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

August, 2020 – June, 2024
Shanghai Jiao Tong University – Shanghai, China
GPA: 3.4/4.0
Bachelor of Science in Electrical and Computer Engineering
Related Course: Intro to Logic Design

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

Software
LaTeX, Git, Matlab, Adobe Premiere, Adobe Photoshop

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

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 - Present)
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

Poetry Generation Based on Images (Nov, 2022 - April, 2023)
Mentor: Andrew Lee (PhD student)

- Pre-process poetry data and fine-tune language model using Python to train it to generate model.
- Analyze the generated results to figure out the possible cause for repetition with Python

May, 2023 – April, 2024

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

Token Reduction in Large Language Models (May,2023 - Present)
Mentors: Prof. Atul Prakash, Haizhong Zheng(PhD student)

- Prune the tokens by changing the transformer's architecture with PyTorch to improve inference time efficiency
- Analyze the activation and the attention of tokens to make token pruning decisions with PyTorch

Large Language Model Moefication (August, 2023 - Present)
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.

February, 2022 – September, 2022

Analyzing Depression's Influence on Imagery and Visual Rumination
Mentors: Prof. Weidong Li, Binglei Zhao (Shanghai Jiao Tong University).

- Utilize computer visualization tool to worked on how depression affects human's imagery and visual rumination
- Collect data by conducting brain-computer interface experiments to analyze the depression

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

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

May, 2022 - August, 2022

Teaching assistant, VG100: Intro to Engineering (Shanghai Jiao Tong University)

- Create recitation class materials to introduce students to the professional technical and communication skills required of engineers and providing them with an overview of engineering at the beginning of their program
- Assist the student teams form their project ideas to make a robot car.
- Guide them to be more professional in technical communications

February, 2022 - April, 2022

Teaching assistant, CHEM2110: Chemistry Lab (Shanghai Jiao Tong University)

- Design class materials for a course on the scientific process
- Assist the students to develop experimenting skills and explore chemistry principles.
- Teach students to analyze and report the experiment results.

February, 2022 - May, 2022, September, 2021 - December, 2021

Teaching Assistant, ENGL1000: Academic Writing I (Shanghai Jiao Tong University)

- Assist the students to develop academic writing skills.
- Guide students on the thesis for their academic research.

Activities

September, 2020 - August, 2022

Minister in SJTU Art Center (Shanghai Jiao Tong University)

- Organize school events like Welcome Party, Singer Competitions and Graduation Party to promote community in the study body.

September, 2021 - August, 2022

Minister in Department Student Union (Shanghai Jiao Tong University)

- Organize sports events for students to encourage exercising to help them learn introductory concepts

December, 2022 - January, 2022

Volunteer teaching in Yunnan, China (Shanghai Jiao Tong University)

- Teach middle school students English and Computer Courses