Lihao Sun

slhleosun@uchicago.edu | slhleosun.github.io

EDUCATION

The University of Chicago *Sep 2021 – Mar 2025*
B.S. in Computer Science & B.A. in Cognitive Science (Honors)
GPA: 3.7/4.0; ML coursework GPA: 4.0/4.0
Research Interests: Mechanistic interpretability of LLMs; representation learning and concept circuits; cognitive modeling and reasoning in AI systems

PUBLICATIONS

- **Aligned but Blind: Alignment Increases Implicit Bias by Reducing Awareness of Race**. **Lihao Sun**, Chengzhi Mao, Valentin Hofmann, Xuechunzi Bai.
  ACL 2025 (Main). ArXiv | Project Page

- **The Geometry of Self-Verification in a Task-Specific Reasoning Model**. Andrew Lee, **Lihao Sun**, Chris Wendler, Fernanda Viégas, Martin Wattenberg.
  Under submission to NeurIPS 2025. ArXiv

RESEARCH EXPERIENCE

Bai's Group, University of Chicago *Research Assistant* *Mar 2024 – Present*
Advisor: Xuechunzi Bai
- Analyzed alignment-induced implicit bias using mechanistic interpretability methods.
- Found alignment reduces explicit bias but worsens implicit bias via racial "blindness."

Insight & Interaction Lab, Harvard University *Research Assistant* *Feb 2025 – Present*
Advisor: Andrew Lee
- Investigated self-verification in task-specific LLMs via GLU vector and attention head analysis.
- Identified minimal circuits for verification using causal intervention and activation geometry.

Social Cognitive AI Lab, Johns Hopkins University *Research Assistant* *May 2024 – Present*
Advisor: Tianmin Shu
- Developed agentic theory-of-mind models for LLM social inference in deduction games.
- Applied Bayesian reasoning to enhance belief modeling in human-agent interactions.

Human+AI (CHAI) Lab, University of Chicago *Research Assistant* *Jan 2024 – Present*
Advisor: Chenhao Tan
- Studied weak-to-strong generalization via affine transformation across tasks and models.
- Found consistent within-model alignment but limited cross-model transferability.

ACADEMIC SERVICE

Reviewer, COLM 2025 (XLLM-Reason-Plan)
Student Volunteer, ACL 2025 Main Conference

AWARDS & HONORS

American Mathematics Competition 12 Distinguished Honor Roll
- World Ranking #91/20,800 (Top 0.5%) *2020*, International
Quad Undergraduate Research Scholar Award (two-time) $10,500 *2024*, UChicago
Data Science Institute Summer Lab Grant $6,000 *2023*, UChicago
Advanced Scholar Award $5,000 *2025*, UChicago

Jeff Metcalf Grant (two-time) $3,000 *2023*, UChicago
Cognitive Science Undergraduate Research Award $500 *2024*, UChicago

TEACHING EXPERIENCE

Teaching Assistant, UChicago MATH 15200 Calculus II *(40+ students)* *2022*
Tutor & Co-Founder, UChicago Data Science Society *(80+ students)* *2022 – 2024*

PRESENTATIONS

Aligned but Blind: Alignment Increases Implicit Bias by Reducing Awareness of Race. Poster Presentation, ACL 2025 (Main Conference), Vienna, Austria

Eliciting and Evaluating Helpful Responses with Weak Models. Poster Presentation, UChicago Undergraduate Research Symposium 2024, Chicago, IL

STARTUP EXPERIENCE

Annotaverse *Founder* *2024 – Present*
Incubator: Beta University Cohort 6
- Building a gamified crowdsourcing platform for high-quality AI data annotation; onboarded pilot users.

SerenAize *Co-Founder* *2022 – Present*
Incubator: UChicago International Leadership Council
- Developed ASMR generation models based on user preferences; 2nd Place, ILC 100 Days Challenge.