Aditya Singh

singh.adityak1@gmail.com | github.com/adsingh-64 | lesswrong.com/users/aditya-singh-1

EDUCATION

The University of Chicago Expected June 2026

BS in Computational and Applied Mathematics | GPA: 3.92/4.00

Coursework: Natural Language Processing (A), Mathematical Methods of Machine Learning (A), Introduction to Causality with ML (A)

PUBLICATIONS & TECHNICAL REPORTS

[1] **Aditya Singh***, Gerson Kroiz*, Senthooran Rajamanoharan, Neel Nanda. "Model Forensics: Determining Whether Concerning Behavior Reflects Misalignment" *ICML 2026 Mechanistic Interpretability Workshop, June 2026.* [arXiv]

[2] **Aditya Singh***, Gerson Kroiz*, Senthooran Rajamanoharan, Neel Nanda. "The Case for Model Forensics" *Alignment Forum, June 2026.* [link]

[3] Gerson Kroiz*, **Aditya Singh***, Senthooran Rajamanoharan, Neel Nanda. "How to Design Environments for Understanding Model Motives." *Alignment Forum, Mar 2026.* [link]

[4] **Aditya Singh***, Gerson Kroiz*, Senthooran Rajamanoharan, Neel Nanda. "Why Did My Model Do That? Model Forensics for Diagnosing LLM Misbehavior." *Alignment Forum, Feb 2026.* [link]

[5] Gerson Kroiz*, **Aditya Singh***, Senthooran Rajamanoharan, Neel Nanda. "Principled Interpretability of Reward Hacking in Closed Frontier Models." *Alignment Forum, Jan 2026.* [link]

[6] **Aditya Singh***, Zihang Wen*, Srujananjali Medicherla*, Adam Karvonen, Can Rager. "Automatically Finding Rule-Based Neurons in OthelloGPT." *NeurIPS 2025 Mechanistic Interpretability Workshop.* [arXiv]

[7] **Aditya Singh**. "Lessons from Studying Model Organisms of Hidden Behavior." *XLab Summer Research Fellowship Report, Aug 2025.* [pdf]

[8] Xiaoyan Bai*, Ike Peng, **Aditya Singh**, Chenhao Tan. "Concept Incongruence: An Exploration of Time and Death in Role Playing." *NeurIPS 2025.* [arXiv]

RESEARCH EXPERIENCE

Research Fellow, Anthropic Aug 2026

- Anthropic Fellow in August 2026 cohort.

Research Fellow, MATS 9.0/9.1 (Neel Nanda) Sep 2025 - Present

- Researched and wrote the field's first paper on Model Forensics, the science of follow-up investigations into concerning behavior. Supervised by Senthooran Rajamanoharan and Neel Nanda.
- Described by Rohin Shah (Head of GDM AGI Safety) as "a great paper he hopes many people read."

Research Fellow, Algoverse AI Safety Fellowship Aug 2025 – Sep 2025

- Made automatic interpretation methods for describing OthelloGPT neurons with decision trees. Supervised by Adam Karvonen and Can Rager.

Research Fellow, Existential Risk Lab (XLab), UChicago Jun 2025 – Aug 2025

- Trained model organisms for interpretability-based auditing of hidden behaviors. Supervised by Adam Karvonen.

SOFTWARE

agent-interp-envs — Framework for studying AI agent behavior in controlled environments for misalignment interpretability research. Multi-provider support (Anthropic, OpenAI, OpenRouter), Docker-containerized agent isolation, checkpointing with resampling support for causal analysis. Includes reward hacking, sandbagging, deception, and whistleblowing environments. Co-developed with Gerson Kroiz.