Present
Graduate Research Assistant, Team Lead — (Funding: IBM HEALS Project) *Troy, NY*

- **Building "TrialBrain" - LLM Augmented clinical trial automation framework** [ArXiv] [Github] [HF]
  - Released Llama-3.2-3B models fine-tuned (PEFT) on 65k+ clinical trials specializing on feature generation task
  - Created "CT-Bench" (1700 trials, 1.6k+ medical conditions) for benchmarking LLMs in feature generation task
  - Identified 3 types of Hallucination in our task and implemented novel hallucination-adjusted metrics for GPT-4/LLaMA-70B evaluations [IEEE BigData'24 (to appear)]
  - Developed end-to-end trial feature generation pipeline achieving 18 percentage point improvement over baseline with RAG-based few-shot examples and 0.85 Cohen's Kappa with human experts [Under Review in ARR]
- **Architected "FRESCA" - a fairness-aware ML-Based patient recommendation framework** [Github]
  - Engineered novel ML-based patient recommendation improving treatment effect estimation accuracy by 4% while reducing demographic disparity by 96%
  - Implemented two-staged trial design framework potentially reducing control patient recruitment cost upto 49%
  - Published in top ML/Healthcare venues, won awards - [RecSys'24] [AMIA'23] [SCT'23 (Best Poster Award)]

CDPHP - RPI Industrial Research Projects May 2020 – Jan 2022
Graduate Research Assistant — (Funding: CDPHP Industrial Research Grant) *Troy, NY*

- **Type-2 Diabetes Health Management Program Evaluation using Machine Learning** [HIMS'22]
  - Developed deep autoencoder for patient matching (35% faster, 40% memory reduction) and multi-stage survival analysis algorithm for health outcome tracking
  - Optimized 9M+ patient record processing pipeline using PySpark/AWS achieving 60% faster processing time
- **Improving Targeted Intervention using Machine Learning** [BIBM'22]
  - Built hybrid ML framework using novel clustering algorithm to identify 3 distinct patient groups from imbalanced dataset (350K control vs 1.6K treatment cases), improving treatment personalization
  - Implemented nearest-neighbor and exact matching techniques for unbiased treatment effect estimation
- **High-risk patient identification using Machine Learning** [BIBM'21] [Github (Non-Proprietary)]
  - Engineered ML pipeline processing 22.5M+ temporal records achieving 95% physician agreement and 30% early detection rate
  - Created hybrid preprocessing pipeline with PCA and downsampling for extreme class imbalance (0.5% positive class), achieving 200x efficiency gain through temporal windowing and sparse data handling

Open-Source Projects (Selected)

BanglaLLM [HuggingFace]: Developing fine-tuned open-source LLMs for reasoning and factual analysis in Bengali
MAMA-GPT [Github]: GPT-4 powered Bengali Voicebot integrating real-time STT/TTS and bidirectional translation
Cerebro [Github][Demo]: Fast, lightweight search engine for exploring papers from major AI/ML venues + ArXiv
Trade-Mind [Github] [Demo]: End-to-end MLOps pipeline for BTC/USD prediction using real-time OHLC features
ChanBOT [Github]: Fine-tuned Llama3.1-8B using PEFT and 4-bit quantization to mimic a fictional TV character

Education

Rensselaer Polytechnic Institute Sep 2019 – May 2025 (Expected)
Ph.D. in Computer Science - ML Systems/LLM/Healthcare, GPA: 3.74 *Troy, NY*
Rensselaer Polytechnic Institute Sep 2019 – May 2021
M.S. in Computer Science - ML Systems/Healthcare, GPA: 3.68 *Troy, NY*