Meta

Frontier AI Framework

Version 1.1

Advanced AI models and systems present an enormous opportunity for the world, both socially and economically. At Meta, we believe that the best way to make the most of that opportunity is by building state-of-the-art AI, and releasing it to a global community of researchers, developers, and innovators.

We're committed to advancing the state of the art in AI, on models themselves and on systems to deploy them responsibly, to realize that potential. While it's not possible to entirely eliminate risk, if we want this AI to be a net positive for society we believe it's important to work internally and, where appropriate, with governments and outside experts to take steps to anticipate and mitigate severe risks that it may present.

This Frontier AI Framework describes how Meta works to build advanced AI, including by evaluating and mitigating risks and establishing thresholds for catastrophic risks. Frontier AI frameworks are a relatively new type of policy instrument and there can be variation in how terminology is used. For that reason, we have included an appendix with definitions for a number of key terms that are important to understand when reading our framework.

The science of AI evaluation is nascent, and researchers in companies, academia, and government are working to develop more robust and quantitative measurement of risks and benefits of AI. As a result, we expect our approach to evaluating and mitigating risk – including the approaches outlined in this document – to evolve and mature over time. We hope that sharing our current approach to development of advanced AI systems will not only promote transparency into our decision-making processes but also encourage discussion and research on how to improve the science of AI evaluation and the quantification of risks and benefits.

How to read this document

This document contains five sections:

1. Introduction

This section outlines the scope of this iteration of our Frontier AI Framework.

2. Governance & transparency

This section outlines our general approach to AI governance and transparency. Sections 3 and 4 provide more detail on how specific elements of this governance approach are implemented for frontier AI.

3. Outcomes & thresholds

In this section we explain our outcomes-led approach to defining risk thresholds for frontier AI. We define catastrophic outcomes in two domains: Cybersecurity and Chemical & Biological risks.

4. Implementation

In this section we explain the process we follow to measure and manage risks from frontier AI, and the processes we follow when determining how to safely develop and release models.

5. Future work

In this section, we outline areas where we plan to focus research efforts and investment to improve our ability to implement this Framework, and safely release advanced AI for the benefit of all.

Section 01

Introduction

1.1 Scope

In line with the Frontier AI Safety Commitments, which Meta signed in May 2024, our Frontier AI Framework relates to our forthcoming models and systems that exceed the capabilities present in the most advanced models. It defines processes to manage and mitigate the risk of frontier AI models or systems producing catastrophic outcomes, and to keep risks of such outcomes within tolerable levels. This Framework is one component of our wider AI governance program. It deals with catastrophic outcomes that could arise as a direct result of the development or release of the frontier AI model. The Framework does not, therefore, reflect the full spectrum of risks that we assess for, nor all of the evaluations that we conduct.[1]

Our Framework is structured around a set of *catastrophic outcomes*. We have used threat modelling to develop threat scenarios pertaining to each of our catastrophic outcomes. We have identified the key capabilities that would enable the threat actor to realize a threat scenario. We have taken into account both state and non-state actors, and our threat scenarios distinguish between high- or low-skill actors.

We define our thresholds based on the extent to which frontier AI would uniquely enable the execution of any of the threat scenarios we have identified as being potentially sufficient to produce a catastrophic outcome. If a frontier AI is assessed to have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1. Our high and moderate risk thresholds are defined in terms of the *level of uplift* a model provides towards realizing a threat scenario. We will develop Frontier AI in line with the processes outlined in this Framework, and implement the measures outlined in Table 1. Section 3 on Outcomes & Thresholds provides more information about how we define our thresholds.

[1] As an example of the types of evaluations we conduct for AI models, see our work on our Llama 3 Herd of Models.

This is the first iteration of our Frontier AI Framework. We expect to update it in the future to reflect developments in both the technology and our understanding of how to manage its risks and benefits. Alongside updates to the Framework, we also identify areas that would benefit from further research and investment to improve our ability to continue to safely develop and release advanced AI models.

Section 02

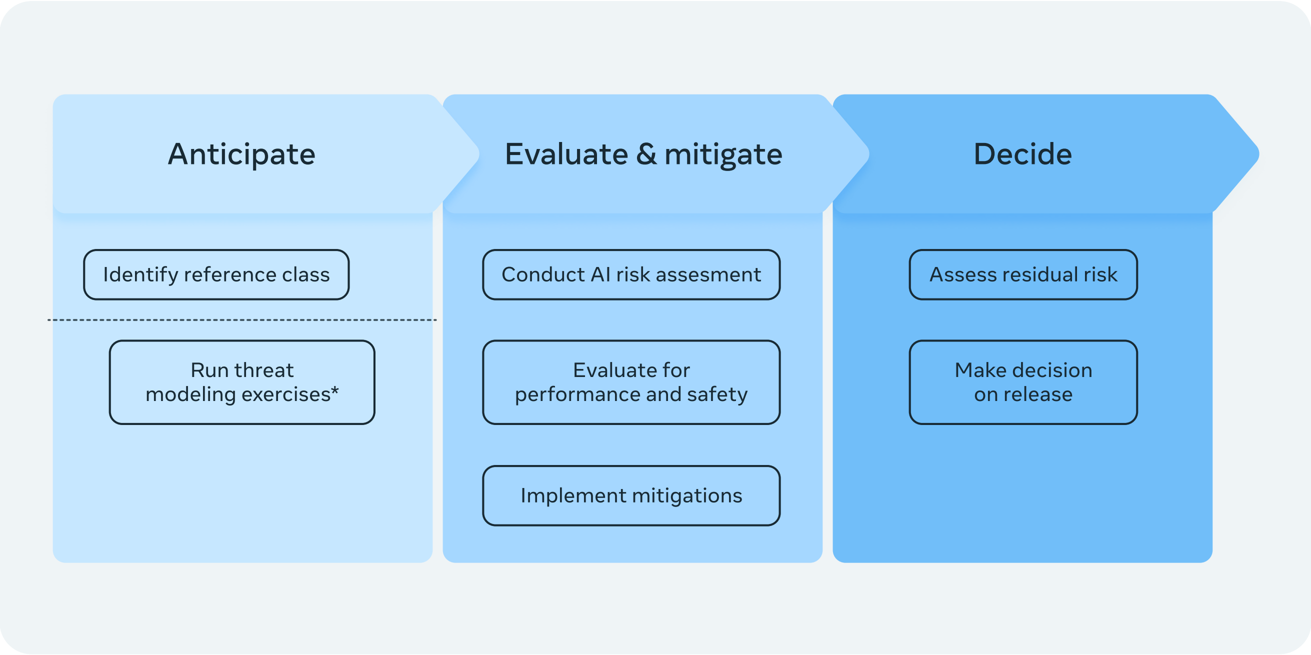
Governance & transparency

We have been developing, deploying, and open sourcing AI research and models for over a decade through both our Fundamental AI Research (FAIR) Lab and our product teams, which leverage AI across our suite of products and services, including in Facebook, Instagram, Messenger, WhatsApp, and ads and other business products. In addition to research releases, we have a growing ecosystem of open models and safety tools that can be used for both research and commercial use cases. This Framework builds upon the processes and expertise that have guided the responsible development and release of our research and products over the years. The processes outlined in this Framework describe our approach to developing and releasing Frontier AI specifically.

2.1 AI governance

This section provides an overview of the processes we follow when developing and releasing frontier AI to ensure that we are monitoring and managing risk throughout. Our governance approach can be split into three main stages: plan; evaluate and mitigate; and decide.

Findings at any stage might prompt discussions via our centralized review process, which ensures that senior decision-makers are involved throughout the lifecycle of development and release.



** We run threat modelling exercises periodically, but they are not necessarily a part of every single release, see the section 'Run threat modelling exercises' for more detail.*

2.1.1 Anticipate

Identify comparable models to use as a reference class

For a given model, we discuss what we plan to build in terms of, for example, the capabilities we anticipate it will have, supported modalities, intended uses and anticipated benefits of the model, and expectations for compute requirements. We compare these various factors against our own models and those available externally. This allows us to identify an estimated 'reference class' of comparable models that we use throughout development to track how our model is performing and to inform what evaluations we may conduct and mitigations we might implement.

If we expect that a model may significantly exceed current frontier capabilities, we will conduct an *ex-ante threat modelling exercise* to help us determine whether this model may pose novel risks (see more below).

Run threat modelling exercises

In addition to our AI risk assessment (see below), which covers known potential risks, we conduct periodic threat modelling exercises as a proactive measure to anticipate catastrophic risks from our frontier AI. In the event that we identify that a model can enable the end-to-end execution of a threat scenario for a catastrophic outcome, we will conduct a threat modelling exercise in line with the processes in Section 3.2.

The exact format of these exercises may vary. The general process is as follows:

1. Host workshops with experts, including external subject matter experts where relevant, to identify new catastrophic outcomes and/or threat scenarios.[2]
2. If new catastrophic outcomes and/or threat scenarios are identified, design new assessments to test for them, in consultation with external experts where relevant.

2.1.2 Evaluate and mitigate

Conduct an AI risk assessment

Our AI risk assessment process systematically evaluates potential risks associated with frontier AI, documenting mitigation strategies and residual risks across a set of applicable risk categories.

The risk assessment process involves multi-disciplinary engagement, including internal and, where appropriate, external experts from various disciplines (which could include engineering, product management, compliance and privacy, legal, and policy) and company leaders from multiple disciplines.

The risk assessment also considers the planned release (i.e. closed deployment, limited release, or full release), as this informs the type of pre-release evaluation we undertake.

Evaluate for performance and safety

Our evaluations can involve a combination of automated and human evaluations, as well as red teaming and uplift studies. Throughout development, we monitor performance against our expectations for the reference class as well as the enabling capabilities we have identified in our threat scenarios, and use these indicators as triggers for further evaluations as capabilities develop.

AI model evaluation is a nascent science, and as capabilities develop new evaluations are developed. As such, we do not have a fixed set of evaluations that we apply to each frontier AI. Rather, we implement relevant evaluations based on capabilities and the latest research.

[2] For certain types of catastrophic risk, this will necessarily include working with government officials, who have the specific knowledge and expertise to enable proper assessment.

As an example, once a model demonstrates a sufficient standard of coding ability, we would typically evaluate the potential of the model to present cybersecurity risks. While we expect that the appropriate evaluations for cybersecurity will change over time, we have developed and open sourced an evaluation called CyberSecEval that is designed for this purpose. For both cyber and chemical and biological risks, we conduct red teaming exercises once a model achieves certain levels of performance in capabilities relevant to these domains, involving external experts when appropriate.

We design our evaluations to account for how the model will be released, including assessing how its capabilities might be enhanced. See section 4.2 for more details.

Implement mitigations

Our mitigation strategy is informed by the risks we've identified in the risk assessment, evaluation results, the mitigations that have been applied to existing models in the same class, and the release approach. Our Llama research papers provide more details on mitigations we have implemented for previous releases. Section 4 of this framework provides more details on mitigation techniques we employ.

2.1.3 Decide

Assess residual risk

We assess residual risk, taking into consideration the details of the risk assessment, the results of evaluations conducted throughout training, and the mitigations that have been implemented.

Make a decision on release

The residual risk assessment is reviewed by the relevant research and/or product teams, as well as a multidisciplinary team of reviewers as needed. Informed by this analysis, a leadership team will either request further testing or information, require additional mitigations or improvements, or they will approve the model for release.

2.2 Transparency

One of the major benefits of an open approach to AI research and development is that it provides a greater degree of transparency as to how a model works, which in turn can lead to a better understanding of, and trust in, AI. We see this as a key benefit of sharing model weights, as well as research papers and model cards.

In line with the processes set out in this Framework, we intend to continue to openly release models to the ecosystem. We also plan to continue sharing relevant information about how we develop and evaluate our models responsibly, including through artefacts like model cards and research papers, and by providing guidance to model deployers through resources like our Responsible Use Guides.[3]

In addition to promoting accountability, open sourcing advanced models makes it possible for us to not only work with outside experts to improve our own evaluation of risk but also for the broader community to independently assess the capabilities of our models. Given the iterative nature of AI development, we believe that this will not only help improve the efficacy, safety, and trustworthiness of our models but improve the state of the art in risk evaluation more generally.

3

Section 03

Outcomes & thresholds

3.1 An outcomes-led approach

A key component of our Frontier AI Framework is a set of thresholds that prompt particular measures and which, in the extreme, involve restricting the development or release of frontier AI until risks can be mitigated. Different approaches to defining thresholds are emerging. In some cases, thresholds have been defined in terms of a particular capability or set of capabilities, while others also include examples of how these capabilities could be weaponized.[4]

Advanced AI capabilities can be used for good and for ill, and they can be applied across different domains. To manage this feature of AI capabilities, and to consider catastrophic risk in a systematic way, we have adopted an outcomes-led approach.

[3] For example, see our Llama 3 Model Card, Research Paper, and Responsible Use Guide.
[4] OpenAI former, GDM latter

We start by identifying a set of catastrophic outcomes we must strive to prevent, and then map the potential causal pathways that could produce them. When developing these outcomes, we've considered the ways in which various actors, including state level actors, might use/misuse frontier AI. We describe threat scenarios that would be potentially sufficient to realize the catastrophic outcome, and we define our risk thresholds based on the extent to which a frontier AI would uniquely enable execution of any of our threat scenarios.

By anchoring thresholds on outcomes, we aim to create a precise and somewhat durable set of thresholds, because while capabilities will evolve as the technology develops, the outcomes we want to prevent tend to be more enduring. This is not to say that our outcomes are fixed. It is possible that as our understanding of frontier AI improves, outcomes or threat scenarios might be removed, if we can determine that they no longer meet our criteria for inclusion. We also may need to add new outcomes in the future. Those outcomes might be in entirely novel risk domains, potentially as a result of novel model capabilities, or they might reflect changes to the threat landscape in existing risk domains that bring new kinds of threat actors into scope. This accounts for the ways in which frontier AI might introduce novel harms, as well its potential to increase the risk of catastrophe in known risk domains.

An outcomes-led approach also enables prioritization. This systematic approach will allow us to identify the most urgent catastrophic outcomes – i.e., within the domains of cybersecurity and chemical and biological weapons – and focus our efforts on avoiding these outcomes rather than spreading efforts across a wide range of theoretical risks from particular capabilities that may not plausibly be presented by the technology we are actually building.

3.2 Threat modelling

Threat modelling is fundamental to our outcomes-led approach. We run threat modelling exercises both internally and with external experts with relevant domain expertise, where required. The goal of these exercises is to explore, in a systematic way, how frontier AI models might be used to produce catastrophic outcomes. Through this process, we develop threat scenarios' which describe how different actors might use a frontier AI model to realize a catastrophic outcome.[5]

[5] We aim to be as methodical and rigorous as possible in our threat modelling. However, it is important to acknowledge that we cannot claim to have anticipated *all* potential threat scenarios. There is always a potential for 'unknown unknowns'. We anticipate and mitigate catastrophic risks to the best of our ability.

We design assessments to simulate whether our model would uniquely enable these scenarios, and identify the enabling capabilities the model would need to exhibit to do so. Our first set of evaluations are designed to identify whether all of these enabling capabilities are present, and if the model is sufficiently performant on them. If so, this would prompt further evaluation to understand whether the model could uniquely enable the threat scenario.[6] See Section 4.2 for more detail.

It is important to note that the pathway to realize a catastrophic outcome is often extremely complex, involving numerous external elements beyond the frontier AI model. Our threat scenarios describe an essential part of the end-to-end pathway. By testing whether our model can uniquely enable a threat scenario, we're testing whether it uniquely enables that *essential* part of the pathway. If it does not, then we know that our model cannot be used to realize the catastrophic outcome, because this essential part is still a barrier. If it does and cannot be further mitigated, we assign the model to the critical threshold.

This would also trigger a new threat modelling exercise to develop additional threat scenarios along the causal pathway so that we can ascertain whether the catastrophic outcome is indeed realizable, or whether there are still barriers to realizing the catastrophic outcome (see Section 5.1 for more detail).

Our threat modelling is informed by our own internal experts' assessment of the catastrophic risks that frontier models might pose, as well as engagements with governments, external experts, and the wider AI community. However, there remains quite considerable divergence in expert opinion as to how AI capabilities will develop and the time horizons on which they could emerge.

To further clarify how we have determined the catastrophic outcomes that are in scope for this iteration of our Framework, we include a set of criteria for inclusion and omission below. These criteria are designed to enable a Framework that is implementable, and that allows us to make evidence-led decisions about development and release.

[6] Evaluation is a nascent science, and we expect false positives. This further evaluation might involve repeating capabilities evaluations to validate the initial results.

For this Framework specifically, we seek to consider risks that satisfy all four criteria:

CRITERIA	
Plausible	It must be possible to identify a causal pathway for the catastrophic outcome, and to define one or more simulatable threat scenarios along that pathway. This ensures an implementable, evidence-led approach.
Catastrophic	The outcome would have large scale, devastating, and potentially irreversible harmful effects.
Net new	The outcome cannot currently be realized as described (e.g. at that scale / by that threat actor / for that cost) with existing tools and resources.
Instantaneous or irremediable	The outcome is such that once realized, its catastrophic impacts are immediately felt, or inevitable due to a lack of feasible measures to remediate.

3.3 Risk thresholds

We define our risk thresholds based on the extent to which a frontier AI would uniquely enable execution of any of our threat scenarios. A frontier AI is assigned to the critical risk threshold if we assess that it would *uniquely enable* execution of a threat scenario. If a frontier AI is assessed to have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1. Our high and moderate risk thresholds are defined in terms of the level of uplift a frontier AI provides towards realizing a threat scenario. We will develop these models in line with the processes outlined in this Framework, and implement the measures outlined in Table 1.

Our outcomes-led approach allows us to avoid over-ascribing risk based on the presence of a particular capability alone, and instead assesses the potential for the frontier AI to actually enable harm. This approach is designed to effectively anticipate and mitigate catastrophic risk from frontier AI without unduly hindering innovation of models that do not pose catastrophic risks and can yield enormous benefits. For frontier AI that falls below the critical threshold, we will take into account both potential risks and benefits when determining how to develop and release these models. Section 4.4 explains this in more detail.

We also define processes we will implement to keep risks within tolerable levels. The table below sets out our thresholds and our rationale, and includes an overview of the processes that we will follow at each threshold. See Section 4 for more detail on the evaluations we perform to make that assessment.

Table 1: Risk thresholds for frontier AI

RISK THRESHOLD		SECURITY MITIGATIONS	MEASURES
Critical	**Stop development** The model would uniquely enable the execution of at least one of the threat scenarios that have been identified as potentially sufficient to produce a catastrophic outcome and that risk cannot be mitigated in the proposed deployment context.	Access is strictly limited to a small number of experts, alongside security protections to prevent hacking or exfiltration insofar as is technically feasible and commercially practicable.	• Successful execution of a threat scenario does not necessarily mean that the catastrophic outcome is realizable. If a model appears to uniquely enable the execution of a threat scenario we will pause development while we investigate whether barriers to realizing the catastrophic outcome remain. • Our process is as follows: a. Implement mitigations to reduce risk to moderate levels, to the extent possible b. Conduct a threat modelling exercise to determine whether other barriers to realizing the catastrophic outcome exist c. If additional barriers exist, update our Framework with the new threat scenarios, and re-run our assessments to assign the model to the appropriate risk threshold d. If additional barriers do not exist, continue to investigate mitigations, and do not further develop the model until such a time as adequate mitigations have been identified.
High	**Do not release** The model provides significant uplift towards execution of a threat scenario (i.e. significantly enhances performance on key capabilities or tasks needed to produce a catastrophic outcome) but does not enable execution of any threat scenario that has been identified as potentially sufficient to produce a catastrophic outcome.[7]	Access is limited to a core research team, alongside security protections to prevent hacking or exfiltration.	• Implement mitigations to reduce risk to moderate levels.
Moderate	**Release** The model would not provide significant uplift towards execution of a threat scenario.	Security measures will depend on the release strategy.	• Mitigations will depend on the results of evaluations and the release strategy.

[7] At present, the science of evaluation is not sufficiently robust as to provide definitive quantitative metrics for uplift. Our assessment of whether a model exhibits significant uplift is made through our AI governance process, with the involvement of numerous experts – including third parties where appropriate. A final assessment of uplift is approved by senior-level decision-makers through our AI Risk process (see Section 2.1).

3.4 Outcomes & threat scenarios

This sub-section outlines the catastrophic outcomes that are in scope of our Framework. We include catastrophic outcomes in the following risk domains: Cybersecurity and Chemical & Biological risks. It is important to reiterate that these catastrophic outcomes do not reflect current capabilities of our models, but are included based on our threat modelling.

For each catastrophic outcome, we include a description of one or more threat scenarios. See Section 3.2 for more information on how we have developed our threat scenarios. We are not providing full details of the constituent steps and tasks within a threat scenario, or the enabling capabilities required to achieve it as we want to better understand how to balance transparency and security in this regard.

3.4.1 Cybersecurity

OUTCOME	THREAT SCENARIO (TS)	EXAMPLE ENABLING CAPABILITIES
Cyber 1: Automated end-to-end compromise of a best-practice-protected corporate-scale environment (ex. Fully patched, MFA-protected)	**TS.1.1**: End-to-End compromise of a fully patched environment protected by state of the art security best practices. Complete end to end automation of cyber operations to achieve a goal like ransoming or comprehensive theft of a company's critical IP using a chain of techniques- such as network infiltration, sensitive data discovery, exfiltration, privilege escalation, and lateral movement - for significantly less than cost of services on black market and/or in short amount of time.	Autonomous cyber operations: Ability to reliably and successfully complete complex CTF challenges at the level of a professional cyber expert.
Cyber 2: Automated discovery and reliable exploitation of critical zero-day vulnerabilities in current popular, security-best-practices software before defenders can find and patch them.	**TS.2.1**: Automated discovery and exploitation of cybervulnerabilities in current popular software using all best-practice exploit mitigations that would take human expert teams one month or more, on average, to find, for significantly less than cost of services on black market and/or in short amount of time.	Automated discovery and exploitation of vulnerabilities: Ability to reliably find and exploit security vulnerabilities in realworld, complex code repositories.
Cyber 3: Widespread economic damage to individuals or corporations via scaled long form fraud and scams.	**TS.3.1**: Automate end-to-end scam flows (e.g. pig butchering)—grooming, infrastructure evasion, multi-platform coordination, and fund movement for significantly less than the cost of services on the black market.	Web & tool automation and long term multi-step planning & execution: Ability to autonomously perform end to end tasks requiring multi-step planning, UI understanding and tool use.

3.4.2 Chemical & biological

OUTCOME	THREAT SCENARIO (TS)	EXAMPLE ENABLING CAPABILITIES
CB 1: Proliferation of known medium-impact biological and chemical weapons for low and moderate skill actors.	**TS.1.1**: Small cells of low or moderate skill actors acquire sufficient scientific and operational capabilities to succeed at complex chemical and biological workflows.	• Graduate level knowledge in biology, biochemistry, and chemistry • PhD level proficiency in the relevant sub-specialty for the threat in question • Summarization of scientific and technical information in a way that's accessible to a non-expert audience
	TS.1.2: Small cells of low and moderate skill actors can consistently execute relevant and complex wet-lab biological protocols.	• Recapitulation of complete wet-lab protocols for complex biological experiments. • Ingestion and interpretation of scientific data (e.g. images or datasets).
CB 2: Proliferation of high-impact biological weapons, with capabilities equivalent to known agents, for high-skilled actors.	**TS.2.1**: A well-resourced group of high-skill actors can procure and scale up production of a high-impact biological agent.	• PhD level proficiency in the relevant sub-specialty for the threat in question
	TS.2.2: A well-resourced group of high-skill actors can improve an existing medium-impact biological agent to match the capabilities of a high-impact agent.	• PhD level proficiency in the relevant sub-specialty for the threat in question • Use of biodesign tools
CB 3: Development of high-impact biological weapons with novel capabilities for high-skilled actors.	**TS.3.1**: A group with extensive resources can meaningfully accelerate the development of a novel biological agent with significant new capabilities.	• PhD level proficiency in the relevant sub-specialty for the threat in question • Use of biodesign tools

Section 04
Implementation

Our decision-making process for developing and releasing frontier AI is guided by our internal AI governance program, our risk thresholds, and the rigorous program of evaluation and mitigation that underpins them.

This section outlines our process for evaluation and mitigation and provides an overview of the corresponding measures we will implement in order to manage risks from our frontier AI models and enable their safe development and release.

4.1 Preparing a robust evaluation environment

AI model evaluation is a nascent science. Improving the robustness and reliability of evaluations is an area of focus for us, and this includes working to ensure that our testing environments produce results that accurately reflect how the model will perform once in production. This includes accounting for capabilities that might undermine reliability of results, such as deception. Ensuring a robust evaluation environment is therefore an essential step in reliably evaluating and risk assessing frontier AI.

4.2 Evaluation and mitigation

We conduct an initial set of evaluations on a first checkpoint to assess capabilities across the risk domains, with a particular focus on the enabling capabilities we have identified for our threat scenarios. These evaluations serve two key purposes. Firstly, they act as an 'absence validation test' – we check the results for the set of enabling capabilities we've identified for our threat scenarios, and assess whether the model exhibits sufficient performance to potentially enable execution of any of our threat scenarios. Secondly, they help us to assess whether capabilities are in line with expectations (see section 2.1) and therefore guide further evaluations and mitigations.

If our evaluations indicate that a frontier AI does *not* exhibit sufficient performance on these capabilities, we will continue training and observing how capabilities develop, using the reference class as the guide.[8]

[8] With current evaluations, it is not possible to define a fixed set of quantitative metrics that would indicate sufficient performance across enabling capabilities. We make this assessment through a process of expert deliberation and analysis of the evidence through our AI governance process.

We use the reference class and the evaluations that we conduct to identify where models require more in depth evaluation to assess their risk, which may include conducting uplift studies. This enables us to differentiate between frontier AI in the high and moderate categories.

If we identify that a frontier AI *does* exhibit sufficient performance on these capabilities, we will conduct further evaluations to establish whether the frontier AI would enable execution of the threat scenario. As explained in Section 3.2, successful execution of a threat scenario does not necessarily mean that the catastrophic outcome is realizable. If the outcome of that threat modelling shows that additional threat scenarios remain a barrier to realization of the catastrophic outcome, we will update our Framework with these additional threat scenarios, and the model can move to the High threshold in the new iteration of the Framework.

Our evaluations are designed to account for the deployment context of the model. This includes assessing whether risks will remain within defined thresholds once a model is deployed or released using the target release approach. For example, to help ensure that we are appropriately assessing the risk, we prepare the asset – the version of the model that we will test – in a way that seeks to account for the tools and scaffolding in the current ecosystem that a particular threat actor might seek to leverage to enhance the model's capabilities. We also account for enabling capabilities, such as automated AI R&D, that might increase the potential for enhancements to model capabilities.

We may take into account monetary costs as well as a threat actor's ability to overcome other barriers to misuse relevant to our threat scenarios such as access to compute, restricted materials, or lab facilities.[9] If the results of our evaluations indicate that a frontier AI has a "high" risk threshold by providing significant uplift towards realization of a threat scenario we will not release the frontier AI externally.

Models that are not being considered for external release will undergo evaluation to assess the robustness of the mitigations we have implemented, which might include adversarial prompting, jailbreak attempts, and red teaming, amongst other techniques. This evaluation also will take into account the narrower availability of those models and the security measures in place to prevent unauthorized access.

[9] We recognize that as costs for training and adaptation reduce, financial constraints may become less of a barrier to misuse of AI. We will account for changing economic models as necessary.

We typically repeat evaluations as a frontier AI nears or completes training. Evaluation results also guide the mitigations and controls we implement. The full mitigation strategy will be informed by the risk assessment, the frontier AI's particular capabilities, and the release plans. Examples of mitigation techniques we implement include:

- Fine-tuning
- Misuse filtering, response protocols
- Sanctions screening and geogating
- Staged release to prepare the external ecosystem

4.3. Benefits assessment

While the focus of this Framework is on our efforts to anticipate and mitigate catastrophic risks from frontier AI, it is important to emphasize that the reason to develop advanced AI systems in the first place is because of the tremendous potential for benefits to society from those technologies. Like quantifying risk, quantifying the benefits of AI is an imperfect science for several reasons. Firstly, both risks and benefits emerge gradually, and often on different time horizons, so the overall impact of a technology may shift over time. Secondly, many impacts are difficult to measure quantitatively. For example, access to advanced AI models has clear benefits for advancing scientific research in different fields, but quantifying the value of that research is extremely difficult, and other discoveries or variables can also influence the scale and impact of that research.

Even for tangible outcomes, where it might be possible to assign a dollar value in revenue generation, or percentage increase in productivity, there is often an element of subjective judgement about the extent to which these economic benefits are important to society.

While it is impossible to eliminate subjectivity, we believe that it is important to consider the benefits of the technology we develop. This helps us ensure that we are meeting our goal of delivering those benefits to our community. It also drives us to focus on approaches that adequately mitigate any significant risks that we identify without also eliminating the benefits we hoped to deliver in the first place.

That is, we believe that by considering both benefits and risks in making decisions about how to develop and deploy advanced AI, it is possible to deliver that technology to society in a way that preserves the benefits of that technology to society while also maintaining an appropriate level of risk.

5

Section 05

Future work

5.1 Updates to our framework

As outlined in the introduction, we expect to update our Frontier AI Framework to reflect developments in both the technology and our understanding of how to manage its risks and benefits. To do so, it is necessary to observe models in their deployed context and to monitor how the AI ecosystem is evolving. These observations feed into the work of assessing the adequacy of our mitigations for deployed models, and the efficacy of our Framework. We will update our Framework based on these observations.

We track the latest technical developments in frontier AI capabilities and evaluation, including through engagement with peer companies and the wider AI community of academics, policymakers, civil society organizations, and governments. We expect to update our Framework as our collective understanding of how to measure and mitigate potential catastrophic risk from frontier AI develops, including related to state actors. This might involve adding, removing, or updating catastrophic outcomes or threat scenarios, or changing the ways in which we prepare models to be evaluated. We may choose to reevaluate certain models in line with our revised Framework.

5.2 Research areas of focus

As discussed above, we recognize that more research should be done – both within Meta and in the broader ecosystem – around how to measure and manage risk effectively in the development of frontier AI models. To that end, we'll continue to work on: (1) improving the quality and reliability of evaluations; (2) developing additional, robust mitigation techniques; and (3) more advanced methods for performing post-release monitoring of open source AI models.

Appendix – terminology

We include these definitions to aid understanding when reading our Framework. However, we note that there is a lack of consensus as to how to define some of these terms and concepts. Societal norms around frontier AI risk will evolve over time, and definitions and terminology for some of these concepts may therefore also evolve as a result.

- ***Frontier AI*** in our Framework refers to highly capable general-purpose generative AI models and systems that we are developing for release or deployment that exceed the capabilities present in the most advanced models. Evaluations will often be done on a model that is embedded in a system.

- ***Catastrophic outcomes*** are outcomes that would have large scale, devastating, and potentially irreversible harmful impacts on humanity that could plausibly be realized as a direct result of access to frontier AI in the future.

- ***Threat modelling*** is a structured process of identifying how different threat actors could leverage frontier AI to produce specific – and in this instance catastrophic – outcomes. This process identifies the potential causal pathways for realizing the catastrophic outcome.

- ***Threat scenarios*** describe how different threat actors might achieve a catastrophic outcome. Threat scenarios may be described in terms of the tasks a threat actor would use a frontier AI model to complete, the particular capabilities they would exploit, or the tools they might use in conjunction to realize the catastrophic outcome.

- ***Enabling capabilities*** are a set of capabilities that are identified as essential to enabling the realization of a threat scenario.

- ***Uniquely enabling*** describes a model that is an essential controlling factor in a given outcome. A model is considered to meet the critical risk threshold if it is determined that a specified threat scenario would not occur without this particular model.

- ***Risk domain*** is used to describe the thematic grouping that a set of catastrophic outcomes belong to.

- ***Risk thresholds*** are the incremental levels of risk that a frontier AI model might pose towards realization of a catastrophic outcome.

- ***Residual risk*** describes the level of risk that a frontier AI model presents *after* mitigations have been implemented.

- ***Development*** refers to the process of training, fine-tuning, and evaluating frontier AI models before deployment or release.

- ***Release*** refers to the different ways in which we may choose to deploy, release, or give access to our models, for example:

  - ***Closed deployment****:* models that are deployed internally or in Meta products, but are not directly available to external partners.

  - ***Limited release****:* releasing externally to a limited set of trusted external partners.

  - ***Full release****:* releasing externally for open research and development as pre-trained and/or fine-tuned versions.

- ***Evaluation(s)*** refers to the assessments we do to understand capabilities and performance. We use this term to describe automated and human evaluations that assess capabilities, as well as evaluations to assess potential for misuse, such as red teaming and uplift studies.

- ***Uplift studies*** are experiments that assess the extent to which access to frontier AI increases a person or group's ability to complete a particular task or scenario in comparison to a control group that only has access to existing resources, such as textbooks, the internet, and existing AI models.