Call for Research Ideas: Expanding the Toolkit for Frontier Model Releases

Foundational Research Grants (FRG) is calling for research ideas that would expand and improve the toolkit for frontier model releases. FRG, a grant program based within the Center for Security and Emerging Technology (CSET) at Georgetown University, funds work on technical topics that bear on the potential national security implications of AI over the long term.

At a glance

- **Topic:** Expanding the toolkit for frontier model releases
- **Award size and duration:** Up to $750,000 per project, to be expended over 3-24 months
- **Selection process:** 1-2 page expression of interest due **July 3**; more details below

Research scope

We are interested in supporting research that **expands and improves the toolkit** available to **AI developers and external observers** who are assessing how to release a new **general-purpose model with cutting-edge capabilities** safely and responsibly**.** More on this below.

By **"general-purpose model"** we mean AI systems that can be used or adapted for a wide range of tasks, rather than being developed for a single purpose—such as foundation models, large language models, and similar. By **"cutting-edge capabilities,"** we mean models that push the envelope of what AI systems can do, or are not far behind that threshold. The term "frontier model" is sometimes used to mean something similar to "general-purpose model with cutting-edge capabilities," so we use both terms interchangeably in this document.

Background and motivation

A wide range of actors—including major tech companies, nonprofit organizations, and informal collectives of individuals—are working to develop increasingly advanced general-purpose AI systems (such as, but not limited to, large language models). An unusual property of general-purpose AI systems is that it is far from straightforward to determine what they can and cannot do, which can make it difficult to assess their likely impacts. This is in contrast to many previous generations of AI and machine learning, in which a given system was generally designed for a single intended task (or a small number of related tasks), making it relatively easy to test how well the system performed at that task.

Recent conversations about release decisions for new models have often simplified the question into one of whether the model should be "open" (i.e. with model weights shared freely online) or "closed"

(i.e. with access to model weights heavily restricted). **In reality, the release-related decisions facing model developers are many and multidimensional**, including not just where on the gradient of openness[1] to fall, but also how to assess risks and benefits in the first place, how to monitor model use (and abuse) after release, what information (if any) to share about the model with other actors, and many other questions.

There is little in the way of agreed-upon best practices for how to make these decisions, and different actors have extremely different estimations of the risks and benefits of different model release approaches. This poses a particular challenge for general-purpose models with cutting-edge capabilities, sometimes called "frontier models," which can bring novel risks and unknown unknowns, in addition to many better-understood (but still unsolved) concerns.

To the extent that model developers are coming up with answers to these questions, **much of this work is happening behind closed doors inside major tech companies such as Google, Meta, OpenAI, and Anthropic**. It is commendable that these companies are dedicating attention to these questions, but the limited transparency around much of this work makes it very difficult for policymakers, civil society, and the broader public to form their own views on companies' decisionmaking. Creating a richer set of tools, concepts, and best practices that are widely available would make it much easier for external observers to hold companies accountable for the decisions they are making around model release. It could also allow less well-resourced AI developers facing similar questions to make better decisions.

Topics of interest

The goal of this call for research ideas is not simply to produce commentary on model release approaches, but to create new tools and frameworks that expand the range of possibilities for making model release decisions and/or improve the quality of existing tools and approaches.

Within these bounds, we are open to a wide range of possible research approaches, topics, and formats that relate to this goal. Some categories of work that we could be interested to support include:

- Frameworks & toolkits for **pre-release risk assessment and decisionmaking**. Some existing work lists categories of risks that could be assessed or describes what mitigations have been implemented, but different organizations seem to handle these questions quite differently. Sub-topics of interest and relevant prior work include:
    - Approaches to accounting for uncertainty, including as it applies to different release approaches (e.g. API access vs. finetuning vs. open weights release), given that it is impossible to perfectly assess how a model will be used (Qi et al. 2023, Davidson et al. 2023).

[1] Solaiman 2023. The Gradient of Generative AI Release

  - Overarching frameworks to assess risk and inform decisionmaking, e.g. responsible scaling policies (METR 2023),[2] marginal risk (Kapoor et al. 2024), safety cases (Clymer et al. 2023).
  - Approaches to assessing novel risks that do not have an empirical track record, e.g. as explored in the literature on experimentalist governance (Wansley 2016).[3]

- Prototypes, demonstrations, or commentary **blurring the boundaries between open and closed releases**. Much ink has been spilled about open vs. closed releases, and about the limitations of framing the question in that way. What options exist (or could be developed) that capture benefits typically associated with both sides of the open-closed dichotomy? What are more productive ways to think about the possibility space for model releases? To what extent are different possibilities compatible with the complex incentives of developers, users, and regulators? Relevant prior work includes research and prototyping on remote audits (OpenMined 2023, Waiwitlikhit et al. 2024), structured access (Shevlane 2022, Bucknall and Trager 2023), limiting harmful fine-tuning (Henderson et al. 2023, Deng et al. 2024), and unlearning hazardous knowledge (Liu et al. 2024, Li et al. 2024).

- Frameworks & toolkits on how to think about **post-release monitoring and mitigations**. Public discussions of model release often focus on pre-release questions, and tend to neglect after-the-fact matters of how a model deployer can monitor the use of their model or system, and what they can do if they discover undesirable use. To the extent monitoring and mitigation are considered, it is often under the assumption that it is a solved problem for closed models and an unsolvable one for open models, though this does not seem to be the status quo in practice (Henderson et al. 2024). What does responsible post-deployment monitoring look like? How should developers navigate privacy and business-model tradeoffs (i.e. customers preferring their data is not retained)? What monitoring and mitigation options are available to developers who have released models publicly? How should AI developers think about—and ideally measure—broader systemic effects of release, rather than isolating their analysis to a single model? Relevant prior work includes writing on "deployment corrections" (O'Brien 2023) and post-deployment reporting (UK Department for Science, Innovation & Technology 2023).

Individual projects can request up to $750,000 for relatively resource-intensive projects, but we are also very open to providing much smaller sums for simpler or shorter projects, e.g. writing position papers addressing the kinds of questions covered above. We are open to supporting work in a range of

[2] Responsible scaling policies and similar have so far been adopted by companies that do not plan to open source their models, and so lack any guidance on when it is and is not responsible to open source a model. We are interested in work that could close this gap.

[3] In general, we welcome work applying more established methodologies from other fields to these questions.

disciplines, including technical toolkits or frameworks but also non-technical analyses of relevant questions. The key consideration will be whether the research outputs are connected to the practical realities of model release decisions.

Submission guidelines and process

Award eligibility, size, and duration

- **Eligibility:** Principal investigators should be based at an institution of higher education or a nonprofit research organization[4]
- **Award size:** up to $750,000 (including no more than 10% indirect costs)
- **Duration:** 6–24 months

Process for submissions

Stage 1: Short expression of interest (due July 3, 2024). Initial submissions should briefly (in 1-2 pages) describe a potential research project relevant to the scope described above. This submission does not need to lock in final details of the project, but should explain roughly what you would be interested to work on. Be sure to clearly include:

- Table of key figures, including: key personnel, host institution, rough cost estimate:

Principal Investigator	
Other Key Personnel	
Host Institution	
Rough Cost Estimate	
Anticipated Period of Performance (duration)	

- Goals of research: ≤200 words
- Research approach: ≤200 words
- Relevance of the work to this call for research ideas: ≤300 words
    - This section should draw explicit connections to the “topics of interest” section of this document.

Send expressions of interest to <csetfoundations@georgetown.edu> with subject line: “Response to 2024 call for research ideas”.

[4] International institutions are welcome to apply, though they may face additional administrative burdens if awarded a grant.

Stage 2: Discussion and development. Selected applicants will be invited to discuss and develop their research idea further with the FRG leadership team over the course of 1–3 informal calls.

Stage 3: Proposal (due September 6, 2024). Selected applicants will be invited to submit a short proposal (around 4-6 pages) including:

- Project scope and goals
- Milestones and timeline
- Budget and brief budget narrative

Stage 4: Final selection. Reviewers will evaluate proposals and select projects to be funded based on feasibility and relevance.

Timeline for submissions

1. Expression of interest: Due July 3, 2024
2. Discussion and development: July 3 through July 26, 2024
3. Proposal: Due September 6, 2024
4. Final selection: Selections made by October 4, 2024

Criteria for selection

Submissions will primarily be evaluated in terms of (a) their relevance to the research scope described in this document and (b) technical feasibility and merit.