Vashisth Tiwari

✉ vashistt@cs.cmu.edu | in www.linkedin.com/in/vashistht/ | 🌐 https://vashistht.github.io/

Education

Carnegie Mellon University *Pittsburgh, PA*
Master of Science in Artificial Intelligence Engineering —ECE (GPA: 4.0/4.0) *December 2024*

- Courses: Advanced NLP, Deep Learning, Machine Learning, AI Systems, Stochastic Processes, Deep Learning Systems*
- Teaching Assistant: Advanced Natural Language Processing (Fall'24) with Prof. Graham Neubig

University of Rochester *Rochester, NY*
Bachelor of Science in Physics, Bachelor of Arts in Mathematics (GPA: 3.97/4.0) *May 2023*

- Awards: Phi Beta Kappa, Semi-Finalist Rhodes Scholarship India, Harry W. Fulbright Prize, Undergrad Teaching Award
- Courses: Modern Statistics & Exploration, Data Structures & Algorithms, Probability, Real Analysis, Honors Linear Algebra

Research Experience

Active Learning For Synthetic Data Generation (Advisor: Prof. Emma Strubell) *Pittsburgh, PA*
Graduate Research Assistant *July 2024—Present*

- Designing and exploring the efficacy of active synthetic data generation for LLMs by incorporating student model feedback

Efficient Machine Learning (Advisor: Prof. Beidi Chen) *Pittsburgh, PA*
Graduate Research Assistant *February 2024—August 2024*

- Demonstrated that speculative decoding significantly improves both throughput and latency in LLM inference, a key insight for performance optimization. Achieved up to 2x speedup for LLaMA-3 inference with batch sizes of 128 and above
- Investigated efficacy of model quantization and sparsity (weight, attention, activation) in speculative decoding draft models

ML for Dark Energy Spectroscopic Instrument (Advisor: Prof. Segev BenZvi) *Rochester, NY*
Research Assistant *January 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 and filtering to preprocess spectral data
- Achieved 95%+ accuracy and high precision for supernovae classification tasks in the DESI data pipeline

Work Experience

Mana Finance Corporation *Hillsborough, CA*
Research Intern *May 2022—August 2022*

- Utilized statistical techniques to analyze stock price distributions and quantify investment risk; developed ML models using Facebook Prophet for assessing expected yields on potential investments
- Prototyped a tool demonstrating direct tracking of Ethereum blockchain data on UniSwap (ingestion, indexing, & visualization)

Los Alamos National Laboratory *Los Alamos, NM*
Research Intern *June 2021—August 2021*

- Modeled complex quantum system using Python; utilized numerical differential equation solvers in Mathematica and Python
- Discovered optimal laser pulse parameters through high-dimensional optimization and parameter estimation using SciPy, CvxPy
- Improved the system performance by 5% beyond the current state-of-the-art pulse parameters

Projects

Pruning while Preserving Reasoning Capabilities [Link] | CMU *March 2024—May 2024*

- Improved upon Bonsai (forward pass only structured LLM pruning, compatible with consumer hardware) in math-reasoning tasks
- Demonstrated that novel task-aware pruning metric better retains reasoning abilities than the standard perplexity baseline

End-to-End NLP System Building [Link] | CMU *February 2024—March 2024*

- Engineered a Retrieval Augmented Generation (RAG) based chatbot on CMU utilizing webpages and semantic scholar data
- Implemented core RAG components: LangChain embedder, Faiss+ColBERT retriever, and reader using open-source LLMs

LLaMA-2 from Scratch | CMU *February 2024*

- Built a 42M LLaMA-2 model and trained on TinyStories dataset: implemented ROPE embeddings, AdamW optimizer, attention
- Continued pre trained on CFIMDB and fine-tuned on SST-5 datasets to enable zero-shot movie review sentiment analysis

MLBareBones [Link] | Personal *August 2023—December 2023*

- Implemented ML algorithms from scratch using NumPy: neural networks, SVMs, linear regression, decision trees, AdaBoost, etc.

Skills

Languages: Python, Java, C++, Mathematica, Bash
Libraries / Frameworks: PyTorch, HuggingFace, TensorFlow, NumPy, Pandas, Scikit-Learn, Spark, Apache Kafka, CUDA
Tools: AWS, Jupyter, Linux, Git/GitHub

Publications

[1] Sadhukhan R, Chen J, **Tiwari V**, et al. "MagicDec-2.0" *Submitted to ICLR* (2024). (*Equal contribution)

[2] Chen J*, **Tiwari V***, Sadhukhan R, et al. "MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation with Speculative Decoding." *Accepted ECCV Efficient Foundation Model Workshop* (2024). (*Equal contribution)

[3] 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).

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