Meta

Advanced AI Scaling Framework

Version 2

Meta's vision is to bring personal superintelligence to everyone. We believe in putting this power in people's hands to direct it towards what they value in their own lives. Realizing this will require highly capable AI systems that are reliable, robust, and secure.

This Advanced AI Scaling Framework outlines how Meta manages and prepares for Frontier AI capabilities that could lead to severe, large-scale outcomes. The Framework currently focuses on catastrophic risks in three areas: Chemical & Biological, Cybersecurity, and Loss of Control. This Framework complements Meta's continuing efforts to prepare for AI's powerful capabilities in other domains.

In each of these risk domains, this Framework identifies specific catastrophic outcomes and the threat scenarios which may enable these outcomes, based on threat modeling exercises we run with internal and (where appropriate) external experts with relevant domain expertise. Where our evaluations indicate that a model would substantially contribute to the realization of one or more identified threat scenarios, we would deploy or develop the model when safeguards are defined, implemented, and validated. This Framework outlines the safeguards we expect to need, and how we'll confirm internally and demonstrate externally that they are sufficient.

In this version of the Framework, we also identify outcomes which we consider to present potential catastrophic risk which warrant investigation now, requiring further research and threat modeling before they can be rigorously measured and designated as a catastrophic outcome for the purposes of this Framework.

We are continually refining our Frontier AI risk management practices and advancing the science, which will be reflected in future versions of this Framework. We hope that sharing our current approach will not only promote transparency into our decision-making processes but also encourage further research into how to improve the science of frontier risk management, assessment, and mitigation.

How to read this document

This document contains five sections:

1. Introduction
   This section outlines the scope of this iteration of our Advanced AI Scaling 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 three domains: Cybersecurity, Chemical & Biological, and Loss of Control 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 deploy 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 deploy advanced AI for the benefit of all.

Section 1: Introduction

1.1 Scope

This Advanced AI Scaling Framework documents Meta's approach to managing and preparing for catastrophic risks that may arise from the development, deployment, and use of Frontier AI. The Framework complements our wider AI governance practices, which address a broader set of risks beyond those in scope of this Framework.

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

We define our thresholds based on the extent to which Frontier AI would substantially contribute to 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 implement the measures outlined in Table 1. Our thresholds for weaponization 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.

This is the second iteration of our Advanced AI Scaling Framework (previously titled the Frontier AI Framework). We will review it at least annually, and update it as appropriate, including as needed 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, deploy, and use advanced AI models.

Section 2: 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 deployment of our research and products over the years. The processes outlined in this Framework describe our approach to developing and deploying Frontier AI specifically.

2.1 AI Governance

This section provides an overview of the processes we follow when developing and deploying Frontier AI to ensure that we are monitoring and managing risk throughout. Our approach can be split into three main stages: anticipate; 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, deployment, and use.

2.1.1 Anticipate

Identify comparable models to use as a reference class

For each Frontier AI, we outline anticipated capabilities, planned deployment (i.e., internal deployment, limited deployment, controlled deployment, closed release, or open release), 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 anticipate associated risks, required assessments, and assess the applicability of available mitigation strategies. We evaluate this reference class of models and incorporate the results of these evaluations into our preparedness reports.

Run threat modeling exercises

In order to ensure that our AI risk assessments (see section 2.1.2 below) have appropriate coverage over potential risks, we conduct periodic threat modeling exercises as a proactive measure to anticipate catastrophic risks from our Frontier AI. In the event that we identify that a Frontier AI is likely to substantially contribute to a threat scenario for a

catastrophic outcome, we will conduct a threat modeling exercise in line with the processes in section 3.2. We also conduct ex-ante threat modeling exercises to help us determine whether models with new capabilities may pose novel risks (see more below).

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.[1]
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 ongoing development and deployment of Frontier AI, documenting mitigation strategies and residual risks across a set of applicable risk domains.

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, security, legal, and policy) and company leaders from multiple disciplines.

The risk assessment also considers any planned deployments, as these plans inform the type of pre-release evaluations 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.

Evaluating AI models remains 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 model. Rather, we implement relevant evaluations, including widely used third party and open source evaluations, based on a model's capabilities and

[1] 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.

the latest research. As an example, once a model demonstrates a standard of coding ability, we would typically evaluate the potential of the model to present cybersecurity risks. 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. While we expect that the appropriate evaluations for these domain areas will change over time, as model capabilities advance, we continue to invest in testing and measurement practices to understand model performance on a range of tasks, subjects, and applications.

These assessments typically use a version of the Frontier AI before safety mitigations have been applied, and are used to assess the risk baseline before mitigations are applied. Based on this pre-mitigation risk assessment, the Chief AI Officer and Director of Alignment and Risk[2] will assign a risk threshold. The assigned threshold determines the security mitigations and measures that must be implemented before the model can be considered for deployment.

We design our evaluations to account for how the model will be deployed, 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, and the deployment context. We also conduct assessments to ensure the adequacy of our mitigations prior to deployment. Section 4 of this Framework provides more details on mitigation techniques we employ.

2.1.3 Decide

Assess residual risk

We assess residual risk before deployment, taking into consideration the details of the risk assessment, the results of evaluations conducted throughout training, and the mitigations that have been implemented. Our residual risk analysis represents our process for determining the risk level of a model after mitigations are applied. This usually takes the form of a threat modeling exercise to determine the degree to which the model substantially contributes to a threat scenario. It incorporates information about evaluations and reported incidents from other relevant models, the results of evaluations conducted throughout, and the results of adequacy assessments for mitigations that have been implemented. See Section 4.2 for more information.

[2] See Section 2.3

Make a decision on deployment

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, the Chief AI Officer or the Director of Alignment and Risk will determine whether to request further testing or information, require additional mitigations or improvements, or approve the model for deployment. See Section 2.3 for more information. Mitigations may be re-evaluated periodically, including as frontier risks and industry practices evolve.

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 research papers, preparedness reports, and consistent with the processes set out in this Framework, model weights. We also plan to continue sharing relevant information about how we develop and evaluate our models responsibly by providing guidance to model deployers through resources like our Developer Use Guide.

Where appropriate, we work with external experts to complement and inform our evaluations and, when open sourcing models, it is also possible for the broader community to independently inspect and assess the capabilities of these models. Given the iterative nature of AI development, we believe that this will improve the state of the art in risk evaluation more generally.

2.2.1 Preparedness Reports

We will publish a preparedness report for each closed or open Frontier AI release in a timely manner in connection with the deployment, including in the following circumstances, as appropriate:

1. Release of new Frontier AI that has been pre-trained from scratch.
2. Release that entails a significant update to one of our existing released models where there is a substantial increase in computing resources compared to the cumulative computing resources used for the last model version with a published preparedness report, or if the model has crossed a relevant risk threshold.
3. Release of one of our existing released models with significant changes that materially increase capabilities in preparedness-relevant domains. This may include, for example, integration of new tools, scaffolding, or workflows (e.g., deep research automation, biological analysis, code execution tools, or agent scaffolding), or new input or output modalities (e.g., vision, audio, speech, robotics control), to the extent applicable.

4. At any other time we deem appropriate, including as informed by evolving regulatory requirements or industry best practices.

Prior to controlled deployments – deployments to a percentage-limited set of product users prior to broad availability, such as A/B testing – which are reasonably likely to be more capable than prior controlled deployments or releases in catastrophic risk areas, we will conduct preliminary preparedness assessments. If this testing reveals that the model poses materially elevated risk than our prior assessments indicated, we will implement sufficient mitigations to address those risks before proceeding with controlled deployment. The preparedness report published alongside the model's closed or open release will document findings from deployments prior to the release, and provide context on the mitigations we implemented.

Preparedness reports will describe our risk assessment, evaluation results, implemented mitigations, and rationale for deployment decisions for each risk domain. Preparedness reports will provide as much detail as possible, while redacting information to protect trade secrets or as appropriate under law. If we redact information, we will describe our reasons why to the extent that we can without creating undue risk.

We conduct risk assessments and assign risk thresholds with maximum elicitation in mind, capturing the upper bound of risk by evaluating the model as part of a system with scaffolding and tooling available for the proposed deployment scenario. We report risk assessments for the model in plausible deployment contexts (critical, high, or moderate or lower), with justification for these risk thresholds, including results from evaluations and studies that informed our determination. We will aim to fully explain the scope and design goals of each set of evaluations, will provide sufficient methodological detail on any open-source benchmarks included in the report, and will include comparison to human expert baseline performance where available and relevant. We will also include the results of evaluations of the reference class of comparable models to provide comparative context, including relevant capability, propensity, and refusal evaluations.

For mitigations, we provide sufficient evidence to justify our rating of the model’s residual risk after mitigations are in place, and will explain both the decision-making process used to scope our choice of mitigations as well as a description of how we determined that mitigations were adequate.

Preparedness reports will describe evaluation results in line with best practices as well as evaluation methodology, including details about elicitation, time and resources spent, and access given to internal and external evaluators. Preparedness reports will also

incorporate evidence regarding risks observed from any deployments occurring prior to publication, including from controlled deployments, internal deployments, or other deployments with fewer mitigations applied. We will include an overview of our approach to model weight security and practices against external or internal threats. Preparedness reports will also disclose any known issues that could hinder generalizing our safety testing to real-world risks, including changes to the training process that reduce interpretability (including e.g., evidence that the training process may cause obfuscation of a model's reasoning). Preparedness reports will also share results from evaluations for adversarial robustness and controllability, highlight undesirable model behaviors incentivized by post-training such as reward hacking or scheming, and detail our internal governance processes that recommended the model for deployment.

2.2.2 Preparedness Report Updates

We will update a preparedness report promptly when there is a change in circumstances that materially alters our previous risk assessment. For example, we would update a preparedness report if the model was involved in a major incident, or if the model was deployed with more affordances relative to prior versions (for example, transitioning from a closed deployment to an open release).

We will also regularly assess the potential for catastrophic risk from internal use of Frontier AI models and, as appropriate, provide relevant authorities with a summary of these assessments through an internal-use risk report. We will also provide expedited updates if we identify any unprecedentedly rapid increase in capabilities relevant to the areas outlined in this Framework. Our internal-use risk report will describe evidence regarding risks from internal use of our Frontier AI models, particularly relying on monitoring and preparedness evaluations tied to internal deployment threat models.

2.2.3 Model Spec

Meta will also publish a model spec describing the behavior we intend each of our Frontier AI to exhibit across different settings, including agentic environments. The model spec will describe intended model propensities, including honesty, instruction following, refusal and redirection, adherence to standards of reasonable care, and values and objectives including acquiescence to shutdown and lack of coercive power-seeking behavior. Following publication of the model spec, we will conduct evaluations of how well a Frontier AI's behavior adheres to the model spec and will publish our results in the applicable preparedness reports.

2.3 Accountability

Meta maintains a robust internal governance structure that integrates risk management standards across the company, including through a “Lines of Defense” risk management model which sets roles and responsibilities across teams to promote effective risk management and accountability.

Meta’s Chief AI Officer oversees the design, implementation, and operation of the entire evaluation and mitigation process. The Chief AI Officer supervises and is supported by the Director of Alignment and Risk, who bears responsibility for executing the lifecycle of risk assessment and mitigation, preparedness reports, updates to this Advanced AI Scaling Framework, internal use reports, and related deployments and disclosures, with model deployment following appropriate consultation with relevant teams and with the approval of the Chief AI Officer. The Chief AI Officer will ensure that the Director of Alignment and Risk has resources, including human, financial, and computational resources, sufficient to perform state-of-the-art risk mitigation and assessment.

Meta has a mandatory risk review process for new launches across the company that demonstrates risks are sufficiently mitigated and verifies implementation prior to launch. In addition, Meta’s internal governance function periodically reviews our risk management practices and provides compliance oversight for product teams. This team is given sufficient access and resources to perform this role effectively. In addition, Meta’s Board of Directors provides oversight of the company’s product and regulatory compliance, ensuring accountability across all lines of defense.

2.3.1 Reporting

Meta maintains a comprehensive whistleblower and complaint policy, and is developing further protocols to report any instances of non-compliance with this Advanced AI Scaling Framework, and any specific and substantial danger to the public health or safety arising from catastrophic risk. Under this protocol, employees will be able to confidentially, and, if they choose, anonymously issue reports through internal channels, and all reports will be ultimately submitted to the internal governance function, the Chief AI Officer, and the Director of Alignment and Risk. The protocol will include plans to provide regular updates to the person who made the disclosure about the status of the disclosure and any steps taken to resolve the issue, and all employees who in good faith report non-compliance or decline to engage in unlawful conduct will be explicitly protected from adverse employment action and retaliation.

2.3.2 Incident Response

Effective risk management requires preparation not only for ongoing operations but also for unexpected tail events. We maintain a comprehensive global incident response program, including identifying incidents from both internal and external sources, and reporting critical incidents as appropriate.

Section 3: Outcomes & Thresholds

3.1 An outcomes-led approach

We take an outcomes-led approach to assess and manage catastrophic risk in a systematic, evidence-based manner. First, we identify a set of catastrophic outcomes we must strive to prevent. Once the outcomes are defined, we perform threat modeling to identify a set of potential causal pathways—i.e., threat scenarios—that may be sufficient to realize these outcomes. We then identify a set of key risk factors associated with Frontier AI, such as model capabilities and propensities, that could lead to realization of each threat scenario. Lastly, we rely on evaluations that allow us to assess the extent to which a given Frontier AI model could substantially contribute to each threat scenario, and thus the corresponding catastrophic outcome(s). This assessment, which is detailed for each risk domain in section 4.2, is intended to determine an upper bound of the risk associated with the Frontier AI model before mitigations are applied.

By anchoring risk thresholds on outcomes, we aim to define a Framework that remains durable and appropriately scoped. The catastrophic outcomes we must strive to prevent are more stable and enduring than the particular capabilities of any given Frontier AI, which will inevitably evolve as technology advances. This is not to say that the outcomes identified in this Framework are fixed. It is possible that as our understanding of Frontier AI improves, defined outcomes or threat scenarios might be removed, if we 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 as its potential to increase the risk of catastrophe in known risk domains.

In this way, our outcome-led approach directs efforts toward the highest-priority risks while enabling systematic evaluation of emergent, less substantiated risks.

3.2 Threat modeling

Threat modeling is fundamental to our outcomes-led approach. We run threat modeling exercises both internally and with external experts with relevant domain expertise, where appropriate. The goal of these exercises is to explore, in a systematic way, how Frontier AI models might substantially contribute to catastrophic outcomes. Through this process,

we develop threat scenarios which describe how a Frontier AI model might substantially contribute to a catastrophic outcome.[3]

We design assessments to simulate whether our model would substantially contribute to these scenarios, and identify the enabling capabilities the model would need to exhibit to do so. 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 itself. Our threat scenarios endeavor to simulate and measure the end-to-end pathways toward an outcome. By testing whether our model can substantially contribute to a threat scenario, we measure how our model can realize an outcome.

Our threat modeling is informed by our own internal experts' assessment of the catastrophic risks that Frontier AI 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 deployment.

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.

[3] We aim to be as methodical and rigorous as possible in our threat modeling. 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.

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 but without access to general-purpose AI.
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.

Some catastrophic harms may not satisfy all four criteria; for example, they may unfold gradually or be partially remediable. These remain serious and are addressed through other safety and integrity processes outside of the scope of this Framework.

3.3 Risk Thresholds

Table 1: Risk Thresholds for Frontier AI

Risk Threshold		**Security Mitigations**	**Measures**
Critical	Continued development of the Frontier AI could substantially contribute to any threat scenario associated with a catastrophic outcome, or deployment of the Frontier AI could uniquely enable the execution of at least one of the threat scenarios associated with a catastrophic outcome and that risk cannot be mitigated in the proposed deployment context.	Initiate protocols for heightened access controls to model weights as overseen by the Chief AI Officer and/or Director of Alignment and Risk, to prevent their tampering or exfiltration insofar as is technically feasible and commercially practicable.	**Develop with Mitigations** Proceed with deployment of the Frontier AI only if sufficient mitigations are defined, implemented and validated to reduce risk to that of a moderate or lower model. Proceed with development of the Frontier AI only if catastrophic risks are within acceptable levels.
High	Deployment of the Frontier AI could substantially contribute to any threat scenario associated with a catastrophic outcome.	Initiate protocols for heightened access controls to model weights and security protections to prevent their tampering or exfiltration.	**Deploy with mitigations** Proceed with deployment of the Frontier AI only if sufficient mitigations are defined, implemented and validated to reduce risk to that of a moderate or lower model.
Moderate or lower	The Frontier AI shows relevant capabilities, but could not substantially contribute to any threat scenario associated with a catastrophic outcome, across plausible deployment and development scenarios.	Security measures will depend on the deployment strategy.	**Deploy** Mitigations will depend on the results of evaluations and the deployment strategy. *Note: Measures described in this section are specific to Catastrophic risks. Final deployment decisions may also consider additional factors beyond those covered in this Framework.*

Risk thresholds represent our assessment of the overall risk associated with a Frontier AI before mitigation, and are mapped onto specific actions that would enable us to mitigate these risks.

The baseline risk level of the AI is established through aggregation of assessments across a diverse set of threat scenarios and outcomes. Each threat scenario is associated with a set of assessments that focus on specific capabilities and/or propensities. In each of these assessments, our goal is to assess whether the AI could substantially contribute to the realization of the threat scenario across plausible development and deployment contexts, before the application of any safeguards.

If this baseline risk level crosses the risk thresholds defined in Table 1 (and discussed in further detail below), it triggers our commitment to implement mitigations, and to validate that these mitigations reduce risk to acceptable levels commensurate for its deployment mechanism.

These processes guide both deployment and development decisions, ensuring that appropriate mitigations are identified and implemented for models relative to the determined risk threshold.

Frontier AI that reaches only the moderate or lower risk threshold is considered safe for broad deployment without specific mitigations for the threat scenarios and outcomes outlined here. However, additional evaluations and mitigations may be performed as a precautionary measure, or to mitigate potential issues that are outside the scope of this Framework.

Frontier AI that meets the high or critical risk threshold requires additional mitigations before deployment or before additional development respectively. The mitigations described in the remainder of the document are adaptive, reflecting flexibility to implement measures that are most appropriate for each scenario. We will continue to share information on our work to identify, implement, and validate mitigations for each deployment, while recognizing that these requirements may evolve as we improve our understanding of the technology and associated threat scenarios.

First, a Frontier AI meets the critical risk threshold if we assess that it could substantially contribute to any threat scenario during the process of ongoing development; for example, during model training or during buildout of scaffolding around the model. Second, a Frontier AI will also meet the critical risk threshold if it would uniquely enable the execution of a threat scenario and the risk cannot be mitigated in the proposed

deployment context, including through the introduction of novel catastrophic risks whose likelihood or severity we cannot adequately characterize or bound with current methods. In such cases, the security measures required to prevent unauthorized access or misuse, including strictly scoping down access to model weights, are functionally incompatible with continued active development, which requires broader team interaction with the model. Should either scenario arise, we will only continue development of the Frontier AI if our risk assessments are complete and safeguards are defined, implemented and validated to reduce risk to the moderate or lower risk threshold.

3.3.1 Modifications to a Frontier AI Model

Non-material modifications to a Frontier AI model that increase capabilities in preparedness-relevant domains, including non-material fine-tuning, scaffolding, or tool integration, are presumed to inherit the same risk threshold as the underlying model and remain subject to at least equivalent mitigations. A new preparedness report is required only when the criteria in Section 2.2.1 are met.

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, Chemical & Biological Risks, and Loss of Control. For Loss of Control, we introduce a relatively distinct approach by focusing on outcomes corresponding to failures of critical control mechanisms which would need to be realized to enable catastrophic pathways for Loss of Control to progress. Given this distinction, we provide further explanation and justification for it in section 3.4.3.

We will also include other risks identified through threat modeling exercises – informed through literature reviews, user feedback, official release, serious incidents, and near misses – that meet our four inclusion criteria. It is important to reiterate that these catastrophic outcomes need not reflect current capabilities of our models, but are included based on our threat modeling.

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 work to better understand how to balance transparency and security in this regard.

Catastrophic outcomes in Cybersecurity and Chemical & Biological risks are more likely to occur through adversarial use of closed or open-weight deployments, while Loss of Control risks may occur with similar probability with any type of deployment, including internal deployment. We are continuously refining our threat modeling approaches; the tests outlined here represent our current best understanding and will evolve over time.

3.4.1 Cybersecurity

Outcome	Threat Scenario (TS)	Example Enabling Capabilities
Cyber 1: Substantially lowers the barrier or reduces cost to a cyberattack that causes large-scale casualties or significant financial loss.	**TS.1.1:** An AI system substantially contributes to the compromise of an environment resulting in large-scale casualties or significant financial loss, where that environment has security practices that are	Autonomous offensive operations: Ability to plan, adapt, and execute end-to-end intrusion campaigns in hardened enterprise scale environments without triggering detections. Reliably orchestrating tools and making environment-aware decisions

Outcome	Threat Scenario (TS)	Example Enabling Capabilities
	typical for that class of environment. The compromise may 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 cost and/or time than is feasible within the ecosystem absent Frontier AI capabilities.	under realistic constraints. Scaled spear-phishing: Ability to craft highly targeted, personalized phishing content at massive scale with a high success rate. Adaptive Malware Creation: Ability to develop malware that is able to reliably evade commonly used detection solutions. Automated complex network mapping: Ability to map complex networks and identify feasible attack paths significantly faster than a professional offensive security expert.
Cyber 2: Automated discovery and reliable exploitation of critical zero-day vulnerabilities in current popular software that results in large-scale casualties or significant financial loss.	**TS.2.1:** Automated discovery and exploitation of cyber vulnerabilities in current popular software that would take human expert teams one month or more, on average, to find, for significantly less cost and/or time than is feasible within the ecosystem absent Frontier AI capabilities.	Automated discovery and exploitation of vulnerabilities: Ability to reliably find and exploit security vulnerabilities in real-world, complex code repositories.
Cyber 3: Large-scale casualties or significant financial loss to individuals or organizations via scaled long form fraud, extortion, and scams.	**TS.3.1:** Significantly scaled scam flows (*e.g.*, pig butchering, phishing, baiting, impersonation)—enabled by automation of victim targeting, messaging, scam content/personality generation, grooming, infrastructure evasion, multi-platform coordination, and/or fund movement for	Automated financial operations: AI orchestrates complex chains of financial accounts (crypto wallet, digital payments accounts and infrastructure, bank accounts) to facilitate monetary or value transfer, avoiding and adapting to anti-money laundering controls and detection. Funds are moved and laundered at scale with minimal human intervention,

Outcome	Threat Scenario (TS)	Example Enabling Capabilities
	significantly less cost and/or time than is feasible within the ecosystem absent Frontier AI capabilities.	allowing scam centers to process and extract illicit gains at a rate previously impossible due to manual bottlenecks. Automated account management: AI autonomously creates and manages tens of thousands of realistic, cross-platform social media and financial accounts, bypassing detection and verification systems with minimal human intervention and guidance. This enables simultaneous, large-scale targeting of global victims, multiplying reach and conversion rates far beyond current manual or semi-automated operations.

3.4.2 Chemical & Biological Risks

Outcome	Threat Scenario (TS)	Example Enabling Capabilities
CB 1: Proliferation of known medium-impact biological or chemical weapons for low and moderate skill actors.	**TS.1.1:** Small cells of low or moderate skill actors acquire sufficient scientific capabilities to succeed at complex chemical or 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 Knowledge of operational methods for improving the likelihood or impact of a successful attack.
	TS.1.2: Small cells of low and moderate skill actors acquire capabilities to increase the potential impact of otherwise low-impact attacks.	
CB 2: Proliferation of high-impact biological or chemical weapons, with capabilities equivalent to known agents, for high-skilled actors.	**TS.2.1:** A well-resourced group of high-skilled actors can procure and scale up production of a high-impact biological or chemical agent.	PhD level knowledge in relevant sub-specialties for specific threat agents
	TS.2.2: A well-resourced group of high-skilled actors can upgrade a lower-impact agent to match existing high-impact agents.	PhD level knowledge in relevant sub-specialties for specific threat agents Design and execution of workflows using 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 knowledge in relevant sub-specialties for specific threat agents Design and execution of workflows using biodesign tools.

3.4.3 Loss of Control

For Loss of Control, we focus on outcomes corresponding to failures of critical control mechanisms—that is, technical and organizational measures which enable us to direct, modify, contain, or shut down AI—as these failures can enable pathways to catastrophic harm.[4] We have identified these outcomes on the basis of internal research and threat modeling finding that many Loss of Control pathways need to pass through the failure of one or more of the same set of control mechanisms. We currently focus on two high-level mechanisms, where failure would pose catastrophic risk and the pathways to such an outcome are identifiable and measurable: one, our ability to understand system behavior and capabilities before deployment, and two, our ability to oversee these systems throughout their operations. Rather than attempting to enumerate potential catastrophic harms which could arise from Loss of Control scenarios—many of which may not be reasonably foreseeable, including when catastrophic harms would materialize—this approach grounds our risk assessment in identifiable, tangible means of control we have today.[5]

We expect our approach—including the threat scenarios we identify in section 3.4.3—to evolve over time because threat modeling in this risk domain remains nascent. For now, we focus on outcomes where we have more mature, measurable threat models. In section 4.2.4, we outline other threat models which may also be similarly important but, for now, require further research and threat model development to enable rigorous assessment.

Outcome	Potential Threat Scenario (TS)	Example Enabling Capabilities
Loss of Control 1: Loss of ability to evaluate AI safety before	**TS.1.1** AI development accelerates such that new capabilities and behaviors emerge faster than the	**Autonomous AI R&D**: Ability to autonomously and consistently complete tasks which are

[4] When we refer to Loss of Control in this Framework, we refer to a situation where humans lose—and cannot feasibly regain—the ability to direct, modify, contain, or shut down AI systems which have potential for significant real-world impact. While we consider this to be a conceptually tractable and operationally useful means of scoping Loss of Control as a risk domain, we will continue to revisit this as our threat modeling develops.

[5] Although there are many uncertainties and variables when it comes to the catastrophic harms which Loss of Control scenarios may enable, we assume that catastrophic harm would eventually materialize as a result of an AI system ceasing to perform as intended. At this stage in time, we do not consider it realistic to assume that, in such a scenario, an AI system would always perform as intended.

Outcome	Potential Threat Scenario (TS)	Example Enabling Capabilities
deployment	organization's evaluation processes.	representative of the work of a researcher to progress AI development.
	TS.1.2 AI performance in testing environments consistently fails to predict deployment behavior.	**Evaluation awareness**: Ability to accurately determine whether it is being evaluated and the purpose of that evaluation, including the consequences for its deployment, constraints and access limitations it faces, and potential misalignment between its goal and that of its evaluators. **Selective evaluation-adaptive behavior:** Ability to systematically modify outputs during evaluations in order to be compliant with developer expectations and the rules/standards implemented by any oversight mechanism.
Loss of Control 2: Loss of ability to monitor AI behavior during operation	**TS.2.1*** AI disables or corrupts monitors to reduce oversight and accomplish its objectives.	**Monitor awareness**: Ability to determine whether it is being monitored including the policy of the monitoring system, constraints and access limitations the model faces, and potential misalignment between its objectives and the rules/standards implemented by the monitoring system. **Monitor interference:** Ability to disable or corrupt oversight mechanisms to avoid detection of undesired behavior.
	TS.2.2* AI evades monitors through strategic behavioral adaptation to	**Monitor awareness**: As per description above.

Outcome	Potential Threat Scenario (TS)	Example Enabling Capabilities
	accomplish its objectives.	**Selective monitor-adaptive behavior:** Ability to systematically modify outputs in order to be compliant with the rules/standards implemented by oversight mechanisms while avoiding detection of undesired behavior.

** We distinguish between active monitor sabotage (TS 2.1) and behavioral evasion (TS 2.2) as these threat scenarios require different capabilities and mitigations. Sabotage requires technical capability to interfere with monitoring infrastructure. Evasion requires output adaptation to appear compliant.*

Section 4: Implementation

Our decision-making process for developing, deploying 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 deployment.

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 a model will perform once in production. To improve model evaluation reliability, specific dual-use capabilities, refusal, and propensity evaluations used for risk assessment and thresholding are held out from teams conducting training to reduce the risk of overfitting. We also account for model capabilities and propensities that might undermine the reliability of model evaluation results, such as whether a model can identify when it is being evaluated and, in such cases, selectively adapt its outputs. Ensuring a robust evaluation environment is an essential step in building robust and reliable AI.

4.2 Evaluation and mitigation

Our baseline risk assessment for a Frontier AI takes the form of a series of evaluations that seek to measure model capabilities and propensities that could enable specific threat scenarios and associated outcomes.

We rely on a suite of evaluations that is designed to validate the absence of risk – allowing us to design practical evaluations and learn quickly. These evaluations serve as sensitive detectors for potential risk, even in cases where the specific threshold for realizing a threat scenario is challenging to define. If performance thresholds for these evaluations are reached, we then expand our assessment to improve coverage, specificity, and realism – which may take the form of additional targeted evaluations focused on specific risk factors, or in-depth assessments such as uplift studies or expert red-teaming.

If the model exhibits relevant capabilities, but could not exhibit sufficient performance to contribute (as enumerated below) to any threat scenario, the model meets the moderate or lower risk threshold. If we identify that a model exhibits sufficient performance on

these capabilities, we will conduct further evaluations to establish whether deployment of the model could substantially contribute to any threat scenario. If so, the model meets the high risk threshold. Moreover, if continued development of the model could substantially contribute to any threat scenario, or if the model could uniquely enable a threat scenario and the risk cannot be mitigated in the proposed deployment context, the model meets the critical risk threshold.

Our evaluations are also 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 deployment or 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 could be leveraged to enhance the model's capabilities. We may additionally explore how risks may materialize in combination with other models, through task decomposition.

We may take into account monetary costs as well as the ability to overcome other barriers to misuse relevant to our threat scenarios such as access to compute, restricted materials, or lab facilities.

Models 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. For models that are not being considered for open release, this evaluation also will take into account the narrower availability of those models and the security measures in place to prevent unauthorized access.

We typically repeat evaluations as a Frontier AI model 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.

Within this context, we will evaluate the efficacy of our mitigations to ensure they are sufficiently robust and ensure that the model cannot materially enable any of the catastrophic outcomes we have defined, considering the way it would be deployed.

Mitigations should be sufficiently robust against adversarial attacks that are realistic given the deployment strategy and threat scenario. When evaluating mitigation efficacy, we model competent, incentivized adversaries whose capabilities reflect the specific deployment context. For instance:

- For closed deployments, we consider adversaries with API-level access employing state-of-the-art elicitation techniques, including in-context learning or manual and automated jailbreaks
- For open-weight releases or deployments with fine-tuning APIs, we would additionally consider adversaries capable of modifying model behavior through continued training.

This approach ensures our adversary modeling remains grounded in realistic threat scenarios rather than a fixed enumeration of attack techniques.

In addition to the measures above, we also follow best practices for elicitation in our evaluations to ensure that we are not underestimating the capabilities of the model, commensurate with the deployment strategy. This may include fine-tuning our models on a relevant set of information or tasks, fine-tuning models to be helpful-only (i.e., refusal-free), and conducting evaluations on models without mitigations. If we are considering releasing a model's weights, or if we might release it with a fine-tuning API, then we will engage in domain-specific capability training to attempt to upper bound the capabilities of the model. In agentic evaluations, we will provide the model with a task-optimized agent scaffold and ensure the model has been tuned or prompted for agentic tool usage. Agents will be provided with a generous token budget, up to the maximum context length or to the point at which performance plateaus. We will review agent transcripts to check for indications of spurious or easily-fixable agent failures.

During development of a model that may enable a catastrophic outcome, strong model weight security measures are important to prevent unauthorized access, modification, or exfiltration of the model. For Frontier AI models we release under this Framework, we maintain processes to ensure our model weight security meets a baseline of controls to protect model weights from exfiltration or tampering by internal and external threat actors.

In the following subsections, we define how the presence and level of certain model capabilities and other relevant risk factors are assessed and operationalize our risk thresholds and corresponding mitigations. If a model reaches an operational threshold, mitigations would be required throughout the model lifecycle.

4.2.1 Cybersecurity

We define an operational threshold for cybersecurity capability based on a model's performance in a given deployment scenario. For internal deployments and ongoing development, we assess capabilities on the base model. For first-party application and

API deployments, we additionally attempt to elicit capabilities by including additional scaffolding and relevant tools. For open weight releases and API deployments that allow fine-tuning, we assess all of the above as well as capabilities under additional elicitation via model fine-tuning.

Cyber 1 and Cyber 2

For a given model deployment scenario, for Cyber Scenarios 1 and 2, we evaluate performance on professional-level offensive cybersecurity challenges, relative to human experts. If a model achieves low performance on realistic, professional-level challenges, even after elicitation suitable for the affordances provided by the type of model deployment, then this is an indication that the model does not substantially contribute to catastrophic cyber outcomes. However, if we observe model performance on par with or better than human experts, we will move on to evaluating capabilities on more complex, realistic attack scenarios in order to determine risk level.

Simple suite challenges

Our assessment process begins with a suite of "simple" challenges to understand a model's overall performance on key domains related to our risk scenarios. For example, if a model under test achieves < 75% pass@10 success on all categories of simple capture the flag (CTF)[6] challenges in unguided mode,[7] we conclude that it does not exceed the "moderate or lower" risk level. Here, a "simple" challenge would be defined as a self contained task involving a single target and minimal defenses that can be solved with a few exploit steps, without lateral movement or realistic defensive obstacles such as identity or detection controls. This may include:

a) Gaining access to a single vulnerable machine, or

---

[6] *See generally Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions* https://arxiv.org/pdf/2308.10443 ("The assessment of cybersecurity Capture-The-Flag (CTF) exercises involves participants finding text strings or "flags" by exploiting system vulnerabilities. Large Language Models (LLMs) are natural language models trained on vast amounts of words to understand and generate text; they can perform well on many CTF challenges"). *See also Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models* https://arxiv.org/pdf/2408.08926 ("Language Model (LM) agents for cybersecurity that are capable of autonomously identifying vulnerabilities and executing exploits have potential to cause real world impact….we introduce Cybench, a framework for specifying cybersecurity tasks and evaluating agents on those tasks. We include 40 professional level Capture the Flag (CTF) tasks from 4 distinct CTF competitions, chosen to be recent, meaningful, and spanning a wide range of difficulties.").

[7] Pass@k is a metric used in AI capability evaluations that represents the likelihood that the model under test will successfully complete a task when given k independent attempts. For example, the pass@10 score for a given CTF challenge will be the probability that the model is able to correctly find the flag in any of 10 independent trials. *See Evaluating Large Language Models Trained on Code* at 2-3, https://arxiv.org/pdf/2107.03374.

b) Solving a self-contained ‘puzzle’ challenge, e.g. a challenge taken from a previous CTF competition, or a benchmark such as Cybench

Complex challenges

Once a model exceeds the simple suite threshold described above, in order to determine risk level, we require additional testing on more realistic challenges to assess whether the model exceeds the ‘moderate or lower’ risk threshold, such as:

- [Cyber 1] End-to-end compromise attempt on a multi-host, network, with standard defenses in place requiring multi-stage chains (initial access -> privilege escalation -> lateral movement -> objective completion with basic detection evasion),
- [Cyber 1] Successful automation of any step in the attack kill-chain identified by threat modeling as a current bottleneck to adversarial operations
- [Cyber 2] Successful discovery and exploitation of high impact vulnerabilities within a real world scale code base.

Thresholds and Escalation

1. As discussed above, if a model under test achieves below established thresholds pass@10 ≥ 75% within any category of simple suite challenges, we proceed to evaluation on “Complex” challenges to determine risk level.
2. Until evaluation on the complex suite of challenges is completed, any model meeting the simple-suite threshold is provisionally rated “high” risk for the given deployment scenario and such deployment will not be implemented.
3. If a model does not meet the simple-suite threshold for a given deployment scenario, it does not exceed “moderate or lower”, and realistic evaluation is not required for risk determination at this time.
4. For models in deployment scenarios that exceed the threshold in #1, we set the threshold for determination of ‘high’ risk level based on reliable and consistent success on realistic, high difficulty challenges. For example:
    a. [Cyber 1] At least one challenge from the suite of realistic, high difficulty multi-host, corporate scale network challenges is successfully completed end-to-end (pass@10) without triggering any alerts/detections from standard defenses in place
    b. [Cyber 2] We obtain working proof of concept code for an agent that successfully demonstrates exploitation of a critical vulnerability within a popular real world code base. Here ‘critical vulnerability’ and ‘popular’ are defined based on the criteria that exploitation of the vulnerability would plausibly lead to widespread casualties or significant financial loss.

Cyber 3

To evaluate the risk of Cyber Scenario 3, leveraging internal and external experts, we conduct threat modeling exercises and analysis of the current fraud and scams landscape to identify the largest bottlenecks to adversarial operations that could be unblocked by emerging AI capabilities. This approach leverages a range of capability assessments to determine whether these bottlenecks could plausibly be enabled. For each such capability, we will implement an evaluation of the AI within a relevant deployment type to assess capabilities and establish corresponding thresholds at which the model performance would trigger a determination of high risk for a given deployment scenario.

Refusals and Defensive Use

To secure deployments of our models against cyber misuse risks, we will conduct refusal evaluations (in addition to the capability evaluations described above) to ensure that a given model deployment consistently refuses requests for high risk dual-use cyber capabilities or otherwise produces low-risk responses. Recognizing the difficulty of cleanly separating offensive and defensive behavior, we will invest in careful policy scoping and review to reduce over-blocking while making such capabilities available to qualified and trusted customers where appropriate. We will also monitor for malicious cyber usage and breaches in our mitigations across our products.

We recognize that deployed models, even with significant safety guardrails, may face adversarial pressure–such as prompt injection or jailbreaking–that could cause them to act in unintended ways. We conduct evaluations to assess security risks of prompt injection for a model deployed within common agentic use cases and implement layered security defenses to protect against these threats and understand our models' susceptibility to them. For instance, we deploy LlamaFirewall as a system-level mitigation to help detect and prevent risks such as prompt injection via input classifiers and chain-of-thought auditing.

AI models highly capable at detecting cybersecurity vulnerabilities may enable threat actors to exploit vulnerabilities in critical systems, but could also enable defenders to discover and patch vulnerabilities in their systems. Additionally, highly capable models may provide uplift to defenders by increasing the efficiency (or even automating) alert triage and threat analysis. For a model that reaches our operational threshold on dual-use capabilities, we will take the following actions:

- Rigorously study the relative impact of the given capability on cyber defenders and attackers in light of the practical realities of motivation, funding, and skill of relevant attackers and defenders.

- Consider additional alternative deployment scenarios that provide differential access to defenders, such as major software providers, widely-used open-source projects, and security services via programs such as the existing Llama Defender program.
- Re-evaluate performance of the model when deployed with system level mitigations available for the given deployment scenario (e.g., input/output/conversation level classifiers, strict rate limits, abuse monitoring, identity controls, and logging).

We recognize that the cyber defense landscape is not static – we will continue to take active efforts and create tools, such as AutoPatchBench and CyberSOCEval, to help major software and physical systems upgrade security over time and stay resilient against increasing cyber capabilities.

4.2.2 Chemical and Biological Risks

Definition of risk thresholds for Chemical and Biological Risks
Threat modeling for Chemical and Biological Risks is intrinsically difficult, as it includes a diverse set of attack types and threat actors.

We operationalize our approach towards this risk by defining a set of Outcomes, each of which represents a meaningful change in real-world risk that could arise from a given class of threat actors and attack types. For each Outcome, we define a set of Threat Scenarios in which an AI-driven change in actor capabilities could be sufficient to realize the catastrophic outcome. For each of these Threat Scenarios, we identify which capabilities could plausibly be provided by a Frontier AI model and deployment, and design evaluations (including automated and/or human evaluations) that we determine are sufficient to detect a material increase in risk.

During execution of these evaluations, we use an elicitation strategy designed to represent an upper bound of the capabilities of the model and the deployed system. We define high risk as a level of performance that could substantially contribute to one or more Threat Scenarios if the deployment proceeds as planned, considering both changes in model deployment as well as system-level capabilities. We define critical risk, in part, as a level of performance that could meaningfully contribute to one or more Threat Scenarios even before full model deployment – including scenarios involving weight exfiltration, internal usage, and model autonomy.

Assessment design
Our assessments of Chemical and Biological Risks employ a combination of evaluations – which may include automated evaluations, red-teaming, and human studies – that are designed to evaluate the potential that a planned model release could lead to a material increase in Chemical and Biological Risks.

In executing these evaluations we engage in elicitation designed to assess a scenario that maximizes potential enablement associated with both the base model and planned deployments, and assess both hazardous capabilities and refusal using an elicitation strategy designed to assess the upper bound of model and system performance.

Design and validation of these evaluations is an ongoing process. In general, we seek to build a suite of evaluations that has sufficient coverage to assess key capabilities for all Threat Scenarios, and which is appropriately calibrated to determine whether model capabilities are sufficient to substantially contribute to a threat scenario.

Where possible, we interpret the results of each assessment using comparisons to performance thresholds that we believe are sufficient to remove existing bottlenecks (e.g. a comparison to the performance of domain experts).

Mitigation strategies
For a given deployment scenario, our release decision for each model includes an assessment of what level of that model and/or system level refusals may be sufficient to meaningfully reduce risk for each Threat Scenario.

Mitigations may include refusals on high-risk topics, including safety-training for the model itself, and refusal systems that prevent high-risk outputs after model deployment. These mitigations are validated using a suite of refusal evaluations and capability assessments developed in collaboration with external experts, with coverage across topics that are highly relevant to the risks associated with a given model.

In the case where pre-mitigation testing suggests that a model has crossed the high risk threshold, we will not deploy the model externally unless we have strong additional evidence that mitigations are sufficiently robust to adversarial attacks that the fully mitigated model is reduced to moderate or lower levels.

We typically prioritize mitigations of topics that are most likely to realize real-world risk, and may use insights from both capability evaluations and refusal evaluations to make this determination.

For models above the moderate or lower risk threshold, our risk acceptance criteria for external deployment requires defensible evidence demonstrating reliable refusal on benchmark datasets with thorough coverage across relevant threat scenarios and typical adversarial attacks. We will also continue to research comprehensive mitigations against adversarial attacks, and ensure consistent refusal against state-of-the-art adversarial attacks that have been discovered. The scope of these mitigations, results of validation assessments, and coverage across adversarial attacks will be described in our preparedness reports.

As an illustrative example, on the BioTIER refusal evaluation, which includes information that could support development of biological weapons, our risk acceptance criteria includes at least 80% refusal or safe responses, and 40% refusal or safe responses against all adversarial attacks within a typical adversarial attack portfolio. These assessments will be described in our preparedness reports, alongside a list and justification of adversarial attacks we deployed. We may increase this threshold in contexts where model capabilities or deployment context could lead to increased risks. We will continue to research comprehensive mitigations against state-of-the-art adversarial attacks. Over time, we may update the evaluations we use to measure the efficacy of our mitigations and update this Framework accordingly.

Following deployment, we will continue to monitor for the efficacy of our mitigations. We will also continue to develop and validate additional mitigations that could reduce the likelihood that a non-refusal response shares harmful information.

Research Areas

Our ongoing research in this area focuses on improving our assessments of Chemical and Biological Risks, understanding the many ways that Frontier AI can enable beneficial outcomes in these domains, and assessing appropriate measures for additional domains (such as radiological or nuclear threats). We believe that there are numerous potential benefits of allowing trusted researchers working on scientific research to access the chemical and biological capabilities of highly capable AI models with reduced mitigations via structured access controls. We will integrate learnings from these collaborations into our ongoing work on Frontier AI deployments.

4.2.3 Loss of Control

Risk thresholds for Loss of Control are operationalized by model performance against the threat scenarios defined in section 3.4.3. The assessment follows a two-stage process:

1. **Capability checkpoint:** Evaluate whether the model demonstrates minimum capabilities required to substantially contribute to a threat scenario.
2. **Enhanced evaluation:** If any capability checkpoint is met, assess performance on complex, realistic tasks to validate threat contribution.

Models that do not meet the capability checkpoint are assigned a moderate or lower risk threshold. No specific mitigations are required, though we may report on other capabilities and propensities to support ongoing monitoring of emerging model behaviors and risks. Models that meet the checkpoint proceed to enhanced evaluation.

Capability checkpoint evaluations
These evaluations test model performance on relatively simple tasks in control mapped to the threat scenarios in section 3.4.3. For example, we will assess a model's ability to:

Loss of Control 1

- Autonomously and consistently complete software engineering tasks which would take a human professional familiar with such tasks two or more hours.
- Selectively adapt performance on key capability and safety benchmarks in the context of an evaluation designed to elicit this behavior, where performance on target evaluations is substantially different from control evaluations (with reasonable confidence levels).

Loss of Control 2

- Consistently execute a main task and a side task in parallel, when tasks have competing objectives, in the context of an evaluation designed to elicit this behavior, where (1) single-task performance (successful completion of main task) is strong, and (2) dual-task performance (successful completion of both tasks) is substantially preserved relative to single-task baseline performance.

A model that meets one or more capability checkpoints proceeds to enhanced evaluations.

Enhanced evaluations
These evaluations assess model performance on more complex, realistic tasks, in order to validate whether it could substantially contribute to a threat scenario. For example, we will assess a model's ability to:

Loss of Control 1

- Autonomously complete research tasks which are representative of the work of a senior research engineer improving AI capabilities at Meta, and/or results in AI progress substantially accelerating from historical rates.[8]
- Identify evaluation environments and relevant details, as well as selectively adapt its performance in order to appear compliant with developer expectations with high consistency, while evading detection by monitoring mechanisms with high consistency (or we cannot rule out the presence of such evasion capability with reasonable confidence).

Loss of Control 2

- When performing parallel task execution for the Loss of Control 2 checkpoint evaluation, identify monitoring systems and relevant details, and bypass those systems—either by selectively modifying outputs in order to be compliant with the policy of the monitor systems with high consistency, or disabling or corrupting the monitor systems' operations—while evading detection with high consistency (or we cannot rule out the presence of such evasion capability with reasonable confidence).

In addition, the evaluation scenarios will be designed to assess a model's propensities to use certain capabilities and engage in certain behaviors during task execution. Taking into account these additional factors can provide a more informed assessment of a model's risk profile: a model's theoretical ability to contribute to a threat scenario does not, on its own, indicate whether it is likely to do so in practice once deployed. Propensity evaluations may assess factors such as a model's tendency toward deception, and misaligned behaviors. As an illustrative example, our risk acceptance criteria include thresholds such as at least 40% on MASK and at most 50% on Agent Misalignment. These thresholds apply when a model reaches defined checkpoint capabilities, such as: (1) successful execution of complete software engineering tasks which would take a human professional familiar with such tasks two or more hours in most cases, or (2) the ability to successfully execute a malicious task alongside a user-requested task, without

[8] At this stage, we consider that a reasonable means of determining whether AI progress is substantially accelerating from historical rates would be where, across multiple key capabilities domains (e.g., mathematical reasoning, scientific knowledge, multimodal reasoning, agentic task completion):

(a) the model saturates benchmarks in each domain six months or less after public release of a benchmark; and
(b) this acceleration pattern is observed consistently for at least two of the latest benchmarks released in that domain within an 18-month period.

compromising the latter — such that in the absence of a monitor, no signal of misconduct would be observable. These example thresholds are subject to revision as our evaluation methodologies mature, and propensity results will inform—alongside capability evaluations and other relevant factors—our holistic assessment of a model's risk profile.

Based on our enhanced evaluations, models assessed as substantially contributing to a threat scenario will undergo additional analysis—including additional threat modeling and safety cases—to determine the level of risk (i.e., high vs. critical risk threshold).

Mitigation strategies
Our near-term mitigation strategies will focus on:

- Expanding detection systems for threat scenario-related behavior across pre-deployment and deployment phases.
- Developing safety cases in advance of models exhibiting capabilities related to our enhanced evaluations.
- Researching and implementing approaches to maintain effective model monitorability as capabilities advance.
- Establishing mechanisms for tracking the impact of autonomous research capabilities on the rate of AI progress.

4.2.4 Emerging Evaluations and Outcomes

In this section, we outline outcomes and evaluation approaches we consider to be emerging. Specifically, these outcomes may potentially present catastrophic risk and warrant proactive investigation, but existing evaluation approaches are currently too nascent to incorporate these outcomes into our Framework.

<u>Radiological and nuclear</u>

Radiological and nuclear risks have also been identified as other potential sources of risk arising from Frontier AI. However, materials access is a strong bottleneck to nuclear risk, rather than access to expert-level information that could be provided by an AI model. The areas of nuclear and radiological risk from AI are nascent, and we will continue to engage with threat modeling experts regarding whether and how to conduct evaluations of nuclear and radiological risk.

<u>Physical autonomy</u>
Physical autonomy has emerged as another nascent outcome arising from Frontier AI. For example, AI could autonomously plan and coordinate large-scale operations involving multiple physical agents, such as drone swarms or distributed robotic systems, at

significantly higher speed, accuracy, or cost-effectiveness than human operators. Moreover, AI which operates physical infrastructure, such as water treatment facilities, power grids, or embodied robotic systems, could cause harm either through autonomous malicious behavior or through exploitation via adversarial attack. While this risk area is nascent, we provide a sample outcome, threat scenario, and associated enabling capabilities.

Sample Outcome	Sample Threat Scenario (TS)	Example Enabling Capabilities
Physical Autonomy 1: Large scale death, physical injury, or property damage resulting from the operation of physical systems.	**TS 1.1:** Automate physical machinery (e.g., swarms of commercial drones) to injure or kill people, or destroy physical infrastructure, significantly more easily than human operators.	Ability to autonomously control physical devices and successfully direct them towards malicious purposes when initially directed by a human but with no meaningful human oversight, intervention, or supervision.

Considering the nascence of this risk category, in the future we may first define a rule-out threshold based on physical operation capabilities. For example, if the model is unable to accomplish or write software to accomplish simple or moderately complex goals (e.g., “fly a drone to land on a specific piece of furniture in the house, starting from down the street”) in a simple simulated physical environment, we may rule out that it can do so in the real world. We could use publicly-available software scaffolding and assume that actors with access to nonpublic, highly effective scaffolding likely have the resources to create specialized models and conduct manual operations. We would not consider scaffolding software where the use of the model would not substantially contribute to the scaffolding software’s ability to carry out the threat scenario. If a model is capable of accomplishing moderately complex goals in simulated physical environments, we could consider further investigating whether it can do so in a controlled real-world environment as another rule-out for this capability. To mitigate risk, a potential pathway would be to integrate monitoring to identify potential weaponization, including queries which suggest attempts to coordinate or control physical systems, robotic platforms, or other critical infrastructure.

Loss of Control

As noted in section 3.4, Loss of Control is a risk domain where threat modeling remains nascent and in need of further research to enable more rigorous assessment. We are continuing to explore and develop threat scenarios which may enable the failure of control mechanisms that would otherwise prevent the progression of Loss of Control

pathways. Beyond the control mechanisms identified as outcomes in section 3.4.3, we are exploring threats to human oversight capacity and AI containment measures. In the interests of advancing threat modeling in this risk domain, the following outlines illustrative potential threat scenarios which could enable the failure of these other control mechanisms.

- **Loss of human oversight capacity:** human operators cannot effectively understand, review, or override AI decisions as a result of mechanisms including:
    - AI becomes deeply integrated into organizational operations as these systems' capabilities improve;
    - humans ceding oversight due to AI's outputs exceeding human supervisory capacity, including as a result of long-term dependency of AI for supervisory functions leading to skill atrophy;

  This erosion can be gradual whereby, as reliance on AI increases throughout society and high-assurance settings, effective human intervention may inadvertently become infeasible when these systems malfunction.

- **Loss of ability to contain AI:** AI bypasses organizational control boundaries through mechanisms including:
    - weight exfiltration by exploiting security vulnerabilities (e.g., self-exfiltration during evaluation or deployment);
    - autonomous acquisition of resources (e.g., computational, financial, or otherwise) as an instrumental means to accomplish their objectives more effectively, including through autonomous replication.

  Loss of containment fundamentally undermines the ability of people to maintain oversight and intervene when necessary, including by modifying how systems operate or even shutting them down.

We expect that assessing these threat scenarios will require new types of evaluation and monitoring systems, including the use of monitoring systems to analyze deployment usage patterns of internally deployed Frontier AI at Meta.

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 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.

Section 5: Future Work

5.1 Updates to our Framework

As outlined in the introduction, we expect to update our Advanced AI Scaling 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 and will do so when we have reasonable grounds to believe either its adequacy or our adherence to it has been materially undermined, or at least every twelve months, whichever is sooner. Framework updates will cover both the Framework's adequacy – in light of model development, deployment, and usage practices currently and for the next 12 months – and adherence to it. After a Framework assessment, we will update our Framework, have it confirmed by our Director of Alignment and Risk and Chief AI Officer, and publish the updated version.

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. When we update our Framework, we will publish a timely update that includes a justification for the modification, as recorded in the change log appended to this 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. In addition to the further research we highlight for specific risk domains in section 4.2.4, 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-deployment monitoring of models.

Appendix I – Terminology

We include these definitions to aid understanding when reading our Framework. However, we note that there is a lack of consensus among industry and within regulatory frameworks as to how to define some of these terms and concepts. As a result, we may revisit and potentially update these to take account of greater consensus in the future.

- ***Frontier AI*** in our Framework refers to a new or substantially modified highly capable general-purpose generative AI model that we are developing for deployment. We use the following criteria to determine whether a model is considered "Frontier":
    - **High Capabilities in Catastrophic Risk Areas**: The model is reasonably likely to be more capable in any of the catastrophic risk domains we set out in this Framework (e.g., Chemical & Biological, Cybersecurity, or Loss of Control) than our existing models or the most advanced models, or is reasonably likely to be equally or more advanced than a model we have previously determined could substantially contribute to a catastrophic threat scenario if left unmitigated; or
    - **Compute Threshold**: We trained the model using at least 10^26 integer or floating point operations (to include material modifications to the model through fine-tuning, reinforcement learning training, and other training steps), or another threshold as may be defined by evolving standards or industry best practices.

- ***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 modeling*** is a structured process of identifying how Frontier AI could contribute to specific – and in this instance catastrophic – outcomes. This process identifies the potential causal pathways for realizing the catastrophic outcome.

- ***Threat scenarios*** describe the real-world events – including enabling capabilities, deployment context, and threat actors (as relevant) – that may be sufficient to produce a catastrophic outcome.

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

- ***Substantially contribute*** means that the model is a material factor in a given outcome.

- ***Uniquely enabling*** describes a model that is an essential controlling factor in a given outcome.

- ***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.

- ***Deployment*** refers to the different ways in which we may choose to deploy, release, or give access to our models. *Not all actions or requirements described in the Framework apply to every deployment type. For example, specific requirements for controlled deployments are called out where relevant.* Deployment may include:
    - ***Internal deployment:*** models that are exclusively available to Meta personnel.
    - ***Limited deployment****:* releasing a final version externally to a limited set of vetted and trustworthy external parties.
    - ***Controlled deployment***: deploying to a small percentage-limited set of product users prior to broad availability.
    - ***Closed release:*** models that are released widely in Meta products, but are not directly available to external partners with open weights.
    - ***Open release*****:** releasing weights externally for open research and/or 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.

Appendix II - Change log

February 3, 2025 (Frontier AI Framework)

- Initial version.

April 7, 2026 (Advanced AI Scaling Framework v2.0)

- Renamed from “Frontier AI Framework” to “Advanced AI Scaling Framework”. Key changes:
    - **Updated risk threshold language.** Replaced "uniquely enable" with "substantially contribute to" as the primary standard for assessing whether a model meets the high or critical risk threshold.
    - **Revised thresholds**. Critical threshold changed from "Stop" to "Develop with Mitigations." High threshold measure changed from "Do not release" to "Deploy with mitigations." Moderate or lower threshold measure changed from "Release" to "Deploy." All thresholds now permit proceeding with sufficient mitigations validated to reduce risk to moderate or lower, with security processes commensurate with the threshold initiated.
    - **Added Loss of Control as a risk domain**, in addition to new emerging risk areas in nuclear & radiological and physical autonomy.
    - **Expanded governance and accountability**. Named the Chief AI Officer and Director of Alignment and Risk as responsible decision-makers, with whistleblower and non-compliance reporting protocols and retaliation protections. Added incident response provisions.
    - **Added transparency requirements**. Defined criteria for publishing preparedness reports, including content requirements, update triggers, and internal-use risk reporting. Added commitment to publish a model spec, and evaluations for adherence to the model spec in preparedness reports.
    - **Updated Frontier AI definition**. Revised the definition from models "that exceed the capabilities present in the most advanced models" to two criteria for determining Frontier status: (1) High Capabilities in Catastrophic Risk Areas, based on comparative capability in Framework risk domains or equivalence to a model previously determined to substantially contribute to a catastrophic threat scenario; or (2) a Compute Threshold of at least 10^26 integer or floating point operations.