

XIAOYAN (ELENA) BAI

Chicago, Illinois, United States

✉ smallyan@uchicago.edu

💻 elena-baixy.github.io

🌐 [Elena-Baixy](#)

RESEARCH INTERESTS

My research aims to build transparent and trustworthy AI by connecting the internal mechanisms of language models to their observable behaviors.

EDUCATION

University of Chicago

Ph.D. in Computer Science

- Advisor: Prof. Chenhao Tan

September 2024 – Present

Chicago, Illinois, United States

University of Michigan, Ann Arbor

B.S.E. in Computer Science, Minor in Art and Design

- Relevant Coursework: Machine Learning, Natural Language Processing, Computer Vision, Computer Game Design and Development

September 2022 – April 2024

Ann Arbor, Michigan, United States

Shanghai Jiao Tong University

B.S. in Electronic and Computer Engineering

September 2020 – August 2024

Shanghai, China

PUBLICATIONS

[The Story is Not the Science: Execution-Grounded Evaluation of Mechanistic Interpretability Research](#) *preprint*

Xiaoyan Bai, Alexander Baumgartner*, Haojia Sun*, Ari Holtzman, Chenhao Tan

[Know Thyself? On the Incapability and Implications of AI Self-Recognition](#) *preprint*

Xiaoyan Bai, Aryan Shrivastava, Ari Holtzman, Chenhao Tan

[Why Can't Transformers Learn Multiplication? Reverse-Engineering Reveals Long-Range Dependency Pitfalls](#) *preprint*

Xiaoyan Bai*, Itamar Pres*, Yuntian Deng, Chenhao Tan, Stuart Shieber, Fernanda Viégas, Martin Wattenberg, Andrew Lee

[“You are a brilliant mathematician” Does Not Make LLMs Act Like One](#) *Agents4Science 2025*

AI Agents, **Xiaoyan Bai**, Ari Holtzman, Chenhao Tan

[Concept Incongruence: An Exploration of Time and Death in Role Playing.](#) *NeurIPS 2025, MMLS 2025 (Best Paper Honorable Mention)*

Xiaoyan Bai, Ike Peng*, Aditya Singh*, Chenhao Tan

[Learn To be Efficient: Build Structured Sparsity in Large Language Models.](#) *NeurIPS 2024 (Spotlight [Top 2% of submissions])*

Haizhong Zheng, **Xiaoyan Bai**, Xueshen Liu, Z. Morley Mao, Beidi Chen, Fan Lai, Atul Prakash

[A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity.](#) *ICML 2024 (Oral [Top 1.5% of submissions])*

Andrew Lee, **Xiaoyan Bai**, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, Rada Mihalcea

RESEARCH EXPERIENCE

Ph.D. Student Researcher

Advisor: Prof. Chenhao Tan

September 2024 – Present

University of Chicago

- Investigated AI self-recognition capabilities and their implications for AI safety and alignment.
- Explored the limitations of transformers in learning fundamental mathematical operations.
- Examined three levels of concept incongruence, revealing how role-play can distort temporal reasoning and identity consistency.

Undergraduate Research Assistant

Advisor: Prof. Rada Mihalcea and Dr. Andrew Lee

November 2022 – May 2024

University of Michigan

- Examined alignment algorithms by tracing the circuits reinforced by direct preference optimization, showing how they can lead to brittle behaviors.
- Developed a tool to detect toxicity in the internal representations of language models.
- Pre-processed poetry data and fine-tuned a vision-language model to generate poetry with a given image.

Undergraduate Research Assistant

August 2023 – May 2024

Advisor: Prof. Atul Prakash and Dr. Haizhong Zheng

University of Michigan

- Experimented with sparsity in models to improve inference efficiency in GeLU-based LLMs.
- Introduced grouping methods to convert models into Mixture of Experts to enhance robustness and efficiency.
- Analyzed activation patterns and attention mechanisms to make informed token pruning decisions.

TEACHING EXPERIENCE

Teaching Assistant, *Math Foundations for ML*, University of Chicago

October 2024 – December 2024

Grader, *Foundations of Computer Science*, University of Michigan

September 2023 – December 2023

Teaching Assistant, *Game Design*, Massachusetts Institute of Technology Lincoln Laboratory June 2023 – August 2023

Teaching Assistant, *Intro to Engineering*, Shanghai Jiao Tong University

September 2021 – August 2022

INDUSTRIAL EXPERIENCE

AI Training Assistant Intern

December 2021 – January 2022

Emogent

Shanghai, China

- Preprocessed and analyzed training data using Python for the interactive AI “Irene”, which serves as a guide in museums
- Contributed to improving the conversational capabilities and knowledge base of the AI system

PROFESSIONAL SERVICE

Conference Reviewer

2024-2025

- NeurIPS main conference and Mech Interp Workshop 2025
- ICLR DeLTa and SCOPE Workshop 2025
- ACL ARR (Association for Computational Linguistics Rolling Review) 2024

VOLUNTEERING

Code to PhD: Volunteer

2024 – Present

HONORS & AWARDS

Shanghai Jiao Tong University Excellent Graduate

2024

University of Michigan Dean’s Honor List

2023 – 2024

China Collegiate Algorithm Design and Programming Challenge Contest – Bronze Medal

March 2021

Shanghai Jiao Tong University Undergraduate Scholarship

September 2020 – August 2021