Zephaniah Roe

Chicago, Illinois | zroe@uchicago.edu | github.com/zroe1

EDUCATION

The University of Chicago, Chicago, IL
Bachelor of Science in Computer Science. Expected, June 2026
GPA: 3.84 / 4.0

EXPERIENCE

University of Chicago Existential Risk Lab, Remote

AI Security Team Lead: (January to September 2025)
- Led the effort to create the **XLab AI Security Guide**, a resource for learning about jailbreaks and adversarial robustness
- Coded the website for the guide and led the effort to design and code the xlab-security python package
- Wrote prose and designed supplemental coding exercises for many sections of the website

Supervised Program for Alignment Research (SPAR), Remote

Research Fellow: (January to September 2025)
- Performed activation patching and trained linear probes to study how diffusion policy robots make decisions
- Researched small autoencoders to understand how neural networks store information and wrote up **results**

JusticeText, Remote

Software Engineering Intern: (June to August 2024)
- Implemented React UI components and performed frontend optimizations to reduce API calls and ensure reliability
- AWS DynamoDB transaction logic and Node.js to implement backend APIs for full-stack applications

University of Chicago Harris School of Public Policy, Chicago, IL
Research Assistant: (January to June 2024)
- Automated a Python process to scrape Afghan government websites to download PDFs for name extraction
- Employed Anthropic and OpenAI APIs to extract names from Afghan government documents to track trends in employment status of individuals over time
- Performed statistical analysis on metadata present in images, PDFs, and Excel files from Afghan ministries

PROGRAMMING

- **Languages**: Python, C, Rust, Javascript, Typescript
- **Projects**:
  - [Toy Models of Superposition Paper Replication](#): a replication of an Anthropic research paper
  - [Neural network from scratch](#): neural network training implemented with only the standard library in C
  - [Blog post on gpt-oss jailbreaks](#): investigation of chain-of-thought jailbreaks on OpenAI open source models
- **Technologies**: OpenAI and Anthropic APIs, git, SVN, Vim, VS Code, gdb, DynamoDB, Linux, LaTeX, etc.

ACTIVITIES

UChicago Existential Risk Laboratory
Club Organizer (August 2024 to present)
- Helped lead the UChicago XLab AI Safety Club, growing the group from 0 to over 60 involved students in one year
- Coded our [website](#), helped create our [intro-curriculum](#), facilitated fellowship meetings, and helped organize events

UChicago Institute of Politics TechTeam / Genesys Works
Project Co-Leader (January 2024 to October 2024)
- Led a team of over 10 undergraduate and masters students to build software for the non-profit Genesys Works
- Oversaw the codebase that will be used in three major US cities to pair hundreds of low-income students with internships

UChicago Alt Protein Project
Research Leader (September 2024 to present)
- Led the computational/ML project for UChicago's alternative protein research accelerator
- Advised a group of 5 undergraduate students and pioneered a [novel approach to protein classification](#)

Other UChicago Activities: UChicago Effective Altruism, UChicago Boxing Club, UChicago Rationality
Other Interests: Classical Latin prose and poetry, prediction markets, boxing, balloon twisting, rap music