

# VASHISTH TIWARI

✉ [vashistt@andrew.cmu.edu](mailto:vashistt@andrew.cmu.edu) | 🌐 [vashistht.github.io/](https://vashistht.github.io/) | 📄 [Google Scholar](#)

## EDUCATION

### Carnegie Mellon University

Pittsburgh, PA

#### Ph.D. in Language Technologies Institute, School of Computer Science

Aug 2025 –

- Teaching Assistant for Inference Algorithms for Language Modeling w/ Prof. Graham Neubig

#### Master of Science in Artificial Intelligence Engineering—ECE (GPA: 3.97/4.0)

Aug 2023 – Dec 2024

- Courses:** Deep Learning Systems, Advanced NLP, Deep Learning, Machine Learning, Stochastic Processes

### University of Rochester

Rochester, NY

#### Bachelor of Science in Physics, Bachelor of Arts in Mathematics (GPA: 3.97/4.0)

Aug 2019 – May 2023

- Courses:** Probability, Hons. Real Analysis, Hons. Linear Algebra, Multi-Dimensional Calculus, Differential Equations
- Awards:** Rhodes Scholarship Semi-Finalist (India), Harry W. Fulbright Prize, Undergrad Teaching Award

## WORK EXPERIENCE

### Google Research (& Google Deepmind)

Seattle, WA

*Student Researcher (Advisor: Giulia DeSalvo)*

May 2025 – Aug 2025

- Post Training for Gemma models, primarily focusing on reinforcement learning and improving the generation of synthetic data for supervised finetuning.
- Extended SoftSRV, a method to generate better synthetic data, to non gemma models.

### Mana Finance Corporation

Hillsborough, CA

*Quantitative Research Intern (Mentor: Max Novendstern, David Kaufman)*

May 2022 – Aug 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.

### Los Alamos National Laboratory

Los Alamos, NM

*Research Intern (Mentor: Dr. Malcolm Boshier)*

Jun 2021 – Aug 2021

- Modeled complex quantum system using Python and utilized Mathematica numerical differential equation solvers.
- Discovered optimal laser pulse parameters through high-dimensional optimization.
- Improved the system performance by 5% beyond the current state-of-the-art pulse parameters through optimized pulses (APS).

## RESEARCH EXPERIENCE

### Active Learning For Synthetic Data Generation

Pittsburgh, PA

*Research Staff (Advisor: Prof. Emma Strubell, Carnegie Mellon University)*

Feb 2025 – May 2025

- Designing a responsive-feedback driven framework where teacher models iteratively refine synthetic data generation based on student model's performance and learning outcomes (for math and reasoning tasks in post training).

### Efficient Machine Learning

Pittsburgh, PA

*Research Assistant (Advisor: Prof. Beidi Chen, Carnegie Mellon University)*

Feb 2024 – Aug 2024

- Showcased how speculative decoding can mitigate the tradeoff between throughput and latency in LM inference.
- Implemented weight pruning, attention sparsity, and activation sparsity techniques for drafting in self-speculation.
- Evaluated compressed models' effectiveness through acceptance rate and speedup metrics in speculative decoding.
- Achieved up to a 2x speedup over autoregressive baseline for LLaMA-3-8B inference at high batch sizes ( $\geq 128$ ) through self-speculation and sparse key-value optimizations.

### ML for Dark Energy Spectroscopic Instrument

Rochester, NY

*Research Assistant (Advisor: Prof. Segev Benzvi, University of Rochester)*

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 to preprocess spectral data.
- Achieved 95%+ accuracy and high precision for supernovae classification tasks in the DESI data pipeline (AAS).

## PUBLICATIONS & PRESENTATIONS

### MACHINE LEARNING

“Energy Considerations of Large Language Model Inference and Efficiency Optimizations”

2025

Fernandez J\*, Na C\*, **Tiwari V\***, Bisk Y, Luccioni S, Strubell E\* et al. (\*Equal contribution). *Accepted to ACL Main*

“MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation”

2024

Chen J\*, **Tiwari V\***, Sadhukhan R\* et al. (\*Equal contribution). *Accepted to ECCV Efficient Foundation Model Workshop. Subsequent work accepted to ICLR 2025*

## PHYSICS

- “Improved Bragg splitting of Bose-Einstein condensates into high-order momenta wave-packets” 2023  
Uzun C, Pandey S, **Tiwari V**, Krzyzanowska K, Boshier M. *American Physical Society Division of Atomic, Molecular and Optical Physics (DAMOP)*
- “High-fidelity splitting of Bose-Einstein condensates into high-order momentum states” 2022  
Uzun C, Pandey S, **Tiwari V**, Krzyzanowska K, Boshier M. *American Physical Society DAMOP*
- “Using Machine Learning to Develop a Transient Identification Pipeline for DESI” 2021  
Wasserman A, **Tiwari V**, BenZvi S.  
📍 Co-Presented at the 237th Meeting of the American Astronomical Society (AAS)

## MATHEMATICS (ALPHABETICAL AUTHOR LIST)

- “Bounds on Zeckendorf Games” 2022  
Cusenza A., Dunkelberg A., Huffman K., Ke D., McClatchey M., Miller S. J., Mizgerd C., **Tiwari V.**, Ye J., and Zheng X.. *Fibonacci Quarterly*, 60 (2022), no. 1, 57–71  
📍 Co-presented the work at the Young Mathematicians Conference with Carl Ye and Kevin Ke
- “Winning Strategy for Multiplayer and Multialliance Zeckendorf Games” 2021  
Cusenza A., Dunkelberg A., Huffman K., Ke D., Kleber D., Miller S. J., Mizgerd C., **Tiwari V.**, Ye J., and Zheng X.. *Fibonacci Quarterly*, 59 (2021), no. 4, 308–318  
📍 Co-presented the work at the Young Mathematicians Conference and UConn Mathematics Conference

## PROJECTS

- Deep Learning Systems** | CMU Aug 2024 – Dec 2024
- Implementing a PyTorch-like Deep Learning library from scratch with auto differentiation, optimizer, and GPU support
  - Adding auto-diff support for Fourier analysis and advanced linear algebra operators (matrix inverse, eigendecomposition)
  - Building efficient tensor operations with CPU/GPU memory management and optimization techniques
- Consumer Hardware Pruning with Preserved Reasoning** | CMU Mar 2024 – May 2024
- Enhanced the Bonsai method (structured LLM pruning for consumer hardware) to better preserve mathematical reasoning
  - Developed novel task-aware pruning metrics that outperformed standard perplexity-based metrics
  - Demonstrated superior reasoning retention in pruned models through comprehensive evaluations
- End-to-End RAG System** | CMU Feb 2024 – Mar 2024
- Engineered end-to-end Retrieval Augmented Generation (RAG) system for university data using LangChain
  - Implemented hybrid search with FAISS vector database, ColBERT dense retriever, and open-source LLM reader
  - Built data processing pipelines to integrate multiple document sources and knowledge bases
- LLaMA-2 Implementation from Scratch** | CMU Feb 2024
- Built a 42M parameter LLaMA-2-style transformer model from scratch using PyTorch
  - Implemented core components: rotary position embeddings, attention mechanisms, and AdamW optimizer
  - Pre-trained on TinyStories dataset and fine-tuned for sentiment analysis on CFIMDB and SST-5 datasets

## TEACHING EXPERIENCE

- Carnegie Mellon University** Pittsburgh, PA
- Teaching Assistant, *Inference Algorithms for Language Modeling (11-763)* Fall 2025
- Teaching Assistant, *Advanced Natural Language Processing (11-711)* Fall 2024
- University of Rochester** Rochester, NY
- Teaching Assistant, *Multiple Courses (Physics & Computer Science)* 2020 – 2023
- Courses: Quantum Theory, Advanced Electromagnetism, Hons. Physics, Introduction to Python

## AWARDS & HONORS

- Semi-Finalist, Rhodes Scholarship**, Indian Consulate 2022
- Phi Beta Kappa**, National Honor Society 2023
- Harry W. Fulbright Prize**, University of Rochester (*Awarded for excellence in experimental physics*) 2023
- Undergraduate Teaching Award**, Dept. of Physics & Astronomy 2023

## SKILLS & INTERESTS

- Programming:** Python (Expert), C++ (Intermediate), Java (Intermediate), Bash (Proficient), Mathematica
- ML/AI:** PyTorch, HuggingFace, TensorFlow, LangChain, Faiss, CUDA, JAX
- Data Engineering:** NumPy, Pandas, Scikit-Learn, Spark, Kafka, AWS (EC2, S3, SageMaker), Docker

## REFERENCES

- Dr. Giulia DeSalvo, Research Scientist, Google Deepmind
- Prof. Emma Strubell, Language Technologies Institute, Carnegie Mellon University
- Prof. Graham Neubig, Language Technologies Institute, Carnegie Mellon University