Xiaoyan(Elena) Bai

smallyan@uchicago.edu • https://elena-baixy.github.io/ • 303-249-2353

Obejective

Current first year PhD student in Computer Science at University of Chicago, advised by Chenhao Tan. I am interested in machine learning, and interpreting language models to make they more trustworthy and less biased

Education

September 2024 - Present

University of Chicago – Chicago, Illinois
PhD student in Computer Science

August, 2022 – April, 2024

University of Michigan – Ann Arbor, Michigan
GPA: 3.866/4.0
Bachelor of Science in Engineering in Computer Science
Minor in Art and Design
Related Course: Machine Learning, Intro to Natural Language Processing, Computer Vision, Game Design and Development

Publications

A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. *ICML(Oral) 2024*
Andrew Lee, **Xiaoyan Bai**, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, Rada Mihalcea

Learn To be Efficient: Build Structured Sparsity in Large Language Models. *Neurips (Spotlight), 2024*
Haizhong Zheng, **Xiaoyan Bai**, Beidi Chen, Fan Lai, Atul Prakash

Technical skills

Programming languages
PyTorch, Python, Java, C++, C, HTML, React, Verilog

Languages
Chinese (native), English (fluent), Japanese (basic)

Working experience

December, 2021 - January, 2022

Data Analyst Internship (Emogent)

- Preprocess and analyze the training data by Python for the interactive AI Irene, which serves in the museum as a guide.

Teaching experience

October, 2024 - December, 2024

Teaching assistant, Mathematical Foundation of Machine Learning (University of Chicago)

- Design course materials, including homeworks, lecture demos and reflection notes
- Hold office hours and recitation class to help student understand the materials

September, 2023 - December, 2023

Grader, Foundations of Computer Science (University of Michigan)

- Grade students' homework for course objective on an introduction to Computer Science theory, with applications

June, 2023 - August, 2023

Teaching assistant, Serious Games and AI (MIT Beaver Summer Institute)

- Create course materials for combining modern methods in machine learning and game-like modeling to quantitatively analyze socially relevant technology and policy questions
- Assist the student teams get their projects working including forming project ideas and debugging Python codes.
- Make sure the students are neither bored nor stressed in online course

Research experience

November, 2022 – April, 2024

Language and Information Technologies lab
Advisor: Prof. Rada Mihalcea (University of Michigan, Ann Arbor).

Interpreting Linear Representations in Language Models (Sept,2023 - Aug,2024)
Mentor: Andrew Lee (PhD student)

- Delve into the current linear representations encoded within large language models to explore explainability in models
- Implement linear probing in language models to intervene the model behaviors

May, 2023 – April, 2024

Prof. Atul Prakash's lab
Advisor: Prof. Atul Prakash (University of Michigan, Ann Arbor)

Large Language Model Moefication (August, 2023 - May, 2024)
Mentors: Prof. Atul Prakash, Haizhong Zheng(PhD student)

- Experiment in the sparsity of models to improve the inference efficiency in GeLU-based LLMs
- Introduce grouping methods to convert the model into Mixture of Experts to robust efficiency.