Andrew Yuan

508-335-0860 | awyuan@umass.edu | linkedin/in/andrewwyuan | github.com/awyuan

Education

University of Massachusetts Amherst — Amherst, MA
M.S./Ph.D. in Computer Science — *Sept. 2023 – Present*

- Advised by Przemyslaw Grabowicz
- Research Topics: ML Mechanistic Interpretability, Explainable/Trustworthy AI

Northeastern University — Boston, MA
B.S. in Computer Science, Minor in Mathematics — *Sept. 2019 – Jan 2023*

- GPA: 3.9/4.0

Worcester Polytechnic Institute — Worcester, MA
High School Dual Enrollment — *Aug. 2018 – May 2019*

- GPA: 4.0/4.0

Experience

Research Assistant — Jan. 2022 – Apr. 2023
Northeastern University — *Boston, MA*

- Investigated mechanistic reasons for memorization in Large Language Models like GPT-2 under Professor David Bau.
- Worked with Professor Alina Oprea and Professor Cheng Tan on poisoning attacks in neural network dropout layers. Implemented several vision models, designed and ran new experiments, and presented results.

Full Stack Engineer Intern — Apr. 2020 – Dec. 2021
AbbVie — *Worcester, MA*

- Developed Clonal Lineage Analysis web application to aid scientists in clustering unseeded clonal lineages.
- Developed a single cell RNA sequencer web application that handles various visualizations with a small team of engineers. Met with scientists monthly to determine project goals and present current progress.

Publications

A. Yuan, A. Oprea, and C. Tan. Dropout attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 26–26, Los Alamitos, CA, USA, may 2024. IEEE Computer Society

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore, December 2023. Association for Computational Linguistics

Teaching

Teaching Assistant — Sept. 2023 – Present
University of Massachusetts Amherst — *Amherst, MA*

- CS589: Machine Learning (Spring 2024)
- CS690F: Responsible Artificial Intelligence (Fall 2023)

Technical Skills

Languages: Python, Java, C, SQL (Postgres), JavaScript/TypeScript, HTML/CSS
Frameworks: React, Node.js, JUnit, GraphQL
Developer Tools: Git, Docker, VS Code, PyCharm, IntelliJ, Eclipse
Libraries: HuggingFace Transformers, pandas, NumPy, Matplotlib, Plotly