Vashisth Tiwari

vashistt@andrew.cmu.edu ❖ (585) 524-8385 ❖ www.linkedin.com/in/vashistht/ ❖ https://github.com/Vashistht

EDUCATION

Carnegie Mellon University **Aug. 2023 – Dec. 2024**

Master of Science in Artificial Intelligence Engineering ***(GPA: 4.00/4.00)*** *Pittsburgh, PA*

- Courses: Advanced Natural Language Processing, Deep Learning, Machine Learning, AI Systems, Stochastic Processes
- **Awards:** Winner, Citadel Securities x CMU Trading Challenge

University of Rochester **Aug. 2019 – Aug. 2024**

Bachelor of Science in Physics, Bachelor of Arts in Mathematics ***(GPA: 3.97/4.00*** *Rochester, NY*

- **Awards**: Semi-Finalist Rhodes Scholarship India (2022); Harry W. Fulbright Prize, Undergrad Teaching Award (2023)

EXPERIENCE

Efficient Inferencing for Large Language Models | *Research Assistant* **Feb. 2024 – Present**

- Optimizing LLM inferencing via speculative decoding; working on ways to increase the alignment between small and large models
- Working on ways to combine different pruning strategies (weight and activation sparsity) in large language models

Mana Finance Corporation | *Research Intern* **May. 2022 – Aug. 2022**

- Utilized statistical techniques to analyze stock price distributions and quantify investment risk, resulting in predictive machine learning models for assessing expected yields on potential investments.
- Developed a demonstrative project prototype illustrating the feasibility of directly tracking data (ingestion, indexing, and visualization) from the block-chain for Ethereum on UniSwap, eliminating reliance on third-party data streams.

Los Alamos National Laboratory | *Research Intern* **Jun. 2021 – Aug. 2021**

- Modeled complex quantum system (atom interferometer beam splitting) using Python and utilized numerical differential equation solvers in Mathematica and Python.
- Discovered optimal laser pulse parameters through high-dimensional data optimization and parameter estimation using SciPy, CvxPy, and improved the system performance by 5% beyond the current state-of-the-art pulse parameters.

Dark Energy Spectroscopic Instrument | *Research Assistant* **Jan. 2020 – May 2021**

- Designed multi-class CNNs for spectral data with TensorFlow, scikit-learn to find galaxies with supernovae
- Enhanced network performance by applying noise-removal techniques like binning, filtering to preprocess spectral data
- Achieved 95%+ accuracy and high precision for supernovae classification tasks in the DESI data pipeline

PROJECTS

Pruning while Preserving Reasoning Capabilities | *Advanced NLP* **Mar. 2024 – May 2024**

- Showed Bonsai (forward pass only LLM pruning) can be improved in reasoning tasks with a more task aware pruning strategy.

End-to-End NLP System Building | *Advanced NLP* **Feb. 2024 – Mar. 2024**

- Built Retrieval Augmented Generation (RAG) chatbot– scrapped, cleaned, embedded, vectorized database; finetuned reader.

LLaMA-2 from Scratch | *Advanced NLP* **Feb. 2024**

- Built a 42 M model based on LlaMA-2 architecture by implemented ROPE embeddings, training on TinyStories dataset.
- Improved classification performance by continued pretraining on SST-5, CFIMDB datasets; enabled zero-shot learning.

ML Bare Bones **Aug. 2023 – Present**

- Repository of NumPy and Torch implementations of basic ML models like SVM, Linear Regression, AdaBoost, etc.

SKILLS & LEADERSHIP

- **Languages:** Python, Java, SQL, Mathematica, Bash
- **Libraries / Frameworks:** PyTorch, TensorFlow, Hugging Face, Keras, NumPy, Pandas, Scikit-Learn, Spark, Apache Kafka
- **Peer Mentor President, ECE Department (Carnegie Mellon University)**
- **President, Society of Physics Students (University of Rochester)**

PUBLICATIONS

[1] Sadhukhan R., Tiwari V., Chen Z., Chen B. "Using In-Context Learning for Improving Speculative Decoding" [Submitted EMNLP] (2024)

[2] Uzun C., Pandey S., **Tiwari V**., et al. "Improved Bragg splitting of Bose-Einstein condensates into high-order momenta wave-packets". APS Division of Atomic, Molecular and Optical Physics (2023)

[3] Wasserman A., **Tiwari V.,** et al. "Using ML to Develop a Transient Identification Pipeline for DESI" AAS (2021)

[4] Cusenza A., Dunkelberg A., Huffman K., Ke D., Kleber D., Miller S. J., Mizgerd C., **Tiwari V.** "Bounds on Zeckendorf Games". Fibonacci Quarterly (2020)