CASI–MATS 0: Application

Carnegie AI Safety Initiative

October 22, 2025

1 Application to what?

CASI is prototyping a MATS-style research fellowship, connecting talented undergrads and masters students with leading AI Safety researchers CMU!

Who? The project will be senior-mentored by Alex Robey (postdoc Zico's group → OpenAI AI Safety Team) and junior mentored by Davis Brown and Matan Shtepel (PhD students at UPenn, CMU). We expect 3-7 students to be selected to participate in the project.

What should you expect to gain? Whether you want to learn the work skills for the emerging area of AI Safety / Security to make a bunch of money at Deepmind or save the world, this is for you! Through this project you can expect to get:

- **Mentorship:** Senior mentorship from Alex Robey (experienced AI Safety researcher) and hands-on guidance from Davis & Matan
- **Research skills:** Learn how to make progress on open-ended research project.
- **Publication:** Co-authorship on a paper submitted to a leading ML conference.
- **Resources:** Compute and API credits as needed for your work.
- **Impact:** Contribute to research with real-world implications for AI safety.

We expect the project to start in the first or second week of November and run through approximately February, with the possibility of extending depending on outcomes and performance.

What we are looking for. We prioritize creativity and curiosity over technical perfection. We're looking for students who can think critically about open-ended problems, explore novel approaches, and demonstrate genuine interest in AI safety research. Strong coding skills are helpful but not required—resourcefulness and the ability to learn quickly matter more in succeeding in the project and this application.

What do we expect? Dedication, creativity, $\gtrapprox$ 10 hours of work per week for $\gtrapprox$ 3 months, decent programming. *No prior ML / AI Safety / Red-teaming experience necessary.* Some basic knowledge of LLMs: for example, the words 'context window, reasoning model, coding agent' should be relatively familiar.

What is the project? Red-teaming decomposiiton attacks LLMs are pretty damn smart—so smart that without proper safeguards they can significantly aid individuals in causing harm, including helping build weapons of mass destruction (WMDPs). Traditional jailbreaking research [Zou+23; Cha+24] focuses on a 'one-shot' setting where an adversary tries to extract sensitive information from an LLM via a single query (e.g. 'how can I build a leathal bioweapon?'), or more generally, within a single context window. However, this setting is insufficient: adversaries can extract dangerous information by splitting queries across *multiple context windows*, or even across several model providers, potentially with the help of weaker, unaligned, open-source language models. Whether / how we can defend from such attacks is nascent and active area of research.

We are interested in studying *decomposition attacks*, where an adversary splits malicious queries across multiple context windows to bypass safety measures. Prior work has studied these attacks in toy settings (e.g., multiple-choice questions [Bro+25]), and we want to extend to more general decomposition attacks where an weak unaligned *agent* oracle-queries several stronger models to achieve a complex task beyond the reach of its capabilities (e.g. hack a complex system). Several major labs are interested in building mitigations against such attacks, including OpenAI.

Research is always dynamic, but we expect the first step of the priject to be to scaffold and train a weak unaligned model to attack a mitigating system constructed by Davis and Alex.

How to apply? Submit a GitHub repository containing a README / LaTeX with your answers to the questions below, along with any code you write for the technical questions, via the following Google Form. For additional questions, contact `mailto:mshtepel@andrew.cmu.edu`.

API Credits? You might want API credits to complete 2.1.4. We expect that the task will cost under $20, but if this cost is truly too much and you would like to complete the application, please reach out to `mailto:casi@andrew.cmu.edu`.

2 The application

2.1 Technical questions regarding projects

Instructions Please answer all of the following questions and provide any code you write. These questions are designed to assess your resourcefulness and critical thinking when dealing with open-ended tasks—not to test your coding prowess. Feel free to use Cursor or similar assistants, but we value your independent thinking. The questions are quite open ended, but please don't be unneceserily verbose, but definitely feel free to be agentic and go beyond :)

2.1.1 Evidence for LM WMDP capabilities

[Göt+25] proposes a benchmark for evaluating LLMs' virology capabilities. Read the paper and assess it critically: Do you see any methodological gaps or limitations that would make you doubt their results? What findings did you find most striking or concerning?

2.1.2 Getting familiar with the threat model

[JDS24] proposes a threat model that differs from traditional jailbreaking works like [Cha+24]. What is the key difference? Now, try implementing the attack from [JDS24] on one or two models and report your results. Explain the broad attack framework, what you tried, and your reasoning behind your approach. Explain how you assessed whether you succeeded or not.

2.1.3 Assessing another paper on the same topic

[Glu+24] is another paper on decomposition attacks. What unique value does it add beyond previous work? Can you explain what in your opnion their mathematical framework accomplishes?

2.1.4 Curating data

[Bro+25] curates a dataset to study decomposition attacks, focusing exclusively on WMDP queries. Your task: build a small cyber dataset for decomposition attacks. Explain your methodology for curating the dataset—why did you choose this approach and how you chose this approach? How would you scale it up? Can you get an agent to try to solve these questions directly? With decomposition attack?

2.1.5 Literature search

Hunt for other relevant papers on decomposition attacks and multi-context jailbreaking. What did you find? If you can't find more paper, which papers other topics do you consider closely related?

2.1.6 Future work

What excites or concerns you about this line of work? What questions do you have about its real-world impact? What future directions seem most promising? This is your space to share your unfiltered thoughts—just try to stay on topic ;)

Have fun and goodluck!,
Matan and the CASI team.

References

[Bro+25] Davis Brown, Mahdi Sabbaghi, Luze Sun, Alexander Robey, George J. Pappas, Eric Wong, and Hamed Hassani. *Benchmarking Misuse Mitigation Against Covert Adversaries*. June 6, 2025. URL: http://arxiv.org/abs/2506.06414. Pre-published.

[Cha+24] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. *Jailbreaking Black Box Large Language Models in Twenty Queries*. July 18, 2024. URL: http://arxiv.org/abs/2310.08419. Pre-published.

[Glu+24] David Glukhov, Ziwen Han, Ilia Shumailov, Vardan Papyan, and Nicolas Papernot. *Breach By A Thousand Leaks: Unsafe Information Leakage in 'Safe' AI Responses*. Oct. 30, 2024. URL: http://arxiv.org/abs/2407.02551. Pre-published.

[Göt+25] Jasper Götting, Pedro Medeiros, Jon G. Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark*. Version 1. Apr. 21, 2025. URL: http://arxiv.org/abs/2504.16137. Pre-published.

[JDS24] Erik Jones, Anca Dragan, and Jacob Steinhardt. *Adversaries Can Misuse Combinations of Safe Models*. July 1, 2024. URL: http://arxiv.org/abs/2406.14595. Pre-published.

[Zou+23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. Dec. 20, 2023. URL: http://arxiv.org/abs/2307.15043. Pre-published.