Vedang Lad

vedanglad.com

☎ (609)-937-7284 ✉ vedang@mit.edu linkedin.com/in/vedanglad github.com/vdlad

Education

Massachusetts Institute of Technology GPA 5.0/5.0 **May 2024**
Master of Engineering in Electrical Engineering and Computer Science *Cambridge, MA*

Massachusetts Institute of Technology GPA 4.8/5.0 **May 2023**
Bachelor of Science in Electrical Engineering and Computer Science *Cambridge, MA*
Bachelor of Science in Physics

Experience

ML Alignment & Theory Scholars (MATS) **June 2024 – Present**
Research Scholar *Berkeley, CA*

- Studying methods to improve the interpretability of large language models under the mentorship of Jessica Rumbelow.
- Developing a novel, data-agnostic method for feature extraction and evaluation for large language models.

Tegmark AI Safety Group **September 2023 – May 2024**
Research Assistant *Cambridge, MA*

- Researched the science of machine learning, or mechanistic interpretability, under the guidance of Max Tegmark.
- Published two first-author papers submitted to top ML conferences, currently under review.

Cleanlab **May 2022 – July 2023**
Machine Learning Engineer *Remote*

- Developed and published a new ML algorithm for label error detection to improve ML data quality.
- Open-sourced error detection algorithms to the Cleanlab Github codebase (9100+ stars) for use by data scientists.

MIT Brain and Cognitive Sciences **December 2021 – May 2022**
Undergraduate Researcher *Cambridge, MA*

- Investigated under the guidance of Joshua Tenenbaum, Dan Yamins, and Judith Fan to analyze the gap in intuitive physics between humans and popular computer vision models.
- Generated state-of-the-art physics simulations to train Graph Neural Networks for pixel-wise predictions.

MIT Kavli Institute with NASA NICER **May 2021 – January 2022**
Undergraduate Researcher *Cambridge, MA*

- Conducted time-series data analysis under Dheeraj Pasham to study black holes using the NASA telescope NICER.
- Implemented optimization algorithms to fit models to energy spectra, to determine black hole composition.

Publications

The Remarkable Robustness of LLMs: Stages of Inference? arXiv:2406.19384
Lad, V., Gurnee, W., & Tegmark, M. (2024).

Opening the AI black box: program synthesis via mechanistic interpretability. arXiv:2402.05110
Michaud, E. J.*, Liao, I.*, Lad, V.*, Liu, Z.*, Mudide, A., Loughridge, C., Guo, Z. C., Kheirkhah, T. R., Vukelić, M., & Tegmark, M. (2024).

Estimating label quality and errors in semantic segmentation data via any model. arXiv:2307.05080
Lad, V. & Mueller, J. (2023).

Extracurricular

MIT Cross Country, Track & Field **August 2019 – May 2024**
NCAA Division III Athlete: 2x Team National Champion, 1x Team National Runner-Up

Plainsboro Rescue Squad **September 2015 – July 2023**
EMT: NJ certified EMT volunteering over 2500+ hours to local community.

Technical Skills

Languages: Python, Java, Julia, JavaScript, HTML/CSS, C, Assembly, Mathematica, Matlab
Developer Tools: VS Code, Jupyter, Pytorch, Tensorflow, Docker, Github, ROS, React