# Xiaoyan Bai

Chicago, Illinois, United States
smallyan@uchicago.edu elena-baixy.github.io Elena-Baixy

## RESEARCH KEYWORDS

Build transparent, efficient, and more accessible tools by understanding model internals and alignment post-training methods; AI safety and alignment; efficient agentic workflow for specialized domain

## EDUCATION

**University of Chicago** **September 2024 – Present**
*Ph.D. in Computer Science* *Chicago, Illinois, United States*

- Advisor: Prof. Chenhao Tan

**University of Michigan, Ann Arbor** **September 2022 – April 2024**
*B.S.E. in Computer Science, Minor in Art and Design* *Ann Arbor, Michigan, United States*

**Shanghai Jiao Tong University** **September 2020 – August 2024**
*B.S. in Electronic and Computer Engineering* *Shanghai, China*

## PROFESSIONAL EXPERIENCE

**Microsoft**, Research Intern **June 2026 – September 2026**
*Downstream Improvement through Auxiliary Learning (Mentor: Swati Sharma and Emre Kıcıman)*

- Built an **end-to-end agentic pipeline (DIAL)** that searches over training preparations, including source-grounded auxiliary tasks, curricula, and bounded training recipes for unknown downstream work. The agentic harness proposed source-grounded tasks and curricula and generated and validated examples from **200 long compliance documents** under a configured 1,500–8,000-example budget per candidate.
- Post-trained Qwen3-8B LoRA adapters from fixed base weights using **Direct Reasoning Optimization (DRO)**, **a GRPO-style RL method** with CISPO loss and 16 reasoning rollouts per prompt; used frozen in-loop evaluations to guide task proposals.
- Identified best observed checkpoints that **improved downstream mean task reward by 26.3%–79.0%** over the base model across three evaluated settings.

**MATS**, Scholar **2026**
*Deceptive model organisms and lie confession in reasoning models (Advisor: Alan Cooney).*

- Post-trained **Qwen3.6-27B model organisms** through synthetic-document fine-tuning on approximately **40K documents**; designed **custom GRPO reward functions** to train behavioral bias and concealment as separate objectives. Comparing two concealment recipes showed that similar visible denials can mask different reasoning about the trained behavior, motivating chain-of-thought consistency as a third evaluation axis.
- **Post-trained separate LoRA confession adapters with SFT and DPO** on a deception corpus that excluded the target behavior; tested transfer across organisms, with SFT raising reasoning-on lie-detection balanced accuracy **from 0.51 to 0.84** on the no-CoT-pressure concealer under identity prompting.

**Emogent**, AI Training Assistant Intern **December 2021 – January 2022**
*Irene: Conversational AI guide for museum visitors.*

- Preprocessed and analyzed training data in Python, then expanded Irene's knowledge base to improve the museum guide's conversational responses.

## PUBLICATIONS

★ ARA: Agent-Native Research Artifacts *NeurIPS 2026, AI4Research and AI4Science Workshops at ICML 2026* (**Spotlight**)
Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, **Xiaoyan Bai**, et al.

Mechanistic Interpretability of Chemical Language Models: Molecular Geometry and Representations in MoLMFormer and SMI-TED *Mechanistic Interpretability Workshop at ICML 2026*
Wanqing Hao, Zetong Xu, Zhiyuan Han, **Xiaoyan Bai**, Chenhao Tan

★ The Story is Not the Science: Execution-Grounded Evaluation of Mechanistic Interpretability Research *AI4Science Workshop at ICML 2026* (**Best Poster Award**)
**Xiaoyan Bai**, Alexander Baumgartner*, Haojia Sun*, Ari Holtzman, Chenhao Tan

Know Thyself? On the Incapability and Implications of AI Self-Recognition *preprint, MSLD 2026*
**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

**Undergraduate Research Assistant** **2022 – 2024**
*Advisor: Dr. Andrew Lee and Prof. Rada Mihalcea* *University of Michigan*

- Constructed **24,576 toxic/non-toxic preference pairs** and post-trained GPT-2 Medium with direct preference optimization (DPO), reducing mean toxicity from **0.453 to 0.208** on 1,199 challenge prompts.
- Developed a tool to probe toxic internal representations and compared model activations before and after DPO, showing that post-training bypassed toxic features rather than removing them. [paper]
- Preprocessed poetry data and fine-tuned a vision-language model to generate poems from images.

**Undergraduate Research Assistant** **2023 – 2024**
*Advisor: Dr. Haizhong Zheng and Prof. Atul Prakash* *University of Michigan*

- Co-developed **Learn-To-be-Efficient (LTE)**, a two-stage post-training method that jointly trains LLMs and expert routers with an efficiency penalty, then adapts the model to sparse inference.
- Grouped feed-forward neurons into experts and trained adaptive Sigmoid routers, bringing structured activation sparsity to models with GeLU and SwiGLU activations.
- Benchmarked LTE against MoEfication and Deja Vu; achieved **1.83–2.59× FLOPs speedup** on generation tasks and **25% lower LLaMA2-7B inference latency** at roughly 50% sparsity. [paper]

## 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**

## PROFESSIONAL SERVICE

**Organizer, Interpretability for Discovery Workshop at NeurIPS 2026** **2026**

**Conference Reviewer** **2024 – 2026**

- Transactions on Machine Learning Research (TMLR)
- COLM 2026
- ICML 2026 (**Gold Reviewer**)
- ICLR 2026; DeLTa and SCOPE Workshops 2025
- NeurIPS 2025; Mechanistic Interpretability Workshop 2025
- ACL Rolling Review (ARR) 2024

## VOLUNTEERING

Code to PhD: Volunteer **2024 – Present**

## HONORS & AWARDS

**ICML Gold Reviewer** **2026**

**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**