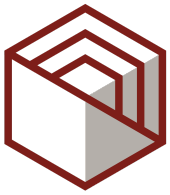

# Scholar Symposium Talk

SERI ML Alignment Theory Scholars Program – Summer 2023

Abstract and visual media are due here on Wed 30 Aug, 6 pm PT
Talks will be held on Fri 1 Sep, 10 am PT

## Task description

The second and final milestone of the in-person research phase of MATS, the Scholar Symposium Talk, will be held on **Friday, 1 September, starting at 10 am PT**. We have allocated 5 min for your talk and 5 min for subsequent audience Q&A. There will be a projector, speaker system, and whiteboard available. Every scholar is required to give a talk by default, but you are welcome to present as a research group. Logistics and team details along with abstracts and supporting visual media are all due on **Wednesday, 30 August by 6 pm PT** (though you can update slides later if necessary). Please message Ryan ASAP for alternative arrangements, including if you believe your research is infohazardous.

The talk is intended to provide the MATS team, grant manager, and broader AI alignment ecosystem with data on MATS research projects, as well as develop scholars' ability to communicate their research. We recommend that scholars exercise reasoning transparency and make use of visual aids in their talk, and aim for clear, concise communication rather than being exhaustive or unnecessarily technical.

The talk should contain:

- A description of a research project you conducted during the MATS program, what you found, and how this might be significant;
- The rationale for why you performed this research that connects it to the broader context of AI alignment/safety;
- If relevant, a brief description of any related future research directions you think are worth pursuing.

We estimate that preparing for the talk should take 3-4 hours, though some scholars might choose to hone their talk for significantly longer. For your talk it is acceptable to, for example:

- Address unknowns and pose questions;
- Use figures and media;

- Use a whiteboard;
- Solicit feedback from peers, your mentor, and Scholar Support;
- Present a single talk as a research team;
- Reference past work, so long as the focus remains on your work at MATS;
- Think about potential Q&A questions and prepare answers;
- Focus on a single project even if you pursued multiple projects within the program;
- Present incomplete results.

If you or your mentor believe that your talk would contain infohazards, please let Ryan know ASAP, and we will discuss alternative arrangements. If you have any other questions, feel free to post them here, send Ryan a message, or ask anonymously.

# Task purpose

This is the second and final milestone of the SERI MATS Program. The Scholar Symposium serves as a way to synthesize and present all the work you have done over the course of the program. If you require help in planning or refining your presentation, we recommend that you consult your mentor, your peers, and the Scholar Support Team.

Along with your recently submitted Scholar Research Plans, the symposium talks serve as an internal impact metric for the SERI MATS Program. Unlike your SRPs, the talks will not impact MATS' evaluations for acceptance into the extension program; however, there is the possibility that funding applications may be influenced, either positively or negatively, by your presentation, as employees of the Long-Term Future Fund, Open Philanthropy, and other funding organizations may be present during your presentations. Additionally, MATS mentors and team members will also be present and your presentation may serve as a means for evaluating further support beyond the MATS Extension Program.

# What to expect

Please arrive at the MATS office on Friday, 1 September by 10 am PT for Scholar Symposium Day. There will be two rooms where presentations will occur simultaneously. You will be assigned a time and room after submitting your abstract and visual media (due Wed 30 Aug, 6 pm PT). At your assigned time, you will be allotted a total of 10 minutes to present. We suggest presenting for ~5 minutes to allow for ~5 minutes of Q&A from evaluators and audience members. There will be a hard cutoff at 10 minutes regardless of where you are in the course of your presentation or Q&A, so we recommend timing your talk in preparation.

Audience members will include MATS peers, MATS team members, and others in AI safety space (including those who might fund or hire you). Your audience will assist in evaluating your talk, so please prepare your content and dress accordingly. Audience members, including fellow scholars, will have access to the standardized rubric below to score your presentation. You will

be able to score the presentations of others when not presenting yourself. **All scholars presenting in-person are required to watch and evaluate at least two presentations**.

We expect to run talks for 2 hours, break for lunch for 1 hour, and resume talks for another 2 hours, however, this could change pending delays and other unforeseen circumstances. Please plan to be in the office for at least 5 hours. Once presentations have concluded, you are free to leave.

Presentation feedback and evaluations will be given by Friday, September 8.

# Resources

- Abstract and visual media submission form
- This post discusses how to be a good podcast guest, but many tips overlap with how to give a good presentation.

## Example Presentations

Please note the examples below are from past cohorts, and the evaluation criteria have since changed. See the rubric below for the current evaluation metrics.

- Example presentation #1
- Example presentation #2

# Rubric (100 points total)

The following rubric is designed to provide a general assessment of each Scholar Symposium Talk based on the criteria and guidelines provided.

| Grade | What? (50 points) | Why? (30 points) | Style (20 points) |
|---|---|---|---|
| 5 - Amazing | Thoroughly explains the project, including details and clear understanding of the project itself.<br><br>Presents findings clearly, concisely and ties them back to project goals in a direct manner. Highlights significance of findings and ties them directly to impact of project.<br><br>Clearly outlines any future research directions related to the project.<br><br>Demonstrates mastery of content. | Effectively and clearly explains how the project is situated in the broader context of a plausible threat model/risk factor.<br><br>Thoroughly explains the potential impact of their work with a comprehensive theory of change.<br><br>Directly addresses failure modes, hazards, and other risks with care.<br><br>Demonstrates thorough commitment to reasoning transparency. | Effective use of visual aids that are useful, informative, maximally utilized, and well organized<br><br>Thoughtfully engages and interacts with the audience beyond answering questions and criticisms. Answers are compelling. Leaves a strong, positive impression.<br><br>Delivery is engaging, confident, and informative.<br><br>Presentation is structured and purposely leaves time for audience Q&A.<br><br>Excellent communication skills. |
| 4 - Good | Demonstrates a clear understanding of the project and outlines some details, but does not go into depth.<br><br>Presents findings with clarity with clear connection to project goals. Mentions significance of findings and loosely ties back to project goals.<br><br>Mentions future research directions related to the project. | Explains how the project is connected to a threat model.<br><br>Presents impact of their project and theory of change is mostly detailed.<br><br>There is some mention of risk analysis, hazards, or failure modes.<br><br>Demonstrates some reasoning transparency. | Includes visual aids that are useful or informative. They are decently utilized and organized.<br><br>Directly engages with questions, criticism, etc. from audience and evaluators.<br><br>Delivery is consistent and mildly engaging.<br><br>Presentation is somewhat structured. Minor deviation from time limit. There is an attempt to leave time for audience Q&A. |

| | | | |
|---|---|---|---|
| | Competently addresses content. | | Good communication skills. |
| 3 - Satisfactory | Presents the project with some clarity, but does not include details.<br><br>Presents findings, but does not demonstrate clear connection to project goals. Significance of findings are mentioned briefly.<br><br>Offers vague ideas about future research directions, or future research directions are somewhat weak or not tractable.<br><br>Attempts to engage with the content, but lacks depth. | Makes a connection to a threat model, though rationale is weak or somewhat unclear.<br><br>Explains the impact of their project, but theory of change is incomplete or briefly mentioned.<br><br>Brief or vague mention of risk analysis, hazards, or failure modes.<br><br>Attempts to demonstrate reasoning transparency, but is mildly inconsistent. | Includes visual aids, though only mildly helpful, informative, or disorganized.<br><br>Engages with questions, criticism, etc. from audience or evaluators. Responses directly address questions, though perhaps weak, not thoughtful, or brief.<br><br>Delivery is consistent despite moments of digression, uncertainty, or irrelevance.<br><br>Presentation is within the bounds of allotted time. Deviation from time limit by a few minutes.<br><br>Fair communication skills. |
| 2 - Subpar | Struggles to convey a clear understanding of the research project.<br><br>Presents findings, but they are unclear, lacking depth, or incoherent with project goals.<br><br>Mentions confusing, unimportant, or disjointed future research directions or mentions future research directions in passing.<br><br>Weakly addresses and engages with the content. | Connection to threat model is incomprehensible, weak, or unclear.<br><br>Attempts to explain the impact of their project and theory of change, but rationale is weak or lacking depth.<br><br>Mention of risk analysis, hazards, or failure modes in passing, or explanation is confusing or incomplete.<br><br>Little commitment to or lacking in commitment to reasoning transparency, or reasoning is illogical. | Includes low-effort visual aids or visual aids are unhelpful, uninformative, or disorganized.<br><br>Hardly engages with questions, criticism, etc. from audience or evaluators. Responses are not thoughtful or dismissive.<br><br>Delivery is lacking, monotonous, or disjointed.<br><br>Presentation is somewhat over or under allotted time. Considerable deviation from allotted time.<br><br>Lacks effective communication and polish. |

| | | | |
|---|---|---|---|
| 1 - Poor | Does not include information about research project or work conducted.<br><br>Does not identify findings from research project. Project findings are confusing or irrelevant.<br><br>Does not mention obvious future research directions.<br><br>Hardly addresses or does not engage with the content, or critical misunderstanding of content. | Does not explain why they chose this project or make an attempt to connect the project to a threat model, or rationale is absent, confusing, or illogical.<br><br>Does not explain the impact of their project, or attempts to explain are incomprehensible, weak, or confusing. Rationale does not exist.<br><br>No commitment to reasoning transparency or reasoning is inherently flawed or opaque. | Does not include any helpful visual aids.<br><br>Does not engage with questions, criticism, etc. from audience or evaluators.<br><br>Delivery is disengaging or incomprehensible.<br><br>Presentation is severely over or under allotted time.<br><br>Poor communication skills, lacking in structure, polish, or engagement. |