# Second Look Research SPAR Stream

University of Chicago XLab

zroe@uchicago.edu | yixiong@gatech.edu | harshul@gatech.edu

## What We Do

Second Look is a research group at the University of Chicago Existential Risk Laboratory that works exclusively on replicating and verifying load-bearing AI safety research. AI safety is a small field in which a comparatively small number of papers shape which problems get prioritized, where talent flows, and what gets communicated to policymakers. Because much of the AI safety community doesn't engage with traditional peer review (and because reviewers rarely run experiments themselves) there is no systematic process for checking whether headline empirical claims hold up. If foundational papers turn out to be misleading, the field loses valuable time building strategy on findings that don't generalize, or imperfect solutions could give labs or the research community a false sense of safety.

Our replications are published openly, written to be brief and readable, and paired with open-source code and interactive visualizations wherever possible, so readers can reach their own conclusions rather than being told what to think.

## Projects We're Excited About

- **Replicating "Confession" work:** One replication which we think would be especially impactful but we don't currently have plans to replicate is OpenAI's work on AI confessions.
- **Replicating and comparing control protocols:** We are interested in work to verify that control protocols behave in similar ways to previous work (either in a standard replication or in one where some aspect of the environment is changed to test the generality of the protocol).
- **Replicating "Assistant Axis":** There is some open source tooling released with the project, but replicating the results on MoE models, more capable models, or with thinking enabled could provide insight into how far the results generalize.
- **Replicating System Card Evaluation results:** Replicating the evaluation figures released in system cards would be extremely useful for verifying the claims of frontier labs.
- **Replicating closed-source frontier lab research:** Frontier labs often produce critical research with broad implications, but don't release models or code for the public to play around with (e.g. Teaching Claude Why). Producing open-source replications of these key results would enable the AI safety community to continue building on this work, and would verify that these highly impactful methods work as intended.

## Replication Criteria

Although by default we will propose a replication for each mentee to work on, we are open to mentee replication proposals as well. The list below outlines factors that could influence if we think a paper is a strong candidate for replication. A paper would not need to score well on all of the below (scoring well on one category could make up for doing badly on some other categories)

- **Narrative influence:** How much does the paper shape how we talk about AI safety to each other and the world? Does it have any important implications for policy or frontier lab practices? Does it reveal technical insights that matter for our threat models?
- **Research dependence:** Does other critical research depend on this paper?
- **Vulnerability:** Does the paper seem unlikely to replicate or generalize? Do authors clearly describe their method, does the papers come with data and code?
- **Difficulty:** How much time and money would the replication take? Is it tractable given our resource constraints and access to relevant models?
- **Recency:** Having replications completed within a few weeks of the paper's release can help the research community iterate quickly. However, ensuring that we do not lower our standards for speed is equally important.
- **Closed Source:** We slightly prioritize findings that are closed source.

Every replication ships with a public repository, an organized commit history, and, where the original work was closed source, whatever code and notebooks are needed for someone else to check our answer in turn.