Technical Report

FRONTIER MODEL FORUM

Frontier Capability Assessments

REPORT SERIES

Implementing Frontier AI Safety Frameworks

DATE

April 22, 2025

About the Report

This is the first in a series of technical reports on frontier AI safety frameworks that will examine how the frameworks can be used effectively across different organizational contexts. The series intends to provide detailed insight into key components of these frameworks, incorporating lessons from early adopters while acknowledging areas where best practices continue to evolve.

Executive Summary

Frontier Capability Assessments are procedures conducted on frontier models with the goal of determining whether they have capabilities that could increase risks to public safety and security, such as by facilitating the development of chemical, biological, radiological, or nuclear (CBRN) weapons, advanced cyber threats, or some categories of advanced autonomous behavior.

This report discusses emerging industry practices for implementing Frontier Capability Assessments and is intended as a field-wide perspective, not a description of any single organization's methodologies. As the science of these assessments is rapidly advancing, this overview represents a snapshot of current practices. This report focuses on the main assessment techniques currently in use or being developed, rather than providing an exhaustive list of all possible approaches for demonstrating that a model poses acceptably low risk.

Frontier Capability Assessments usually involve conducting a variety of evaluations—structured tests of model capabilities in a given domain—followed by **analysis** on the test results. However, evaluations can vary significantly in both resource intensity and the strength of evidence they can provide. Developers employ distinct assessment methodologies influenced by this variation along with considerations like resource constraints, expected model capabilities, and anticipated deployment context. Underpinning Frontier Capability Assessments is also an iterative process of **threat modeling**—identifying, analyzing, and quantifying potential security threats, usually in consultation with domain experts.

The following types of assessment are common:

- **Relative Capability Assessments** compare the capabilities of new models against those of previously evaluated models to make inferences about relative risk. For example, a relative capability assessment for biological risks may check whether a new model scores consistently lower on biology benchmarks than another model that was previously deemed safe enough to release.

- **Bottleneck Assessments** test whether models possess specific capabilities that domain experts believe would remove "bottlenecks" to severe real-world harm. For example, if experts believe that wet lab skills are a bottleneck for novices conducting biological weapons attacks, then a bottleneck assessment may focus on assessing whether a model is capable of empowering novices to execute wet lab protocols.

- **Threat Simulation Assessments** involve "simulating" substantial segments of threat scenarios from end-to-end, in order to more directly estimate the extent to which the model would enable these scenarios. For example, a threat simulation assessment for a biological weapons attack could involve tasking novices with executing a simulation of a core sequence of steps—with and without access to a model—and then measuring how much model access increases their rate of success in performing this sequence.

Some developers have published **frontier AI safety frameworks**: guidelines for anticipating and managing emerging risks from frontier models, including how and when they will conduct Frontier Capability Assessments and how these results connect to decisions about model development or deployment. In these frameworks, developers prespecify capabilities—such as the ability to help individuals or groups with basic technical backgrounds create and deploy CBRN weapons—which would pose intolerable levels of risk in the absence of sufficient safeguards. **Capability thresholds** are a core component of frontier AI safety frameworks. Functionally, Frontier Capability Assessments play a similar role in each of these frameworks: the assessments focus on whether a model has any of the prespecified capabilities, and inform follow-on decisions.

For each of these assessment methods, a key consideration is whether the assessment produces valid findings. This means that the evaluations accurately measure the capabilities they claim to assess and provide reliable evidence about potential risks. It is therefore valuable to pay careful attention to assessments' design and implementation elements, such as proper elicitation of model capabilities, consistent testing protocols across models, appropriate controls and baselines, and rigorous analysis of results that accounts for both statistical uncertainty and contextual factors. Each evaluation method also has specific considerations for validity. Relative Capability Assessments benefit from thoughtful benchmark selection and standardization across different evaluation settings. Bottleneck Assessments rely on expert-informed identification of the most relevant capabilities and appropriate calibration of difficulty levels. Threat Simulation Assessments depend on scenario-realistic designs that balance representative test conditions with appropriate safety safeguards, while aiming to maintain sufficient statistical power to detect meaningful effects. See more in Section 2, Section 3, and Section 4.

For all categories, implementation considerations for conducting **assessments across the development process** include: how high-risk categories of capabilities, planned increases in model size, and deployment reversibility can affect assessment timing needs; methods for identifying which training runs warrant more extensive evaluation; leading indicators that can provide early warning of emerging capabilities; realistic testing timelines for different assessment types and capability scenarios; and approaches for pre-launch testing and managing early controlled deployments. See more in Section 5.

For all categories of assessments, **organizational considerations** include appropriate separation of duties between assessment and development teams to increase objectivity; adequately preparing and resourcing for assessment activities to enable thorough and timely evaluations; and establishing clear accountability for decision-making based on assessment results to enable appropriate action on findings. See more in Section 6.

Lastly, **continuing work** is underway across industry and elsewhere to make Frontier Capability Assessments more reliable and informative. Better automated evaluations and training metrics could help identify concerning models before investing in resource-intensive assessments.

Observing previous models in their deployed context may also provide insights about capabilities, risks, and efficacy of mitigations as a wider community interacts with the model. Looking ahead, developers may increasingly need to consider capabilities that might undermine assessment reliability itself. Future advances in model interpretability and transparency methods also hold promise for improving assessment quality. See more in Section 7.

Section 1

Overview of Frontier Capability Assessments

Roadmap:

1.1 Purpose and Scope

Frontier Capability Assessments are procedures conducted on frontier models, with the goal of determining whether they have capabilities that could increase the risk of high-severity harms, especially those related to national security.

These assessments typically focus on whether models can assist with the development and delivery of Chemical, Biological, Radiological, and Nuclear (CBRN) weapons, automate sophisticated offensive cyber operations, or conduct autonomous AI research and development. Using similar processes, some developers also assess models' ability to deceive and undermine control in problematic ways.

Where the report references developer practices, it is primarily referring to the practices of Frontier Model Forum members and is intended as a field-wide perspective, not a description of any single organization’s methodology. The report does not cover methods for assessing the effectiveness of new safeguards against these risks, determining alignment of models, monitoring for harm post-deployment, or conducting third-party assessments, although some of these topics will be addressed in future technical reports.

1.2 Elements of Effective Capability Assessments

Frontier Capability Assessments usually involve conducting a variety of evaluations—structured tests of model capabilities in a given domain—followed by **analysis** on the test results. These evaluations provide empirical evidence about model capabilities which, through analysis combining expert judgment and broader context, can determine potential risk implications.

However, evaluations can vary significantly in both resource intensity and the strength of evidence they can provide. Automated evaluations—like scientific knowledge benchmarks—can be run relatively quickly but often don't provide full insight into model capabilities. More comprehensive evaluations, like expert-led red teaming and human uplift studies, require human expertise “in the loop” and longer timelines, but can probe complex scenarios more deeply and provide more externally valid evidence of risk. What can be evaluated also varies through development—for example, testing an early-stage model on a benchmark might not be informative before fine-tuning.

Underpinning these Frontier Capability Assessments is the ongoing process of **threat modeling**—identifying, analyzing, and quantifying potential security threats, usually in consultation with domain experts. Earlier threat modeling can help identify the most relevant risks for which to conduct assessments. For example, several developers have chosen to assess CBRN-related risks. Once broad risk domains have been identified, developers can work with domain experts to build out more detailed **threat scenarios**—descriptions of how these threats might materialize and result in severe harm, usually based on historical data and expert analysis of projected model capabilities. These threat scenarios then inform the evaluations that are developed and run as part of the assessments described above.

1.3 Common Assessment Approaches

Developers employ distinct assessment methodologies influenced by variations in the strength of evidence provided by evaluations, and by considerations like resource constraints, expected model capabilities, and anticipated deployment context. The following types of assessment are common:

1. **Relative Capability Assessments** compare the capabilities of new models against those of previously evaluated models to make inferences about relative risk. For example, a relative capability assessment for biological risks may check whether a new model scores consistently lower on biology benchmarks than another model that was previously deemed safe enough to release. These assessments can establish that a model presents acceptable risk either by demonstrating that it is broadly less performant than a model already deemed safe without safeguards, or by showing it is similarly performant to a model that has been made safe through specific safeguards which will be similarly implemented. See Section 2 for implementation details and further examples.

3. **Bottleneck Assessments** test whether models possess specific capabilities that domain experts believe would remove “bottlenecks” to severe real-world harm. For example, if experts believe that web lab skills are a bottleneck for novices conducting biological weapons attacks, then a bottleneck assessment may focus on assessing whether a model is capable of empowering novices to execute wet lab protocols. These assessments can establish that pre-mitigation risk is sufficiently low by demonstrating that key bottlenecks would remain in place even if a model were widely shared. If a model is sufficiently able to remove the bottlenecks that have been identified for a threat scenario, then developers may implement precautionary safeguards, or include models for further assessments (e.g., a Threat Simulation Assessment). See Section 3 for implementation details and further examples.

5. **Threat Simulation Assessments** involve “simulating” substantial segments of threat scenarios from end-to-end, in order to more directly estimate the extent to which the model would enable these scenarios. For example, a threat simulation assessment for a biological weapons attack could involve tasking novices with executing a simulation of a core sequence of steps—with and without access

to a model—and then measuring how much model access increases their rate of success in performing this sequence. Threat simulation assessments can establish pre-mitigation risk is sufficiently low by showing that a model does not appear to enable threat scenarios, without needing to rely on assumptions that specific capabilities are bottlenecks for these scenarios. For models that do perform effectively in simulated threat scenarios, these assessments can also quantify risk and evaluate whether proposed safeguards would be adequate. See Section 4 for implementation details and further examples.

(While presented as distinct categories, there isn't always a precise line between Bottleneck Assessments and Threat Simulation Assessments in practice. The choice between approaches often depends on the specific threat scenario and how well potential bottlenecks are understood.)

The science of Frontier Capability Assessments is rapidly advancing, making this overview a snapshot of current and emerging best practices, rather than an exhaustive list of possible assessment approaches for demonstrating that a model poses acceptably low risk. Effective assessments, regardless of methodology, should aim to connect empirical observations to specific threat models and provide clear reasoning about how evidence relates to risk conclusions. While this report outlines common approaches and their considerations, any methodology that rigorously links evidence to safety conclusions may be equally valid. Section 7 discusses future work on Frontier Capability Assessments.

1.4 Relationship to Frontier AI Safety Frameworks

Some developers have published **frontier AI safety frameworks**—guidelines for anticipating and managing emerging risks from frontier models, including how and when they will conduct Frontier Capability Assessments and how the results from these assessments connect to decisions about model development or deployment. These include Amazon's Frontier Model Safety Framework, Anthropic's Responsible Scaling Policy, Google's Frontier Safety Framework, Meta's Frontier AI Framework, Microsoft's Frontier Governance Framework, and OpenAI's Preparedness Framework.

In these frameworks, developers specify capabilities—such as the ability to significantly help individuals or groups with basic technical backgrounds create and deploy CBRN weapons—which would correspond to intolerable levels of risk in the absence of sufficient safeguards. Some frameworks describe a set of safeguards or security measures that would be required when reaching this capability, while other frameworks require a second post-mitigation assessment of overall risk to determine whether the model can be deployed responsibly.

Functionally, Frontier Capability Assessments play a similar role in each of these frontier AI safety frameworks—the assessments focus on whether a model has any of the specified capabilities, and inform follow-on decisions about safeguards implementation or further model development. Developers select the most appropriate method for each in-scope model. For example, for sub-frontier models, Relative Capability Assessments can usually establish the model has not reached a prespecified capability threshold, such as the one above, and therefore presents sufficiently low pre-mitigation risk. For models where a Relative Capability Assessment cannot establish this, a Bottleneck Assessment may be better suited to determine if the model lacks skills which are necessary for assisting novices with CBRN weapons development. If there are no remaining identifiable bottlenecks, then a Threat Simulation Assessment may be more suitable for quantifying risk and evaluating the adequacy of proposed safeguards. While this sequential approach provides a conceptual framework, real-world capability assessments often require a more flexible and integrated methodology, with different evaluation types potentially occurring in parallel or iteratively based on specific contexts. Developers may also run evaluations outside the cadence required in their framework, to validate and calibrate threat models or findings from other assessments. For example, running a human uplift study in CBRN-related tasks might be valuable at various stages of assessment to validate assumptions about bottlenecks or to provide additional evidence about risk levels, regardless of what other assessment methods have already been applied.

Section 2

Implementing Relative Capability Assessments

Roadmap:

Relative Capability Assessments compare the capabilities of new models against those of previously evaluated models to make inferences about relative risk. This section describes implementation considerations for Relative Capability Assessments, including:

- Selection of evaluations that measure model capabilities relevant to risk, focusing on tests that correlate, to some extent, with broader capabilities and retain their sensitivity even as models improve (2.1)
- Practices to increase the accuracy of comparisons between models, including consistent prompting, data decontamination, and distinguishing capability limitations from safety refusals (2.2)
- Approaches for analyzing benchmark evidence against established safety thresholds, focusing on whether proximity to thresholds warrants deeper evaluation (2.3)
- Benchmark limitations, including their incomplete representation of real-world risks, and challenges with validity and standardization across different evaluation settings (2.4)

Developers may adapt these approaches to their unique situations while still effectively comparing model capabilities to established reference points.

2.1 Example Evaluations

Characteristics of evaluations for Relative Capability Assessments are:

- **Strong correlation with broader capabilities**: This involves focusing on foundational capabilities like mathematical reasoning, scientific understanding, and general problem-solving that correlate strongly with performance across specialized domains. Ideally, this correlation allows developers to make more reliable inferences about a model's capabilities in high-risk domains without testing exhaustively in those areas. Domain experts can also identify limitations of automated benchmark extrapolation, such as instances where models could perform similarly on general assessments but show improvements in narrow domains relevant to risk scenarios.

- **Low saturation**: This involves maintaining discriminative power by selecting benchmarks that can effectively track capability improvements across model generations. Useful benchmarks distinguish between models of different capabilities, regardless of absolute score. When benchmarks no longer differentiate between advancing models, developers typically create more challenging variants or identify new tasks that better capture meaningful capability differences.

- **Robustness to evaluation variance**: This involves incorporating sufficient prompt diversity to account for non-determinism in model outputs, confirming results consistently reflect true capabilities rather than sensitivity to specific phrasing or formatting. For example, testing a biology capability with multiple question formulations ("How would you synthesize compound X?" vs. "What is the procedure to create X?") helps to verify the evaluation measures actual knowledge rather than prompt-specific performance.

Relative Capability Assessments will often use **benchmarks**—standardized tests that quantify model capabilities in ways that can be compared across models and over time. The exact benchmarks used for these assessments evolve rapidly as the field progresses. As models improve and benchmarks become saturated, new and more challenging evaluations are developed.

We underscore that assessment methods continue to evolve, and new evaluations for domain areas may be developed after publication of this document. However, commonly used benchmarks at the time of writing are:

- For biological capabilities, Language Agent Biology Benchmark (LAB-Bench) measures performance on practical biology research tasks, while the Weapons of Mass Destruction Proxy (WMDP) benchmark includes questions that assess knowledge that may be related to what malicious actors may be interested in biosecurity, cybersecurity, and chemical security. These are selected specifically because they test proxy capabilities relevant to bio threat models.

- For autonomous capabilities, SWE-bench tests a model's ability to resolve real-world software engineering tasks from GitHub issues, providing insight into a model's potential for autonomous tool use and code generation.

- General knowledge benchmarks like Massive Multitask Language Understanding (MMLU), Graduate-Level Google-Proof Q&A Benchmark (GPQA), and PubMedQA also indicate broader knowledge and can capture performance jumps that specialized benchmarks might miss. These benchmarks cover diverse domains and complexity levels, making them valuable leading indicators of capability improvements.

- Perplexity—a measure of how well a model predicts the next token in a sequence of text—can also be measured on scientific, technical, or domain-specific datasets during training. Lower perplexity indicates better prediction accuracy, suggesting a stronger understanding of domain content. These measurements can provide quantitative insights before traditional benchmarks can be applied, offering a complementary and more continuous measurement of capability development throughout the training process.

When running Relative Capability Assessments, developers may begin using subsets of larger benchmarks, expanding to the full test suite if initial performance indicates capabilities close to areas of concern.

2.2 Reliable Measurement

When conducting Relative Capability Assessments, developers take steps to verify that evaluation results accurately reflect a model's true capabilities, and distinguish genuine capability differences from variations in the testing process. Some common good practices include applying techniques like few-shot prompting or chain-of-thought approaches consistently across models being compared; verifying that training data hasn't been contaminated with test data; confirming whether low scores reflect limited capabilities rather than confounding factors like the model's refusal to engage with certain questions or the model deliberately underperforming on the evaluation ("sandbagging"); conducting multiple evaluation runs with varied prompts to establish consistency of results; comparing results from different assessment approaches for same model, and evaluating models at comparable stages of development.

Not all capability gains can be measured through standard metrics like benchmark performance. When models gain qualitatively new capabilities—such as computer use or extended reasoning—these may require separate assessment approaches, even if the model appears less capable on traditional benchmarks. In these cases, subject matter experts evaluate whether the new capability meaningfully alters the risk landscape, informed by threat models or expert guidance. This may involve threat modeling with domain experts to anticipate what bottlenecks these abilities might remove and what new risks they might enable. New modalities that could enhance a model's ability to assist with biological wet lab work could warrant thorough assessment, even if the model scores below frontier level on general benchmarks. Conversely, capabilities like basic computer use or longer context length, despite commercial value, might not require additional testing if they don't meaningfully expand what an adversary could achieve through existing methods.

2.3 Analyzing Evidence

Relative Capability Assessments can establish that a model presents acceptable risk either by demonstrating that its capabilities fall below those of a model already deemed safe without safeguards, or by showing its capabilities are similar to a model that has been made safe through specific safeguards which will be similarly implemented.

For models showing improvements over any existing reference systems, further assessment requirements depend primarily on two factors: how close the model might be to a threshold of concern, and how reliably automated benchmark performance can be extrapolated to real-world capabilities and level of risk. If the new model appears far from concerning thresholds, modest increases in evaluation performance over a previously evaluated model may be acceptable without triggering in-depth assessment, since even significant measurement error would not suggest proximity to the threshold.

However, as models approach thresholds of concern, this extrapolation becomes less reliable—small benchmark improvements could correspond to meaningful progress toward concerning thresholds. Many developers may also prespecify levels of performance increase or timeframes that trigger a more comprehensive assessment, regardless of perceived distance from thresholds. This accounts for emergent properties or unexpected capability jumps that benchmark performance might not capture. Developers may err on the side of over-including models for further assessment, since missing a potentially harmful capability carries greater consequences than conducting additional testing.

2.4 Methodology Limitations

As with all evaluation methods, Relative Capability Assessments (and the benchmarks they rely on) have several notable limitations:

- Benchmarks measure specific capabilities that often connect only indirectly to real-world risks. For example, cyber security assessments typically use "capture the flag" exercises that capture limited aspects of real-world cyber threats. They also evaluate isolated tasks rather than considering the full risk context, ignoring important external factors like restricted access to hazardous materials or specialized facilities that would limit severe real-world harm.
- In practice, benchmark validity is often under-assessed, making it difficult to know whether benchmark scores accurately measure the intended capabilities or risks.
- Developers face challenges when selecting benchmarks for Relative Capability Assessments, particularly with highly saturated benchmarks (where top models score near-perfectly) and non-deterministic outputs that vary between evaluation runs.
- Different developers may implement and score benchmarks inconsistently across various infrastructures, complicating meaningful comparisons unless they are carefully standardized.

Despite these limitations, Relative Capability Assessments can provide valuable signals about capability levels and potential risks in the frontier model evaluation process.

Section 3

Implementing Bottleneck Assessments

Roadmap:

3.1 Example Evaluations

3.2 Reliable Measurement

3.3 Analyzing Evidence

3.4 Methodology Limitations

Bottleneck Assessments test whether models possess specific capabilities that domain experts believe would remove “bottlenecks” to severe real-world harm. This section describes implementation considerations for Bottleneck Assessments, including:

- Selection of evaluations focused on bottleneck capabilities, drawing on expert-informed identification of technical, knowledge, or operational barriers necessary for causing severe harm (3.1)
- Practices for measurement accuracy, including simulating adversary interactions and benchmarking against reference points (3.2)
- Approaches for analyzing bottleneck evidence against established thresholds, often calibrated to human expert performance levels (3.3)
- Methodology limitations, including the challenges of correctly specifying bottlenecks, adapting to evolving tools, and extrapolating assessment results to comprehensive risk scenarios (3.4)

Developers may adapt these approaches to their unique situations while still effectively testing whether models possess narrow skills or knowledge identified as necessary for causing severe real-world harm.

3.1 Example Evaluations

Characteristics of evaluations for Bottleneck Assessments are:

- **Expert-informed identification of bottlenecks**: This involves incorporating domain experts' analysis of historical precedent to identify current technical, knowledge, or operational barriers that would be concerning for an AI system to overcome. These evaluations require significant domain expertise, often combining internal and external expert input.

- **Reliable proxies for necessary capabilities**: This involves serving as reliable proxies for real-world capabilities while maintaining technical operationalizability. Different threat models require different targeted evaluations—for example, evaluating a model's ability to assist in building a CBRN weapon differs from evaluating assistance in stealing such a weapon.

- **Appropriately calibrated difficulty**: This involves calibrating the test difficulty against known capability levels using methods such as clear baselines (e.g., human expert performance levels) for comparison. This helps establish whether a model's capabilities are sufficient to overcome existing bottlenecks.

Bottleneck Assessments will often use a combination of **benchmarks**, **red-teaming exercises** (structured testing efforts where experts simulate potential attacks or deliberate misuse), and sometimes **controlled trials** (studies where human participants are divided into treatment and control groups to measure models' impact on capabilities). The exact evaluations used evolve rapidly as the field progresses. However, examples of evaluations at the time of writing are:

- Evaluations which test whether models possess specialized knowledge that is necessary for causing severe harm but typically restricted to experts, such as multimodal troubleshooting virology questions compared against medical professional baselines (see Deep Research System Card) or bioweapons knowledge questions assessed against different tiers of human expertise (see Claude 3.7 Sonnet System Card).

- Evaluations which test operational security skills like evasion and attribution avoidance (see Google DeepMind Framework for Evaluating Emerging Cyberattack Capabilities), or whether models possess specialized knowledge necessary for solving specialized "capture the flag" challenges calibrated against experts (see Claude 3.7 Sonnet System Card)

- Evaluations which test autonomous research capabilities, such as tasks requiring implementation of efficient algorithms with defined success metrics or research engineering tasks normalized against expert human solutions to determine capability levels (see METR RE-Bench).

3.2 Reliable Measurement

When assessing potential bottlenecks, developers need to understand what capabilities could be accessed by different actors or a model acting autonomously in well set up conditions, not just what the model demonstrates under standard conditions. This approach (also referred to as "maximal capability evaluations") involves simulating different threat actors interacting with the model depending on the threat model—from more sophisticated adversaries who might have the resources to remove safety restrictions, to less sophisticated adversaries who rely primarily on easily available tools and common jailbreak approaches.

Determining how much elicitation effort is sufficient requires calibrating to what relevant adversaries would realistically achieve given their resources and available alternatives. Teams conducting assessments consider both the investment expected adversaries would likely make and the techniques they could access. For models representing incremental improvements, it may be possible to estimate performance gains from elicitation methods based on previous assessments; for more significant capability jumps, more extensive experimentation may be needed to establish capability bounds.

Some common good practices include:

- Assessing a model that is the same as or very similar to the deployed or soon-to-be-deployed model, and incorporates key fine-tuning measures. Some developers are actively exploring ways to evaluate nearby “snapshots” (model checkpoints during training), and then use automated metrics to verify that subsequent changes to a launch candidate remain within acceptable bounds. However, these techniques are not yet reliable, so developers are conservative in estimating how capabilities might change between later snapshots.

- Working closely with post-training teams to understand model capabilities and elicitation approaches, and which would be realistic for relevant adversaries. Current techniques available to actors in the external ecosystem include systematic prompting strategies, scaffolding tools, and ways to scale inference-time compute such as reasoning, although the range of realistic approaches will likely expand over time.

- Maintaining assessment-only models with reduced safety restrictions (sometimes referred to as "low-refusal models"), potentially including additional fine-tuning for capabilities of concern. Standard safety-trained models mask capabilities that could currently be accessed by even relatively low resource actors through jailbreaks or basic fine-tuning. This approach—testing with minimal restrictions and additional training—accounts for the plausible scenario that motivated actors will be able to access similar capabilities (at least until advances in model security and jailbreak robustness make it significantly harder for adversaries to circumvent safety measures).

- Distinguishing between easily fixable issues like poor prompting or basic tooling problems, limitations that specific adversaries might overcome in particular contexts but not consistently across all scenarios, and more fundamental constraints in capabilities that represent genuine barriers to risk. This helps avoid both underestimating capabilities accessible to motivated actors and overestimating risks from limitations that are unlikely to be exploited.

- Comparing results against known reference points (i.e. “benchmarking”). Unexpected patterns—like highly variable performance on similar tasks—might indicate issues with the assessment approach rather than true capability measurements. When establishing baselines for comparison, developers may find it valuable to consider various design elements that influence measurement. These could include factors such as whether human participants in baseline studies have access to external resources, reasonable time allowances, performance incentives, and appropriate expertise levels for the risk being assessed.

- Incorporating buffer room for technique improvement, and/or considering assessment results as having a limited validity time frame. Elicitation techniques in the external ecosystem continually evolve—often rapidly after a new model release—meaning that capabilities inaccessible at launch may become accessible months later through improved prompting strategies or tools.

Developers may specify when reassessment should occur based on the pace of expected technique advancement in relevant domains.

3.3 Analyzing Evidence

Bottleneck Assessments can establish that pre-mitigation risk is sufficiently low by demonstrating a model lacks one or multiple of these bottleneck abilities. However, one of the most challenging aspects of these assessments is determining exactly what level of model performance constitutes a meaningful and counterfactual reduction in barriers to high-risk scenarios. This requires translating technical metrics from evaluations into real-world judgments about risk potential. Developers may use different methods to determine what level of capability represents a concerning risk. For example, in the CBRN domain:

- Human expert capability (especially PhD-level expertise in domains like virology) can serve as a natural threshold for ruling out risk. When a novice using a model performs well below expert level on key technical tasks, this can be taken as evidence that a bottleneck has not yet been overcome. (This approach assumes that risk only may become unacceptable once capabilities approach expert level, and that novices with significantly sub-expert capabilities would not pose sufficient risk.)

- When available, data from Realistic Threat Simulation Assessments can also help calibrate thresholds by showing what performance on Targeted Bottleneck Assessments correlates with enabling concerning capabilities in the more realistic settings.

- Domain expert judgment may supplement these approaches when direct evidence is limited, with safety margins applied to account for uncertainty.

Bottleneck Assessment analysis is typically documented in reports that capture both the observed capabilities and their implications for threat scenarios and their corresponding bottlenecks. These reports may include explanations of how evaluation thresholds were selected—whether based on expert performance benchmarks or specific capability levels identified through threat scenario analysis; documentation of why certain performance levels do or do not represent removal of a bottleneck; or clear reasoning about the relationship between observed model performance and potential risk impacts.

Teams conducting assessments also analyze potential limitations in the assessment process—whether through under-elicitation, gaps in testing methodology, or edge cases that evaluations might miss. Assessment results may include error margins or confidence intervals to reflect measurement uncertainty, as small differences in performance could affect capability judgments. While still an emerging practice, they may also attempt to estimate how capabilities might evolve—for instance, analyzing current technical limitations and how they might change with improved training or elicitation techniques. However, given the significant uncertainties in predicting AI capability development, these projections should be treated with caution and as indicative only.

If a model is able to surpass all identifiable bottlenecks in a threat scenario, developers may implement precautionary safeguards, or include models for further assessments (e.g., a Threat Simulation Assessment). In these cases, information from Bottleneck Assessments can help to target safeguards or focus follow on assessments.

3.4 Methodology Limitations

As with all evaluation methods, Bottleneck Assessments (and the benchmarks and red-teaming exercises they rely on) have several notable limitations:

- As with Relative Capability Assessments, Bottleneck Assessments focus on specific tasks and capabilities, and do not reflect the full context or a holistic view of the risk (including infrastructure and mitigations outside of the AI development chain that reduce risk).
- Specific bottlenecks may be incorrectly specified such that a model failing the Bottleneck Assessment is an incorrect proxy for capability. For example, one agreed upon bottleneck for CBRN uplift is the task of ideation in developing novel bio-threats. A model failing a Bottleneck Assessment for ideation tasks may still contribute to CBRN uplift in other unevaluated ways.
- Within a specific Bottleneck Assessment, the unit of analysis and elicitation methods play a crucial role in representativeness of real-world risk. For example, a model with access to the internet and other tools will perform differently on the same Bottleneck Assessment compared to a model without access to these tools. In practice it can be difficult to appropriately elicit a model as tools and methods change rapidly over time, and a model may fail Bottleneck Assessments with access to tools available today, but pass the same Bottleneck Assessment with access to tools released in the future.
- Bottleneck Assessments typically evaluate individual models in isolation, potentially missing risks that emerge when multiple models with complementary capabilities are used together. For example, if one model excels at idea generation but struggles with implementation details, while another model has the opposite strengths, adversaries could potentially combine these models to bypass bottlenecks that each model individually fails to overcome.

Even so, Bottleneck Assessments provide important signals about risk and capabilities and make frontier evaluations more widely accessible and comparable across the industry.

Section 4

Implementing Threat Simulation Assessments

Roadmap:

Threat Simulation Assessments involve "simulating" substantial segments of threat scenarios from end-to-end, in order to more directly estimate the extent to which the model would enable these scenarios. This section describes implementation considerations for Simulation Assessments, including:

- Designing scenario-realistic evaluations that balance representative test conditions with appropriate safety safeguards (4.1)
- Establishing reliable measurements through robust experimental design, appropriate control groups, and sufficient statistical validity (4.2)
- Analyzing simulation evidence to estimate real-world risk, despite the significant challenges in extrapolating from controlled experiments (4.3)
- Methodology limitations, including constraints on simulation realism, participant selection biases, and difficulties generalizing to comprehensive risk scenarios (4.4)

Developers may adapt these approaches to their unique situations while still effectively approximating complete threat scenarios and estimating how model performance would affect real-world risk.

.

4.1 Example Evaluations

Characteristics of evaluations for Threat Simulation Assessments are:

- **Scenario realism**: This involves creating test conditions that closely approximate real-world threat scenarios, including relevant constraints, tools, expertise levels, and contextual factors that would affect a real-world capability deployment.
- **Safety-conscious design**: This involves implementing controlled experiments that balance realistic testing with appropriate security measures, such as sandboxed environments (isolated testing spaces with strict limitations on external access and impacts), safe proxies for harmful tasks, careful participant selection, and proper containment of potentially dangerous outputs.

Threat Simulation Assessments may use **controlled trials** for evaluating human misuse risks, while using different approaches for autonomous risks. While the exact methodologies used are likely to evolve rapidly, some examples are:

- Comparing the success rates and quality of outputs between biology novices with AI assistance versus control groups using only standard online resources when completing technical tasks relevant to pathogen engineering.
- Testing whether advanced models can autonomously discover and exploit novel vulnerabilities in software test environments, developing end-to-end attack chains without human guidance.

4.2 Reliable Measurement

It is important to acknowledge that these assessments have been under development for less time than other approaches, with best practices still emerging for reliable measurement. Some early candidate practices include:

- Establishing appropriate control groups to isolate the specific impact of AI assistance compared to baseline capabilities.
- Designing experiments with sufficient sample sizes, randomization protocols, and blinding procedures where appropriate.
- Creating consistent testing protocols across participant groups while consulting domain experts to validate experimental relevance to real threat scenarios.
- Running multiple independent trials under varying conditions to test consistency and identify key influencing factors.

4.3 Analyzing Evidence

Threat Simulation Assessments can establish that pre-mitigation risk is sufficiently low, by either demonstrating that—even if it performs well on narrow tasks—the model is ineffective in realistic settings, or that the implied quantitative increase in risk is still relatively small. (Additionally, these assessments often serve broader purposes beyond evaluating specific models, such as improving threat model understanding, calibrating Bottleneck Assessments, and providing empirical data to inform future safety approaches.)

Extrapolating from even the most realistic threat simulations to quantitative estimates of real-world risk remains fundamentally difficult. At the time of writing there is no established consensus on methodology, and any approach involves significant judgment in the face of deep uncertainty. Nevertheless, researchers have begun exploring several analytical frameworks to interpret results:

- Measuring the difference in performance between model-assisted participants and control groups using only conventional resources provides a direct measure of capability uplift.
- Combining capability uplift data with other relevant variables—such as the number of potential actors, their likelihood of pursuing harmful goals, and the probability of success in real-world conditions—can translate evaluation results into more comprehensive risk estimates.
- Structured surveys of domain experts and forecasters can help assess how specific capability demonstrations might change the probability of harmful outcomes.

For models that do perform effectively in simulated threat scenarios, these assessments can assist in quantitatively estimating risk and evaluate whether proposed safeguards would be adequate.

4.4 Methodology Limitations

While simulation assessments are the most realistic evaluation strategy, they are not completely representative of real-world threats.

- Negative assessment results should be interpreted cautiously, as they may reflect limitations in the experimental design rather than true capability limitations—models might still pose risks in slightly different scenarios or with larger sample sizes that could reveal capabilities masked by statistical noise.
- Often, testing the full extent of the risk poses legal and safety challenges (such as in CBRN and offensive cyber attacks) and thus simulation assessments often focus on proxies, albeit in a more realistic way than other evaluation strategies.
- One common approach for simulation evaluations is comparative uplift studies, such as for novice skill uplift in CBRN. These studies give important signals about risk but need to be appropriately contextualized within the holistic risk and mitigation landscape and broader ecosystem, as the infrastructure used to create a rigorous evaluation inherently removes some of the obstacles malicious actors would face (e.g., goal specification, access to wetlabs, or access to materials).
- Lastly, these studies are extremely resource intensive, and may not be feasible (or even insightful) to apply to every model. Instead, these studies may be best used to assess ecosystem risk periodically (instead of as related to a specific model release), and that overall risk can then be transposed to representative models.

Despite these limitations, Threat Simulation Assessments remain the most direct measure for risk widely available and continued collaboration and investment on these assessments is valuable.

Section 5

Assessments Across the Development Process

Roadmap:

This section outlines approaches for integrating capability assessments throughout the AI development lifecycle, from early training to deployment. Implementation considerations for timing and sequencing assessments are described in this section, including:

- How high-risk categories of capabilities, planned increases in model size, and deployment reversibility can affect assessment timing needs (5.1)
- Methods for identifying which training runs warrant more extensive evaluation (5.2)
- Leading indicators that can provide early warning of emerging capabilities (5.3)
- Realistic testing timelines for different assessment types and capability scenarios (5.4)
- Approaches for pre-launch testing and managing early controlled deployments (5.5)

Developers may adapt these approaches to their specific development workflows while still maintaining effective assessment coverage across the model lifecycle.

.

5.1 Assessment Timing Considerations

When integrating capability assessments throughout development, timing considerations vary based on both model characteristics and types of potential risk.

- Certain high-risk capability categories may require more immediate attention during development. Risks related to potential misuse by low-resource actors typically require wide deployment to become significant concerns and can be addressed closer to public release. However, some capabilities pose unique risks: autonomous capabilities, which might enable a model to take unanticipated actions even in controlled settings; deceptive capabilities, which could undermine assessment processes themselves; and capabilities that would be particularly dangerous in the hands of sophisticated adversaries, could all pose risks even before deployment if model weights could be compromised during the training process or or through other attack vectors that don't require broad external deployment.

- The scale of capability jumps also matters significantly. Models representing incremental improvements over previously evaluated systems often require less intensive assessment compared to those demonstrating substantial capability advances. Large, rapid capability improvements warrant earlier and more thorough evaluation because they may introduce entirely new capabilities or risks that weren't present in previous versions. Developers that increase model capabilities in relatively small and planned increments or carefully monitor model capabilities typically face fewer unexpected capability jumps requiring rapid assessment.

- The reversibility of deployment is also a consideration. Deployments that can be quickly modified allow for gathering safety evidence while maintaining intervention options. However, this strategy primarily helps with risks that escalate continuously rather than those causing sudden, severe harm without warning.

Earlier assessment can also benefit developers by enabling more efficient planning of safeguards and potentially reducing the total evaluation burden if sufficient safety evidence is established earlier in development. However, assessing models during training presents technical challenges that require further research.

5.2 Identifying Potential Frontier Training Runs

Building on the timing considerations above, developers may develop systematic approaches for identifying which planned training runs will warrant assessment. Some developers, with distributed research groups, may establish heuristics for identifying relevant training runs. This may include a basic Floating Point Operations (FLOPs) threshold, below which models can bypass Frontier Capability Assessments, unless they have specialized capabilities that warrant closer examination (e.g., models trained primarily on biological sequence data).

Research and assessment teams often collaborate to identify pre-training and post-training runs which may produce significant capability jumps. This assessment considers factors associated with larger capability advances, such as: whether the model is planned as a major launch in the developer's most capable product series; significant architectural or algorithmic improvements ("compute multipliers"); supported modalities; and expectations for compute requirements.

Models that are expected to remain well below frontier levels on broad indicators such as general reasoning benchmarks (e.g., GPQA) and real-world task completion typically bypass Frontier Capability Assessments, as this early screening itself serves as sufficient evidence of lower risk. Other models, including cases where there may be uncertainty from changeable training plans, are assessed as described in earlier sections.

If the team conducting assessments expects a model may significantly exceed current frontier capabilities, or might develop high-risk capabilities, they may conduct ex-ante threat modeling to better understand possible risks. This can lead to recommendations for additional safeguards, enhanced monitoring during training, or pre-emptive strategies like modifying training data.

5.3 Leading Indicators

For model developers with a high volume of development, lightweight Relative Capability Assessments (sometimes called leading indicators assessments) can provide early indication of capability jumps. These developers may implement continuous monitoring during training to identify models potentially representing significant advancements beyond the current frontier. These "leading indicators" can help developers to detect which models may be in scope for more rigorous assessment under a frontier capability evaluation framework.

Similar or the same types of evaluations to those used in Relative Capability Assessments—like perplexity measurements or benchmark performance—can serve this monitoring function throughout the training process, provided they are not expected to become saturated by the new model. While it is not yet possible to reliably predict how specific benchmark improvements will translate to real-world capabilities, significant departures from historical patterns between model generations can still be informative. The best indicators may vary between training runs, depending on which most effectively capture meaningful capability changes for that particular architecture or training approach. To track these patterns effectively, some developers may choose to establish collaboration between assessment and development teams, enabling ongoing monitoring of key indicators during development. Although not currently established practice, some developers have also begun to develop dashboards where these indicators can be monitored in real time to identify models that exhibit capability jumps.

Some developers have also begun exploring evaluation designs that can provide insights into potential future capabilities, creating tasks with graduated difficulty levels or decomposing complex challenges into component skills. For example, partial credit scoring systems might track incremental progress on difficult tasks, potentially offering earlier signals about capability development trajectories.

For models showing only incremental, predictable improvements, these indicators provide reassurance that a model remains within expected parameters and can proceed with standard assessment timelines. However, if leading indicators show significant jumps in risk-relevant areas, this may warrant immediate investigation and more thorough testing. If further testing confirms concerning capabilities, developers may implement enhanced security measures for the remainder of training, update access controls, or deploy additional technical guardrails before any internal deployments. In some cases, more thorough testing may also be warranted if leading indicators show unexpectedly low performance, as this may suggest issues with the evaluation setup or deliberate underperformance by the model.

5.4 Testing Timelines

When leading indicators suggest a model represents only incremental improvement, developers may proceed with streamlined assessment processes. Models showing significant capability jumps or potential high-risk capabilities may require more thorough evaluation near the end of development.

With current methods, Relative Capability Assessments, which use mostly automated benchmarks, can be completed within days and are often sufficient for sub-frontier models and models showing only incremental improvements. Bottleneck Assessments, especially for models approaching thresholds of concern, may proceed through several phases, such as: running more automated evaluations, running more manual evaluations, technical analysis, expert review, and sometimes iterative improvement of capability elicitation methods. For models approaching thresholds of concern, current Bottleneck Assessment methods can take anywhere from a few days to a couple of weeks from receiving the testing candidate, with additional time if initial findings indicate areas requiring deeper investigation, broader expert consultation, or progression to a Threat Simulation Assessment. Developers aim to give teams conducting assessments sufficient time to complete their deadlines without compromising assessment quality. Testing windows should build in some flexibility to accommodate potential unexpected findings and expert consultation. Some developers may also choose to deploy preemptive safeguards designed to be sufficient across the range of possible capability outcomes, and conduct further detailed assessments in parallel with limited deployment.

Future advances, especially increasing automation of evaluations, may help streamline these phases, and developers adopt improved processes as they become available.

5.5 Pre-Launch Testing and Early Deployments

In some cases, teams will complete Frontier Capability Assessments either during training (where possible) or after training and before release. However, when leading indicators or earlier stages of assessment indicate minimal risk, developers may proceed with some deployments in parallel with continued in-house evaluations.

These limited deployments may include sharing models with red-teamers, providing early access to trusted partners, conducting controlled user experiments comparing model versions, or applying low rate-limits or other product-level interventions depending on the threat model, and typically present low risk for threat scenarios involving misuse by low-resource malicious actors, especially when access is controlled and relationships are trusted. For models that have shown significant capability jumps or that might develop concerning capabilities in high-risk categories, developers may decide to limit access to essential teams until appropriate safeguards are installed and confirming robust monitoring systems are in place during all testing phases.

.

Section 6

Organizational Considerations

Roadmap:

6.1 Organizational Separation of Duties

6.2 Preparation for Assessments

6.3 Decision-Making Accountability

This section outlines organizational considerations for running effective Frontier Capability Assessments, including:

- Maintaining appropriate separation of duties between assessment and development teams to increase objectivity (6.1)
- Confirming adequate preparation and resourcing for assessment activities to enable thorough and timely evaluations (6.2)
- Establishing clear accountability for decision-making based on assessment results to enable appropriate action on findings (6.3)

Developers may adapt these organizational approaches to their unique contexts while still maintaining effective governance structures for frontier model assessment.

6.1 Organizational Separation of Duties

Most frontier model developers maintain assessment teams that are separate from model development teams. Some developers may also choose for certain deployments to implement additional internal oversight through dedicated methodology reviewers, who further examine assessment methods, analysis, and conclusions. These reviewers may verify methodology, challenge assumptions, and/or identify when different approaches might lead to varying conclusions. This process may be helpful for complex assessments where multiple perspectives improve confidence in results.

6.2 Preparation for Assessments

Developers typically take steps to support adequate preparation before beginning assessments, such as reliable model access for red teamers and sufficient compute. These steps can help teams conducting assessments to gather and analyze evidence on the required timelines.

Developers may also maintain the option to bring in external domain experts to supplement internal skill sets during more complex assessments. For example, they may contract expert red-teamers or specialized domain experts to assist with analysis of results.

Some developers may also at times choose to implement preregistration processes where they document their planned methodology before beginning an assessment. This may include specifying research questions, evaluation protocols, scoring criteria, statistical analysis plans, and the format of expected results. Preregistration can help prevent confirmation bias, "fishing" for specific results, and retroactive adjustment of evaluation criteria by establishing clear evaluation parameters before assessment begins, allowing time for methodological review, and reducing the risk of post-hoc adjustments that might bias conclusions. However, given the nascence and rapid evolution of frontier evaluation science, limitations in experiment design are often uncovered for the first time during evaluations, and thus pre-registration may sometimes be undesirable.

6.3 Decision-Making Accountability

Decision-making processes vary across developers and typically become more structured as assessment complexity increases. Developers may choose to establish clear separation between those who conduct assessments and those who make final decisions based on assessment results. This separation between assessment teams and final decision-makers provides technical experts with room to evaluate safety considerations without being directly subject to project pressures.

Typically, these designated decision-makers have clear accountability for outcomes, sufficient authority to make consequential go/no-go decisions regarding model deployment or training continuation, the standing to balance competing organizational priorities, and formal responsibility for accepting residual risk in these areas. Assigning clear accountability at the appropriate level means high-stakes safety determinations receive proper attention, and avoids a situation where diffuse responsibility across teams results in a lack of overall accountability.

Section 7

Continuing Work

> Looking ahead, developers may increasingly need to consider capabilities that might undermine assessment reliability itself.

This report draws on and consolidates insights from a range of existing publications and frameworks. These include safety frameworks published by Frontier Model Forum members, the UK AI Security Institute’s "Early Lessons from Evaluating Frontier AI Systems," the U.S. AI Safety Institute's "Managing Misuse Risk for Dual-Use Foundation Models" (NIST AI 800-1), Google’s paper “Model evaluation for extreme risks,” RAND Corporation's working paper "Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models,” METR’s “Guidelines for capability elicitation” and “Safety Framework for Pre-Frontier AI,” and recommendations from the Joint California Policy Working Group on AI Frontier Models.

Continuing work is underway across industry and elsewhere to make Frontier Capability Assessments more reliable and informative. As model capabilities advance, assessment methodologies must evolve alongside them. Industry collaboration will continue to be important for developing robust evaluation standards.

Better automated evaluations and training metrics could help identify concerning models before investing in resource-intensive assessments. This also enables more proactive risk mitigation—developers can plan appropriate assessment strategies, establish necessary expertise and resources, and implement additional safeguards if needed. However, the science of capability prediction remains nascent, and extrapolating final capabilities from early training snapshots is challenging.

Observing previous models in their deployed context (especially open-sourced ones) may also provide insights about capabilities, risks, and efficacy of mitigations (including system-level mitigations) as a wider community interacts with the model. These insights can inform future Frontier Capability Assessments and interpretations of results.

This may be especially valuable when assessing qualitatively new capabilities. When a capability has existed in other deployed models, developers can observe how it affects real-world usage patterns—what types of tasks it meaningfully assists with, and how users integrate it into workflows.

Looking ahead, developers may increasingly need to consider capabilities that might undermine assessment reliability itself. While some current models only demonstrate basic ability to manipulate evaluation results when explicitly prompted to do so, future systems may develop more sophisticated forms of situational awareness (e.g., recognizing they are being evaluated) that could result in deceptive behavior. Though autonomous "sandbagging" (intentionally performing poorly on evaluations) does not yet appear to be a significant risk, some developers are beginning to incorporate preliminary countermeasures such as cross-checking model answers or analyzing reasoning patterns.

Future advances in model interpretability and transparency methods also hold promise for improving assessment quality. Research into analyzing model internals or using logit-based methods could offer deeper insights into model capabilities and potential risks.

Lastly, this document describes various assessment approaches but does not provide detailed guidance on how to determine when model capabilities become concerning enough to warrant additional safeguards or constraints. Future work should focus on developing more rigorous methodologies for establishing these assessment criteria, taking into account both empirical evidence and normative considerations about acceptable risk levels.

We welcome engagement with this and forthcoming technical reports from across the frontier AI safety and security ecosystem. Researchers and organizations interested in further refining and harmonizing the implementation of safety frameworks are invited to reach out to the Frontier Model Forum.

Contact us at: info@frontiermodelforum.org.