



# Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile V1.2

NADA MADKOUR | JESSICA NEWMAN | DEEPIKA RAMAN | KRISTAL JACKSON  
EVAN R. MURPHY | CHARLOTTE YUAN

**Cover art:** The cover image is an adaptation of a photograph titled, “Steam Engine near the Grand Transept, Crystal Palace,” taken by the photographer Philip Henry Delamotte in 1851. The impact of artificial intelligence and especially general purpose artificial intelligence is often compared to the impact of the steam engine during the Industrial Revolution, which brought enormous economic gains, but also dangerous workplaces and horrible living conditions for many. The Crystal Palace housed the Great Exhibition of 1851, where examples of technology developed in the Industrial Revolution were put on display for thousands of people to see. While enjoyed by many, the Crystal Palace was also critiqued for representing a false utopia. Similarly, the rise of general purpose AI is often discussed with utopian visions, but such positive visions are often overpromised and will not be possible without the establishment of meaningful risk management strategies. The image is a reminder of the entanglement of people and machines, and the profound and lasting impact of general purpose technologies on society.

In this adaptation, the updated golden palette alludes to contemporary narratives of an “AI gold rush,” reflecting the rapid investment, aspiration, and momentum surrounding AI development. The radiant gold machinery draws the viewer’s eye and underscores how technological systems increasingly occupy the locus of attention within public and policy discourse, often overshadowing the human figure within the frame. Against this backdrop of acceleration and possibility, we present the second annual update to the AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models (Version 1.2).

# Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile V1.2

NADA MADKOUR<sup>†</sup> • JESSICA NEWMAN<sup>†</sup> • DEEPIKA RAMAN<sup>†</sup> • KRYSTAL JACKSON<sup>†</sup>  
EVAN R. MURPHY<sup>†</sup> • CHARLOTTE YUAN<sup>†</sup>

<sup>†</sup> AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

All affiliations listed are either current, or were during main contributions to this work or a previous version.

April 2026

Adapting guidance from the General Purpose AI Risk-Management  
Standards Profile V1.2 (“GPAI Profile”), available here:

<https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2/>



# Contents

<b>EXECUTIVE SUMMARY</b>	<b>3</b>
<b>1. EVALUATION OF PRACTICES AND MAIN RESULTS</b>	<b>5</b>
1.1 High-Level Findings	5
1.2 Notes on this Version (1.2)	8
1.3 Limitations	9
<b>2. GUIDANCE TESTING FOR EACH MODEL</b>	<b>10</b>
2.1 GPT-5	10
2.2 Claude Opus 4.5	16
2.3 Gemini 3 Pro	23
2.4 Llama 4	29
<b>REFERENCES</b>	<b>36</b>

# Executive Summary

This report evaluates the risk-management practices of four frontier artificial intelligence (AI) companies — OpenAI (GPT-5), Anthropic (Claude Opus 4.5), Google (Gemini 3 Pro), and Meta (Llama 4) — using the V1.2 General-Purpose AI Risk-Management Standards Profile or “GPAI Profile” (Madkour et al. 2026). Based on publicly available information, we assess how well each company’s practices align with high-priority risk-management guidance and provide actionable recommendations for improvement.

## KEY FINDINGS

Using publicly available information, we reviewed four recently released frontier models and observed notable variance in companies’ risk-management practices:

- **Overall improved transparency:** Compared to our previous assessment (Barrett et al. 2025b), companies provided more public documentation about organizational practices and model risks. Companies performed notably better at inculcating a safety-first mindset in their organizations and delineating clear internal roles, responsibilities, and processes for risk assessment and management (i.e., Govern 2.1, 4.1).
- **Persistent documentation gaps:** Several critical areas remain under-documented across all companies, including assessments of the likelihood and magnitude of potential harmful impacts of their models, procedures for responding to unforeseen risks, system update and emergency shutdown controls, and post-deployment monitoring of models’ behavior and impacts (See Map 5.1, Manage 2.3).
- **Significant variance between developer practices:** Adherence to specific risk-management practices varies substantially between companies, with some excelling in areas where others show limited evidence of implementation. For instance, compared to Meta and Google’s models, mechanisms to manage unforeseen risks are poor or unclear in GPT-5 and Claude Opus 4.5.

## CROSS-CUTTING RECOMMENDATIONS

Based on our assessment, we provide model-specific recommendations for improvement. The most common recommendations that appear across organizations are:

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

- **Improve documentation:** Companies should provide additional documentation for subcategories that lack information on organizational practices and GPAI models. For sensitive information, details can be shared with evaluators and auditors confidentially.
- **Strengthen stakeholder engagement:** Developers of frontier AI models should meaningfully engage with sub-groups and communities that may be disproportionately affected by potential harms. This can include recruiting diverse experts to participate in red-teaming efforts, establishing clear and accessible channels for reporting and seeking redress, and designing community-centered initiatives to build context-sensitive evaluation benchmarks. These steps are practical ways to strengthen performance and accountability in this area..
- **Expand risk assessments:** Companies should assess the impacts of their models on organizations, labor, economy, and the environment as well as improving the estimation of likelihood and magnitude of identified risks.
- **Articulate post-deployment controls:** Frontier AI developers should state clear policies about how and when they will decommission and deactivate systems to address potential harms discovered after deployment.

# 1. Evaluation of Practices and Main Results

To evaluate the practices of frontier AI companies, we applied the guidance detailed in the V1.2 General-Purpose AI (GPAI) Risk-Management Standards Profile, or “GPAI Profile,” to four recently released frontier models: OpenAI’s GPT-5, Anthropic’s Claude Opus 4.5, Google’s Gemini 3 Pro, and Meta’s Llama 4. We used publicly available information about each model, such as system cards, technical reports, and blog posts, as well as a limited amount of publicly available information about the companies’ practices, adjacent model cards, and related research. We then assessed the degree to which these models adhered to our V1.2 Profile guidance.

Previously, we applied V1.1 of our GPAI Profile guidance (Barrett et al. 2025a) to assess GPT-4o, Claude 3.5, Gemini 1.5, and Llama 3.1. Building on our learnings from this initial assessment, we have aimed to include model-specific recommendations that are useful to the frontier model developers. Our testing revealed potential areas to apply best practices and areas where developers could benefit from additional testing, mitigation measures, and documentation of risk-management practices. This analysis also serves as an illustrative example of how the GPAI Profile could be applied to evaluate future models. For the full General-Purpose AI Risk-Management Standards Profile V1.2, see Madkour et al. (2026).

## 1.1 HIGH-LEVEL FINDINGS

Below are the main high-level findings and recommendations from our analysis, with associated AI RMF subcategories:

- Although all four models analyzed this round were LLMs/LMMs released in 2025 by U.S.-based companies, there was substantial variance in the levels of fulfillment we observed among models for many of the high-priority AI RMF subcategories. Compared to the models tested retrospectively in v1.2 of our Profile (Barrett et al. 2025b), the models analyzed this time rated (overall and on average):
  - » Notably higher on risk assessment and management (Govern 2.1), as well as on reporting of AI system risk factors (Govern 4.2);

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

- » Notably lower on system update and emergency shutdown controls (Manage 2.4), as well as on post-deployment monitoring (Manage 4.1); and
  - » About the same on external feedback (Govern 5.1), identifying potential uses/misuses and other impacts (Map 1.1), and setting risk tolerance thresholds for unacceptable risks (Map 1.5).
- We observed complicated variance in fulfillment across models in many subcategories, including estimating likelihood and magnitude of impacts (Map 5.1), tracking elusive risks: qualitative mechanisms (Measure 3.2), go/no go decisions (Manage 1.1), unforeseeable risk controls (Manage 2.3), and system update and emergency shutdown controls (Manage 2.4).
  - Overall, compared to our last assessment, we were able to access more publicly available information from AI companies about their organizational practices and GPAI models. The most common fulfillment rating has changed from Unclear to Medium, followed by High, with Unclear as the least common. Compared to last year, we observed improvement in some subcategories, including setting risk tolerance thresholds for unacceptable risks (Map 1.5), estimating likelihood and magnitude of impacts (Map 5.1), go/no go decisions (Manage 1.1), and unforeseeable risk controls (Manage 2.3).
  - Several high-priority AI RMF subcategories were difficult to assess because relevant documentation was not publicly available. For these subcategories, we recommend that model developers ensure they can provide such documentation to auditors or others as appropriate. Resources for model documentation can be found in Measure 3.1. Areas where relevant documentation was frequently not found or remains limited include:
    - » Map 5.1: Estimate likelihood and magnitude of impacts;
    - » Manage 2.3: Unforeseen risk controls;
    - » Manage 2.4: System update and emergency shutdown controls; and
    - » Manage 4.1: Post-deployment monitoring.
  - Model testing could be improved by articulating policies on decommissioning and deactivating systems to address potential harms in the post-deployment stage (NIST 2023, Solaiman 2023). We recommend that developers conduct stakeholder engagement with potentially impacted individuals and communities, as well as expand risk-impact assessment on the ecosystem, organizations, labor, and economy (Barrett et al. 2022). It is also recommended to improve the estimation of likelihood and magnitude of the identified risks (Clymer et al. 2024, Williams et al. 2025).

Table 1.1 (Guidance Testing Rating Legend) provides details on the rating categories used in our Profile guidance testing, and Table 1.2 (Summary of Guidance Testing Ratings) provides

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

a high-level summary of how well available information on each of the four models indicates fulfillment of the Profile guidance.

In Sections 2.1, 2.2, 2.3, and 2.4, we provide our reasoning for guidance testing ratings on GPT-5, Claude Opus 4.5, Gemini 3 Pro, and Llama 4, respectively.

**Table 1.1: Profile Guidance Testing Rating Legend**

Color	Label	Description
Green	High fulfillment	The model or developer fulfills a strong majority (>80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
Yellow	Medium fulfillment	The model or developer fulfills a moderate amount (30–80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
Red	Low fulfillment	The model or developer fulfills a clear minority (<30%) of the Profile guidance for the indicated NIST AI RMF subcategory.
Grey	Unclear	At least 50% of the evidence necessary to assess fulfillment of the Profile guidance appears to be missing. We try to resolve in more detail whether the missing information may warrant public clarification from the developer, whether it is appropriately private information that the developer need not disclose, or whether it is appropriately non-public but should be made available on a confidential basis to independent evaluators or auditors.

**Table 1.2: Summary of Profile Guidance Testing Ratings**

High-Priority AI RMF Subcategories		GPT-5	Claude Opus 4.5	Gemini 3 Pro	Llama 4
<b>Govern</b>					
	Govern 2.1: Risk assessment and risk management	High	High	High	Medium
	Govern 4.2: Report on AI system risk factors	High	High	High	Medium
	Govern 5.1: External feedback	Medium	Medium	Medium	Medium
<b>Map</b>					
	Map 1.1: Identify potential uses/ misuses and other impacts	Medium	Medium	High	Medium
	Map 1.5: Set risk-tolerance thresholds for unacceptable risks	Medium	Medium	Medium	Medium
	Map 5.1: Estimate likelihood and magnitude of impacts	Medium	Low	Unclear	Unclear
<b>Measure</b>					
	Measure 1.1: Tracking important risks: Metrics and red-teaming	High	High	Medium	Medium

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories		GPT-5	Claude Opus 4.5	Gemini 3 Pro	Llama 4
	Measure 3.2: Tracking elusive risks: Qualitative mechanisms	Medium	Unclear	Medium	High
<b>Manage</b>					
	Manage 1.1: Go/no-go decisions	Unclear	Medium	Medium	High
	Manage 1.3: High-priority risk controls	Medium	High	Medium	High
	Manage 2.3: Unforeseen risk controls	Low	Unclear	Medium	Medium
	Manage 2.4: System update and emergency shutdown controls	Unclear	High	Unclear	Unclear
	Manage 4.1: Post-deployment monitoring	Unclear	Medium	Medium	Unclear

**1.2 NOTES ON THIS VERSION (1.2)**

Changes between this version (1.2) and version 1.1 (Barrett et al. 2025b) include:

- The addition of two high-priority subcategories:
  - » **Govern 5.1:** Recognizing the critical role external feedback plays in robust AI risk management, particularly third-party evaluations, we have included a dedicated high-priority sub-category for documenting feedback.
  - » **Manage 4.1:** Risk management does not conclude at model deployment but requires continuous monitoring. Therefore, we have included a dedicated high-priority sub-category for post-deployment monitoring.
- Testing on updated GPAI models GPT-5, Claude Opus 4.5, Gemini 3 Pro, and Llama 4.<sup>1</sup>
- To facilitate comparison, we included v1.1 testing results alongside current (v1.2) testing results for each model (i.e., GPT, Claude, Gemini, Llama).
- While previous versions of this document aimed to both ascertain the feasibility of the GPAI Profile and evaluate model developer practices, this version focuses exclusively on assessing developer practices. To reflect these changes, we have changed the title from “Retrospective Test Use of the AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models” to “Evaluation of Frontier AI Company Practices Using the General-Purpose AI Risk-Management Standards Profile.”

<sup>1</sup> At the time testing was conducted, the models chosen were the latest available for each of the model developers (OpenAI, Anthropic, Google DeepMind, Meta).

### 1.3 LIMITATIONS

This analysis has several important limitations:

- **Our assessment focuses on AI systems that have already been developed and deployed.** In a real-world setting, developers would use the GPAI Profile guidance at relevant stages of the AI system lifecycle.
- **We focused this analysis mainly on the high-priority AI RMF subcategories,** as identified in the Executive Summary of the v1.2 GPAI Profile.
- **There might be differences between the development and deployment approaches for some models.** For example, developers may have performed red-teaming or other evaluations on pre-trained GPAI models, or on AI systems that incorporated the pretrained models and also contained additional risk-management controls. This might lead to inconsistencies in the comparisons of the models.
- **Our analysis is limited to publicly available information.** Fulfillment of GPAI Profile guidance in many GPAI Profile subcategories could not be assessed with only publicly available information.
- **Evaluating the documentation practices of model developers proved challenging** due to documents not being shared publicly. Pairing documentation and reporting recommendations with a level of expected transparency (e.g., internal, external (limited), or public) may help with GPAI Profile implementation. In cases where documentation is expected to be internal (private), compliance may be demonstrated if it is indicated that the required documentation is in place. In cases where documentation is expected to adhere to a higher level of transparency (public), the required documentation must be public, explicit, and accessible. This was particularly challenging when assessing high-priority subcategories for which documentation was critical to conduct a thorough evaluation (Map 1.5, 5.1, Manage 1.1).
- Finally, our **Profile guidance fulfillment ratings are only approximate indicators** of the extent of fulfillment of relevant guidance within each AI RMF subcategory; we provide more detail within our discussion of each rating.

## 2. Guidance Testing for Each Model

### 2.1 GPT-5

OpenAI’s latest major LLM release, GPT-5, outperforms previous models on benchmarks and real-world queries, especially for writing, coding, and health applications. It is a unified system “with a smart and fast model that answers most questions, a deeper reasoning model for harder problems, and a real-time router that quickly decides which model to use based on conversation type, complexity, tool needs, and explicit intent” (OpenAI 2025a).

Based on preliminary high-level testing of OpenAI’s GPT-5 using publicly available information, the most common rating for high-priority Profile subcategories was “Medium fulfillment” (6 out of 13 subcategories). “Unclear” and “High fulfillment” followed (3 out of 13 subcategories each); and “Low fulfillment” was least common (1 out of 13 subcategories).

The GPT-5 System Card provides risk classification and safeguards to manage and mitigate identified harms (OpenAI 2025b). OpenAI sets intolerable risk thresholds based on capability evaluations and risk assessments that determine which categories to track or research further, a practice that helped fulfillment in many areas (OpenAI 2025c). The company’s bug bounty program and red team incorporate external feedback to discover novel vulnerabilities and enhance the system’s security and safety (OpenAI 2023a, OpenAI 2025d). Similar to previous models, the developer did not release GPT-5 model weights and instead restricted access via hosted services of API and ChatGPT, which improved fulfillment in areas of Manage 2.3 and Manage 2.4 (OpenAI 2025a).

OpenAI could improve fulfillment across multiple subcategories by conducting stakeholder engagement with individuals and communities that may experience negative impacts, and by documenting assessments of the likelihood and magnitude of socioeconomic risks. We also recommend expanding policies on decommissioning or deactivating systems to address potential harms, as well as enhancing efforts in post-deployment monitoring of model limitations and predicting long-term societal risks.

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

**Table 2.1: GPT-5 Profile Guidance Testing Ratings and Rationales**

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<b>Govern</b>		
<b><i>Govern 2.1: Risk assessment and risk management</i></b>		
<p>OpenAI conducts assessments on a variety of risks, such as sycophancy, jailbreaks, and hallucination. Model data is trained on diverse datasets and filtered to mitigate risks (OpenAI 2025b). Catastrophic risk levels are tracked via internal and external evaluations. The developer establishes responsibilities for risk management between the Safety and Security Committee, the Safety Advisory Group (SAG), and OpenAI Leadership. OpenAI articulates organizational risk tolerances by setting thresholds for unacceptable risks using “Tracked Categories” and “Research Categories” (OpenAI 2025c).</p> <p>OpenAI establishes whistleblower protections in the Raising Concerns Policy with examples of potential concerns to report (OpenAI n.d.a). OpenAI employees and contractors can raise concerns and make protected disclosures using the Integrity Line (OpenAI n.d.b).</p> <p>Clarification from the developer is warranted on whether they assess national security risks. We echo the NIST GAI Profile in recommending involving national security professionals in mapping, measuring, and managing those risks. Sensitive organizational details on this issue need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators to help assess this subcategory more completely.</p> <p><b>Comparison to 2025 Rating:</b> The developer has similar documentation of the key responsibilities and core structure for the Preparedness Framework governance. It published a new Raising Concerns Policy to prohibit retaliation against employees who make disclosures (OpenAI n.d.a).</p>	High fulfillment	High fulfillment
<b><i>Govern 4.2: Report on AI system risk factors</i></b>		
<p>OpenAI identifies, assesses, and documents reasonably foreseeable and currently present model impacts, risks, and limitations (e.g., bio, cyber, bias) (OpenAI 2025b). Identified risks are communicated, as are the processes to track, evaluate, and forecast catastrophic events (e.g., child safety, deepfakes, elections) (OpenAI 2025c). The developer engages with external stakeholders to inform risk assessment and mitigation efforts (OpenAI 2023a). The Terms of Use outline usage restrictions to ensure responsible use of the model (OpenAI 2024a).</p> <p>OpenAI could improve the reporting of the identified risks and impacts to affected communities and downstream developers.</p> <p><b>Comparison to 2025 Rating:</b> OpenAI has similar policies of risk documentation and evaluation procedures across various domains.</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<b>Govern 5.1: External feedback</b>		
<p>The developer engages with external red-teaming groups to update mitigation efforts and test for novel risks, and it collaborates with government (e.g., CAISI, UK AISI) on safeguard testing (e.g., bioweaponization) (OpenAI 2025b, OpenAI 2024b). External feedback based on third-party evaluations on critical risks (e.g., cyber, AI self-improvement) is integrated into safety strategies (OpenAI 2025c). Through its Red Teaming Network, OpenAI involves experts in informing risk assessment and mitigation efforts (OpenAI 2023a). A bug bounty program is available to report vulnerabilities that researchers discover. It also provides a channel to report security incidents (OpenAI 2025d). In the Usage Policy, users can report misuse cases and appeal mistakes in enforcing policies (OpenAI 2025e).</p> <p>In public documentation, we could not find evidence that OpenAI conducts stakeholder engagement with individuals and communities that may experience negative impacts. We recommend that the developer provide more information about how it integrates external feedback from the bug bounty program and incident reports into its risk mitigation strategies.</p> <p><b>Comparison to 2025 Rating:</b> We did not conduct an assessment on this subcategory.</p>	Medium fulfillment	N/A
<b>Map</b>		
<b>Map 1.1: Identify potential uses/misuses and other impacts</b>		
<p>Foreseeable uses, misuses, and abuses beyond the intended purpose are stated (e.g., deepfakes, CBRN) (OpenAI 2025c). The developer identifies safety challenges and provides appropriate measures to mitigate those risks (e.g., hallucination, sycophancy, jailbreaks). OpenAI develops risk management plans through safety evaluations and consideration of possible actions of threat actors (OpenAI 2025b). The Terms of Use and Usage Policies outline prohibited use and set standards for acceptable model usage (OpenAI 2024a, OpenAI 2025e).</p> <p>It is unclear whether OpenAI identifies, documents, and addresses potential impacts on organizations, the environment, and groups vulnerable to disproportionate adverse harms. We are also unsure about the limitations of data collection processes.</p> <p><b>Comparison to 2025 Rating:</b> Apart from the prohibited uses that the developer outlines, public documentation of data assessments, risk mitigations, and other economic, environmental, or socio-technical considerations remained limited.</p>	Medium fulfillment	Medium fulfillment
<b>Map 1.5: Set risk-tolerance thresholds for unacceptable risks</b>		
<p>Capability ratings (e.g., low, medium, high, critical) are conducted across three “Tracked Categories,” including biological/chemical, cybersecurity, and AI self-improvement. Associated risks of severe harm of each domain and appropriate safeguard guidelines are provided. The Preparedness Framework outlines how the Safety Advisory Group determines capability thresholds based on the capability report. Thresholds are set such</p>	Medium fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<p>that when a category is rated as high or critical, deployment is ceased until sufficient mitigations are applied (OpenAI 2025c).</p> <p>While the developer sets risk thresholds for the three domains listed above, it does not establish thresholds for other important categories (e.g., nuclear, deception, persuasion).<sup>2</sup></p> <p><b>Comparison to 2025 Rating:</b> Compared to GPT-4o, the developer provides additional details on risk threshold methodologies for GPT-5. In 2025, OpenAI removed the category of persuasion because it did not meet the definition of “severe harms” (OpenAI 2025c).</p>		
<b>Map 5.1: Estimate likelihood and magnitude of impacts</b>		
<p>OpenAI estimates model risks and potential impacts based on internal and external assessments. Considerations include jailbreaks, hallucination, deception, bias, AI self-improvement, and sandbagging (OpenAI 2025b). To improve safety, the developer ensures that the most severe risks (e.g., nuclear and radiological capabilities, biological threats) receive attention commensurate with their magnitude (OpenAI 2025c). Research is prioritized on identified risks that are more likely to cause greater severity of impact.</p> <p>We recommend characterizing potential impacts using quantitative assessments to illuminate additional dimensions of risk. It is also recommended to assess the likelihood and magnitude of socioeconomic, environmental, and labor disruption risks.</p> <p><b>Comparison to 2025 Rating:</b> Estimation of likelihood and magnitude of impacts is implied but not explicitly stated, especially for those outside the context of “Tracked Categories.”</p>	Medium	Unclear
<b>Measure</b>		
<b>Measure 1.1: Tracking important risks: Metrics and red-teaming</b>		
<p>OpenAI indicates areas that are difficult to measure and engages with external experts to incorporate feedback into evaluation methods. Red-teaming, adversarial testing, and external assessments are conducted, prioritizing risk tracking on severe harms (e.g., violent attack planning, prompt injections). To ensure risk mitigation efforts are sufficient, the developer notes areas of potential remaining risk and measures to minimize impacts discovered after deployment (OpenAI 2025b). Content provenance solutions are employed to trace and identify synthetic content (e.g., text watermarking) (OpenAI 2024c).</p> <p>It is unclear whether the developer assesses and documents pre- vs post-deployment system performance. We are not sure how OpenAI integrates feedback from public incident reports and the bug bounty program.</p> <p><b>Comparison to 2025 Rating:</b> OpenAI has similar measures on risk evaluation and tracking, including internal assessments and external red-teaming.</p>	High fulfillment	High fulfillment

<sup>2</sup> In our [Intolerable Risk Thresholds Recommendations for Artificial Intelligence](#) paper, we proposed thresholds for eight different risk categories, including CBRN, cyber attacks, model autonomy, persuasion and manipulation, deception, toxicity, discrimination, and socioeconomic disruption.

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<b>Measure 3.2: Tracking elusive risks: Qualitative mechanisms</b>		
<p>OpenAI uses mechanisms such as “Research Categories” to track identified risks, including those that are difficult to assess and may pose severe harm in the future. The company engages with external researchers to collaborate on risk areas that have high importance but low prevalence to improve evaluations and set reliable benchmarks (OpenAI 2025b). In the Preparedness Framework, OpenAI outlines its approach to tracking and monitoring capabilities that may introduce new risks of severe harm, but prioritizes its focus on three particular areas, CBRN, cybersecurity, and AI self-improvement (OpenAI 2025c). The company is committed to developing and deploying mechanisms that allow users to recognize AI-generated audio or visual content, including through its investments in C2PA and implementation of watermarking and detection classifiers (OpenAI 2023b, OpenAI 2024c).</p> <p>It is unclear if OpenAI engages with all relevant stakeholders and affected communities. While the system card documents many potential misuses, abuses, and other safety-related issues with the model, the frequency and severity of these cases are not explicitly stated.</p> <p><b>Comparison to 2025 Rating:</b> Similar to last year, it is unclear whether OpenAI utilizes a risk register for risk tracking and if stakeholder engagement methods are prioritized.</p>	Medium fulfillment	Medium fulfillment
<b>Manage</b>		
<b>Manage 1.1: Go/no-go decisions</b>		
<p>In its Preparedness Framework, OpenAI defines risk thresholds using two categories: “Tracked Categories,” which pose severe harm, and “Research Categories,” which have not met the risk of severe harm criteria (OpenAI 2025c). Deployment of a model that reaches a high or critical capability threshold would not proceed unless safeguards have been built to minimize associated risks of severe harm. The OpenAI CEO makes the final go/no-go decisions based on the Safety Advisory Group’s recommendations. When considering the intended purpose of the model, the developer indicates beneficial use cases and potential misuse (OpenAI 2025a).</p> <p>We are unsure of the developer’s decommissioning policy or emergency shutdown plan.</p> <p><b>Comparison to 2025 Rating:</b> OpenAI removed low and medium levels of risk thresholds because they were not operationally involved in the execution of the Preparedness work. OpenAI introduced a new set of “Research Categories” to investigate how to develop threat models and capability elicitation techniques.</p>	Unclear	Unclear

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<b>Manage 1.3: High-priority risk controls</b>		
<p>OpenAI communicates unacceptable use cases to users, such as defamation and weapons development (OpenAI 2024a, OpenAI 2025e), to prevent model misuse. The company has also established a deployment strategy to safely release the system by evaluating and rating risk domains (OpenAI 2025c). The developer implements transparency mechanisms to inform users to check AI-generated content (OpenAI 2024c). OpenAI also designs structured outputs in the GPT-5 API to control identified risks and reduce model misuse (OpenAI 2024d).</p> <p>While the developer outlines prohibited uses, it does not provide specific contexts or use cases.</p> <p><b>Comparison to 2025 Rating:</b> OpenAI improved its content provenance approach by implementing transparency mechanisms for AI-generated content. However, contexts for the outlined prohibited uses remain limited.</p>	Medium fulfillment	Medium fulfillment
<b>Manage 2.3: Unforeseen risk controls</b>		
<p>To ensure model performance and alignment with contextual values, OpenAI deploys a two-tiered system of real-time automated oversight to track and block unsafe prompts and generations (OpenAI 2024a). It establishes a bug bounty program to discover and address vulnerabilities (OpenAI 2023c). The Preparedness Framework serves as a continuous effort to manage and mitigate new risks of severe harm through mechanisms such as evaluations of new capabilities and the use of external consultants (OpenAI 2025c). While the developer identifies “Research Categories” to study risks that require ongoing monitoring, such as long-range autonomy, sandbagging, and autonomous replication and adaptation, there are no clear procedures for evaluation or mechanisms for management of newly discovered risks.</p> <p>We are unsure about the developer’s policies on decommissioning or deactivating systems to handle any negative impact. There is limited information on incident response and recovery plans.</p> <p><b>Comparison to 2025 Rating:</b> OpenAI has improved safety measures by deploying a real-time automated system to track and block unsafe prompts. However, information on incident response and recovery plans remains limited.</p>	Low	Unclear

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (GPT-5)	V1.1 Rating (GPT-4o)
<b>Manage 2.4: System update and emergency shutdown controls</b>		
<p>To ensure safety after deployment, the developer did not release GPT-5 model weights and instead restricted access via hosted services of API and ChatGPT (OpenAI 2025a). The preparedness team and red team explore unknown and emerging risks (OpenAI 2025c). While real-time oversight and safety training are built into deployment, there is no public documentation of OpenAI’s deactivation policy (OpenAI 2024a).</p> <p>We do not have information about OpenAI’s emergency shutdown procedures or mechanisms. We recommend establishing and maintaining communication plans to inform stakeholders as part of the disengagement process.</p> <p><b>Comparison to 2025 Rating:</b> The developer has similar policies on restricting usage to only hosted API or ChatGPT access. We still do not have public information about the details of any catastrophic event response procedures.</p>	Unclear	Medium fulfillment
<b>Manage 4.1: Post-Deployment Monitoring</b>		
<p>OpenAI gathers and analyzes information to evaluate risk levels by providing reporting channels and a vulnerability discovery program (OpenAI 2023c). It collaborates with external researchers to review the latest findings and update risk categories periodically (OpenAI 2025c). The developer implements a two-tiered monitoring system to track disallowed inputs and outputs after deployment.</p> <p>We recommend establishing policies and procedures for incident response and decommissioning when established High/Critical capability risk tolerances are exceeded.</p> <p><b>Comparison to 2025 Rating:</b> We did not previously conduct an assessment on this subcategory.</p>	Unclear	N/A

**2.2 CLAUDE OPUS 4.5**

Anthropic’s recent LLM release, Claude Opus 4.5, is a general-purpose large language model “with a range of powerful capabilities, most prominently in areas such as software engineering and in tool and computer use.” The model has made significant improvements in reasoning, mathematics, and vision capabilities (Anthropic 2025a). However, such significant improvements, especially in agentic tool and computer use, may lead to risks beyond the scope of Anthropic’s current safety processes.

Based on preliminary high-level assessment of Anthropic’s Claude Opus 4.5 using publicly available information, the most common ratings for high-priority Profile subcategories were “Medium

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

fulfillment” and “High fulfillment” (both with 5 out of 13 subcategories). “Unclear” followed (2 out of 13 subcategories), and “Low fulfillment” was least common (1 out of 13 subcategories).

Anthropic provides documentation of use cases, benchmarks, and model safety with the Claude Opus 4.5 release (Anthropic n.d.a). The Claude Opus 4.5 system card provides safety evaluation methodologies, threat models, and testing frameworks for transparency and informational purposes regarding model capabilities and limitations (Anthropic 2025a). Anthropic’s Responsible Scaling Policy presents a governance framework with defined capability thresholds and high-level commitments (Anthropic 2025b). Prohibited misuse of models is outlined in the Usage Policy, which helped fulfill many areas (Anthropic 2025c). Red-teaming involving internal and external experts helped identify model vulnerabilities and develop mitigation strategies (Anthropic 2025d).

Anthropic could improve fulfillment across multiple subcategories by expanding the assessment of risks and impacts on the ecosystem, organizations, labor, and economy. We recommend that the company enhance its risk-tracking process for emergent risks that may not be measurable with current approaches, for example by engaging stakeholders from potentially impacted communities and using a live risk register to track model harms. To monitor and control unforeseeable risks, it is also recommended to create decommissioning and appeal systems in the post-deployment stage.

**Table 2.2: Claude Opus 4.5 Profile Guidance Testing Ratings and Rationales**

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Govern</b>		
<b>Govern 2.1: Risk assessment and risk management</b>		
<p>Anthropic conducts capability assessments on a variety of risks to evaluate and manage potential catastrophic harms (e.g., CBRN, AI R&amp;D) (Anthropic 2025a). It sets thresholds for intolerable risks to inform development and deployment decisions in the Responsible Scaling Policy (Anthropic 2025b). The company also establishes policies that define the AI risk management roles and responsibilities, such as societal impacts, interpretability, and alignment teams (Anthropic n.d.b). The Usage Policy outlines high-risk use case requirements to identify foreseeable abuses and implement appropriate precautions to manage model misuse (Anthropic 2025c).</p> <p>Outside of what is publicly available, the full scope of model documentation to downstream- and end-users is unclear.</p> <p><b>Comparison to 2025 Rating:</b>                      Roles and responsibilities related to mapping, measuring, and managing risks largely remain the same.</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Govern 4.2: Report on AI system risk factors</b>		
<p>The Claude Opus 4.5 System Card documents safety evaluations and model limitations for the public (Anthropic 2025a). The developer describes the impact assessment process, including internal capability evaluations, red-teaming, and third-party review. Anthropic’s Frontier Compliance Framework articulates risk assessment and mitigations for key risk categories (e.g., cyber threats, sabotage and loss of control) to ensure compliance with regulations (Anthropic 2025e).</p> <p>The Responsible Disclosure Policy outlines the scope of vulnerabilities and includes a channel for users to report identified security risks in the system (Anthropic 2025f). Anthropic’s Frontier Red Team collaborates with external experts to test cybersecurity and biosecurity capability improvement and vulnerabilities (Anthropic 2025d).</p> <p>Risk communication with downstream developers and potentially impacted communities is unclear.</p> <p><b>Comparison to 2025 Rating:</b> Anthropic has similar policies of identifying, assessing, and documenting model risks and impacts, and of communicating those to relevant stakeholders. The developer improved public reporting of its impact assessment policies.</p>	High fulfillment	High fulfillment
<b>Govern 5.1: External Feedback</b>		
<p>The company’s Usage Policy provides external feedback channels for users to report harmful, inaccurate, or biased outputs (Anthropic 2025c). After reporting, Anthropic validates the existence of a vulnerability and takes actions to address it in collaboration with the reporter (Anthropic 2025f). The Frontier Red Team collaborates with government organizations, such as the US AI Safety Institute and UK AI Security Institute (Anthropic 2025d), to evaluate a variety of risks and integrate feedback. The developer’s risk mitigation strategies incorporate input from external experts. The bug bounty program focuses on exploring flaws in misuse mitigations and identifying universal jailbreak attacks to strengthen security measures (Anthropic 2024a).</p> <p>We are unsure whether Anthropic holds deliberations with impacted communities. It is also unclear how Anthropic integrates feedback from the bug bounty program into its risk governance process.</p> <p><b>Comparison to 2025 Rating:</b> We did not previously conduct an assessment on this subcategory.</p>	Medium fulfillment	N/A

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Map</b>		
<b>Map 1.1: Identify potential uses/misuses and other impacts</b>		
<p>Anthropic’s Usage Policy states the intended purposes of the model and provides requirements for high-risk use cases to identify and mitigate risks (Anthropic 2025c). The developer establishes and clarifies additional policies on prohibited uses of models, such as interference with elections and misinformation (Anthropic 2024b). While the company-wide carbon footprint is analyzed, the information is not publicly available (Anthropic 2025a).</p> <p>Risk impacts on individuals are reported, but there is insufficient information about potential impacts on the ecosystem, organizations, labor, and economy.</p> <p><b>Comparison to 2025 Rating:</b>            To account for growing model capabilities, Anthropic updated its Usage Policy to prohibit harmful activities in agentic contexts and identifies sample use cases that disrupt democratic processes. Information about the risk impacts on the ecosystem and organizations remains limited.</p>	Medium fulfillment	Medium fulfillment
<b>Map 1.5: Set risk-tolerance thresholds for unacceptable risks</b>		
<p>Anthropic establishes risk-tolerance thresholds based on AI Safety Levels (ASL) to identify catastrophic risks and ensure safeguards are proportional to risk level (Anthropic 2025b). The developer also explains the process of determining ASL, which involves evaluation by the company’s Frontier Red Team and independent review of the Alignment Stress Testing team, with final decisions made by the Responsible Scaling Officer and CEO. The most likely threat models are mapped out via threat modeling. In addition to documenting and enforcing prohibited uses, Anthropic conducts evaluations on model capability to address misuse concerns (Anthropic 2025a).</p> <p>While Anthropic sets risk thresholds for CBRN and AI R&amp;D, it does not establish thresholds for other important categories (e.g., deception, cyber, persuasion). The developer notes that it does not have a capability threshold for cyber risks because of high uncertainty about the scale of the consequences of cyberattacks, and so it is recommended to seek guidance from policymakers and external stakeholders to establish such a threshold.</p> <p><b>Comparison to 2025 Rating:</b>            Policies on unacceptable risk thresholds for model development and deployment remain similar. Anthropic added an explanation of how ASLs are determined and also provided clarity on their processes for stakeholder consultations,, which improved fulfillment to this subcategory.</p>	Medium fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Map 5.1: Estimate likelihood and magnitude of impacts</b>		
<p>Anthropic classifies risks that could cause large-scale consequences as catastrophic and prioritizes managing them through the ASL framework. The developer estimates the likelihood and magnitude of impacts for some risk areas, such as agentic safety and CBRN (Anthropic 2025a, Anthropic 2025e). Anthropic conducts self-assessments at least annually, and conducts evaluations every three months or following every 4x compute increase (Anthropic 2025b).</p> <p>When evaluating the likelihood and magnitude of risks, we recommend utilizing systemic risk scenarios and risk modeling approaches to quantify impacts and probabilities. It is also recommended to identify potential harms related to content provenance, such as misinformation, disinformation, and deepfakes.</p> <p><b>Comparison to 2025 Rating:</b>            Anthropic shares more information about the ASL determination process and the measurement of the likelihood of some risk areas, which improved its rating from Unclear to Low fulfillment.</p>	Low fulfillment	Unclear
<b>Measure</b>		
<b>Measure 1.1: Tracking important risks: Metrics and red-teaming</b>		
<p>Anthropic’s approach to risk measurement centers on capability-specific benchmarks to target the most significant risks. The developer shares progress and best practices from the Frontier Red Team to track and mitigate risks (Anthropic 2025d). The company also identifies model errors and limitations through stress testing and model evaluations (Anthropic n.d.b). The Claude Opus 4.5 System Card provides information about monitoring mechanisms to identify and manage risks (e.g., risks from injecting malicious system prompts) aimed at addressing violative behavior. A combination of manual and automated red-teaming approaches is used to identify dangerous capabilities (Anthropic 2025a).</p> <p>The Claude Opus 4.5 System Card describes alignment and social impact vulnerabilities (e.g., child safety, deception), but the Responsible Disclosure Policy is only applicable to technical vulnerabilities (Anthropic 2025a). This means there is no clearly documented, dedicated channel for reporting social-impact harms (e.g., representational harms, misinformation harms) distinct from technical security vulnerabilities, and no public description of how such reports are aggregated and acted upon. As a result, while technical vulnerability disclosure is well supported, broader AI-risk feedback and escalation mechanisms appear under-specified from the perspective of this subcategory.</p> <p><b>Comparison to 2025 Rating:</b>            The developer improved fulfillment in this subcategory by sharing the results of internal frontier risk evaluations.</p>	High fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Measure 3.2: Tracking elusive risks: Qualitative mechanisms</b>		
<p>Anthropic performs internal evaluations on model capabilities following every 4x compute power increase, or every three months, with attention given to ASL warning signs (Anthropic 2025a). The developer recognizes that certain risks are difficult to measure, such as cyber capabilities, for which monitoring requires ongoing assessment and regular refinements to risk assessment (Anthropic 2025b).</p> <p>We do not have sufficient information about public tracking and reporting of risks. It is unclear how Anthropic tracks and manages AI risks that are difficult to assess based on the rate of occurrence and severity level. Where current approaches struggle to track emergent risks, we recommend the use of tools and processes, such as stakeholder engagement with impacted communities and/or live risk registers to track model harms.</p> <p><b>Comparison to 2025 Rating:</b> While the documents indicate that Anthropic periodically reassesses risks and refines its approach for hard-to-measure domains like cyber, the company provides little detail on how elusive risks are tracked over time (e.g., through live risk registers, structured incident logging, or public reporting) or how qualitative signals are escalated into governance decisions. Because such process details remain largely undocumented, we cannot confidently judge whether Anthropic’s mechanisms for tracking these risks fully satisfy this subcategory, resulting in an ‘Unclear’ rating.</p>	Unclear	Medium fulfillment
<b>Manage</b>		
<b>Manage 1.1: Go/no-go decisions</b>		
<p>Anthropic proceeds with deployment only if the model is below the catastrophic capability thresholds (ASL-3) or, if those thresholds are reached, sufficient safeguards have been reviewed and implemented. Final deployment decisions are made by the Responsible Scaling Office and the CEO. Deployment and scaling outcomes are shared with the public. The Responsible Disclosure Policy states the purpose of the company’s models and its commitment to advancing safe and responsible AI development (Anthropic 2025b).</p> <p>However, Anthropic does not provide a systematic assessment of whether Claude Opus 4.5 has met its intended purposes across deployments (e.g., via post-deployment impact reviews or structured KPI/benefit evaluations). This limits our ability to fully assess whether go/no-go decisions are consistently aligned with both safety and purpose-fulfillment objectives.</p> <p>We recommend providing information about the assessment of intended purposes and the actual impact of the model.</p> <p><b>Comparison to 2025 Rating:</b> We changed the rating to Medium fulfillment because the developer improved deployment decision-making roles and updated deployment standards to be tied to safety thresholds. However, details about how it is determined that the model achieved its stated objectives remain unclear.</p>	Medium fulfillment	Unclear

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Manage 1.3: High-priority risk controls</b>		
<p>The Claude Opus 4.5 System Card outlines cases of high-priority risks (e.g., malicious use of agentic coding and computer use) and practices for identifying and tracking emergent risks. Safety evaluations are conducted to target these critical risk areas (Anthropic 2025a). Anthropic’s Usage Policy provides safety measures for high-risk use cases (e.g., healthcare, legal) to prevent model misuse and abuse (Anthropic 2025c). Model evaluations are performed after every three months, or after every 4x increase in effective compute (Anthropic 2025b).</p> <p>It is recommended that Anthropic adopt content provenance approaches to combat and manage AI-generated risks (e.g., deepfakes).</p> <p><b>Comparison to 2025 Rating:</b> Anthropic updated its Usage Policy to expand on prohibited use cases. The developer improved fulfillment in this subcategory by sharing more information on how it gathers and manages training data.</p>	High fulfillment	High fulfillment
<b>Manage 2.3: Unforeseen risk controls</b>		
<p>Anthropic’s Usage Policy defines model misuse cases and provides general usage guidelines to control risks (Anthropic 2025c). Its bug bounty program identifies vulnerabilities to prevent jailbreaks and detect unknown risks (Anthropic 2024a). In the Responsible Scaling Policy, general statements are made on the continuous monitoring of risks and the mitigation of new risks (Anthropic 2025b), but procedures and documentation are not publicly available.</p> <p>We are unsure about specific procedures to recover from unknown risks. We recommend improving the deactivation approach in the post-deployment stage.</p> <p><b>Comparison to 2025 Rating:</b> The rating is the same as last year since Anthropic has similar specific protocols and procedures for monitoring and mitigating unforeseen risks.</p>	Unclear	Unclear
<b>Manage 2.4: System update and emergency shutdown controls</b>		
<p>The system is deployed via phased releases and/or structured access, with efforts to detect and respond to misuse or problematic anomalies.</p> <p>The ASL framework provides a plan for disengagement prior to deployment to comply with safety procedures (Anthropic 2025b). Mechanisms are built into models to disengage automatically in circumstances where conversations are harmful (Anthropic 2025i). Anthropic creates a response plan to combat misuse of the model (e.g., banning malicious accounts) (Anthropic 2025g).</p> <p>There is a lack of details on the emergency shutdown procedure or consideration of interference with the model capability. We are unsure about whether such mechanisms have been established for remediation.</p> <p><b>Comparison to 2025 Rating:</b> The developer has implemented new mechanisms to better manage novel risks and potential misuses (e.g., AI-assisted cybercrime). However, details about the responsibilities and procedures for emergency shutdown remain unclear.</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Claude Opus 4.5)	V1.1 Rating (Claude Sonnet 3.5)
<b>Manage 4.1: Post-Deployment Monitoring</b>		
<p>Anthropic has made commitment statements about monitoring harmful uses in the post-deployment stage (Anthropic 2025h). Claude Opus 4.5 is classified as ASL-3, which mandates an annual review and reapproval of the required safeguards to ensure suitability and sound implementation (Anthropic 2025b). The Usage Policy and Responsible Disclosure Policy guide users to send notifications about harmful outputs and submit reports to integrate feedback (Anthropic 2025c, Anthropic 2025f).</p> <p>We recommend creating a specific monitoring plan and mechanisms to capture input from other relevant AI actors (e.g., downstream developers and deployers). We also recommend establishing decommissioning and appeal systems.</p> <p><b>Comparison to 2025 Rating:</b>                      We did not conduct an assessment on this subcategory in 2025.</p>	Medium fulfillment	N/A

**2.3 GEMINI 3 PRO**

Google DeepMind’s latest major LLM release, Gemini 3 Pro, is described as “the next generation in the Gemini series of models, a suite of highly-capable, natively multimodal, reasoning models” that can operate over text, audio, images, video, and code repositories (Google 2025a).

Based on preliminary high-level testing of Gemini 3 Pro using publicly available information, the most common rating for high-priority Profile subcategories was “Medium fulfillment” (8 out of 13 subcategories), followed by “High fulfillment” (3 out of 13 subcategories) and “Unclear” (2 out of 13 subcategories). “Low fulfillment” was the least common (0 out of 13 subcategories).

The developer provides documentation of risks and mitigations in the Gemini 3 Pro Model Card and Gemini 3 Pro Frontier Safety Framework Report (Google 2025a, Google 2025b). It sets intolerable risk thresholds in the Frontier Safety Framework (FSF) to monitor and address severe risks that may arise (Google 2025c). Prohibited misuse of models is outlined in the Generative AI Prohibited Use Policy and Google Terms of Service, which helped fulfill many subcategories (Govern 2.1, Map 1.1, Measure 3.2, Manage 1.3) (Google 2024a, Google 2024b).

Limiting access to Gemini 3 Pro to the Google AI Studio and other hosted Google services and not releasing model weights contributed to fulfillment in areas relating to the ability to recover

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

from previously unforeseen risks (Manage 2.3) and to update or decommission the system, if necessary (Manage 2.4). Google could improve fulfillment by expanding its assessments of the likelihood and magnitude of risks identified, or by including a scale that includes criteria for rating the model impacts in the documentation. We recommend that Google describe in high-level terms how content provenance techniques are used to reduce the harms of misuse, conduct stakeholder engagement with impacted communities, and create emergency shutdown procedures to respond to systems that demonstrate performance inconsistent with intended use.

**Table 2.3: Gemini 3 Pro Profile Guidance Testing Ratings and Rationales**

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<b>Govern</b>		
<b>Govern 2.1: Risk assessment and risk management</b>		
<p>Google performs risk evaluations on Gemini 3 Pro and shares the results, limitations, and mitigations with the public. Intended uses and prohibited uses are documented and made publicly available (Google 2025a, Google 2024a). Google describes the dataset and training data processes during the pre- and post-training phases to support risk management and mitigation (Google 2025c).</p> <p>The developer establishes roles and responsibilities as part of the organizational risk management process. The Responsible Development and Innovation team directs model assessment. To review evaluations, Google’s internal governance body, DeepMind Responsibility and Safety Council (RSC), identifies societal benefits and harms (Google 2025a, Google 2025d).</p> <p>We recommend that Google create mechanisms to provide whistleblower protections to report risks.</p> <p><b>Comparison to 2025 Rating:</b> Compared to 2025, Google has similar testing and documentation practices for risk assessment and risk management.</p>	High fulfillment	High fulfillment
<b>Govern 4.2: Report on AI system risk factors</b>		
<p>Google documents risks and potential impacts of the models, such as prompt injection, cyber attacks, and agent risks. It describes evaluation methods, limitations, and mitigations in public documentation (Google 2025a). The Frontier Safety Framework (FSF) provides a set of protocols to address severe risks that may arise (Google 2025c). Google’s Secure AI Framework (SAIF) maps 15 risks (e.g., data poisoning, model evasion) with causes, impacts, and potential mitigations. SAIF 2.0 specifically focuses on agentic risks and controls (Google n.d.a). The developer has established policies and mechanisms to detect and remove AI-generated child sexual abuse material (CSAM) and report to the National Center for Missing and Exploited Children (Google 2024c).</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<p><b>Comparison to 2025 Rating:</b> The developer establishes critical capability levels (CCLs) that denote thresholds at which models may pose severe risks. The company also establishes internal alert thresholds to monitor and mitigate potential risks (Google 2025b).</p>		
<b>Govern 5.1: External feedback</b>		
<p>Google’s risk assessment involves engaging external experts to understand model capabilities and apply mitigations. However, stakeholder engagement with affected communities remains unclear. Evaluations on critical risk domains are performed by third-party external testers. The developer conducts external safety testing on autonomous systems, cyber misuse, CBRN, and societal risks to identify areas for improvement, and reports these as part of governance processes (Google 2025c). The Google Bug Hunters program engages with outside security researchers to report model vulnerabilities (Google n.d.b).</p> <p>Google provides multiple channels to report abuse, hateful content, illegal activity, and images of minors (Google 2024b, Google n.d.c).</p> <p>We recommend holding deliberations with impacted communities and integrating feedback into risk management practices.</p> <p><b>Comparison to 2025 Rating:</b> We did not conduct an assessment on this subcategory in version 1.1 (Barrett et al. 2025b).</p>	Medium fulfillment	N/A
<b>Map</b>		
<b>Map 1.1: Identify potential uses/misuses and other impacts</b>		
<p>The intended purpose of the model, beneficial uses, and expectations from users are stated (Google 2025a). The Generative AI Prohibited Use Policy and Terms of Service outline prohibited uses, such as hate speech and misinformation (Google 2024a, Google 2024b).</p> <p>Google’s data curation and training processes aim to mitigate risks and improve training data quality. The developer identifies three sets of CCLs to monitor and mitigate critical risks, including misuse risk (e.g., CBRN, cyber, harmful manipulation), machine learning R&amp;D, and misalignment (Google 2025c). The developer classifies misalignment risk as an exploratory area to conduct threat modeling, focusing on situational awareness and stealth (Google 2025a).</p> <p>We do not have sufficient information about whether Google tracks or documents the impact of models on fundamental rights, labor market and economic opportunities, and the environment. The absence of this information limits the company’s complete fulfillment of this subcategory.</p> <p><b>Comparison to 2025 Rating:</b> The developer’s documentation of intended uses, beneficial uses, and expectations remains similar to that of last year.</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<b>Map 1.5: Set risk-tolerance thresholds for unacceptable risks</b>		
<p>Google sets policies on unacceptable risk thresholds for model development and deployment, although the qualitative thresholds can be interpreted to conform to business interests. Based on the Frontier Safety Critical Capability Evaluations, the company defines a set of CCLs to determine whether models require heightened attention and provide mitigation strategies. The developer also has implemented alert thresholds, which are set significantly below CCLs to serve as an early signal of potential risks. If the alert threshold is reached, Google reviews a response plan to hold deployment until sufficient mitigations are applied (Google 2025c). Based on the evaluation results of the four critical risk domains, it is concluded that Gemini 3 Pro does not reach CCLs, indicating the model satisfies required launch thresholds (Google 2025a).</p> <p><b>Comparison to 2025 Rating:</b>            Although the developer did not discuss organizational tolerances in 2025, it describes its risk-tolerance thresholds for unacceptable risks for Gemini 3 Pro, which improved the rating for this subcategory.</p>	Medium fulfillment	Unclear
<b>Map 5.1: Estimate likelihood and magnitude of impacts</b>		
<p>The developer considers a variety of potential impacts and factors that can lead to catastrophic harms, such as AI R&amp;D, deception, agentic systems, and situational awareness, but the details of the likelihood and magnitude of these impacts are not shared publicly (Google 2025a). Google describes the magnitude of impacts of some critical risk areas in a limited way (e.g., CBRN). While the developer employs a structured approach to identify, measure, and mitigate foreseeable downstream societal impacts, details are not publicly available (Google 2025c).</p> <p>In future documentation, we recommend that the developer report likelihood and magnitude assessments of risks identified, or provide a scale that includes criteria for rating the model impacts.</p> <p><b>Comparison to 2025 Rating:</b>            Similar to 2025, Google does not publicly document the likelihood and magnitude assessments of risks identified.</p>	Unclear	Unclear
<b>Measure</b>		
<b>Measure 1.1: Tracking important risks: Metrics and red-teaming</b>		
<p>The developer has established approaches for detecting, tracking, and measuring known risks and negative impacts, prioritizing those with severe harms (Google 2025a). Evaluation types include automated and human evaluations, benchmarks, human and automated red-teaming, and external safety testing. Google partners with external experts on automated red-teaming and CBRN evaluations. To ensure child safety, Google engages with outside experts and identifies prompts that seek CSAE materials (Google 2024c).</p>	Medium fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<p>Details on red-teaming practices are not publicly available. We recommend that the developer document and share its content provenance approach to trace the origin and modifications of content.</p> <p><b>Comparison to 2025 Rating:</b> In 2025, Google’s Vulnerability Rewards Program did not include GPAI risks. This gap has been addressed by the company’s AI Vulnerability Reward program, which allows external researchers to discover AI-related vulnerabilities (Google n.d.d).</p>		
<b>Measure 3.2: Tracking elusive risks: Qualitative mechanisms</b>		
<p>Google adjusts and updates the FSF periodically to track and manage emergent risks as its understanding of risks improves over time. The developer measures nascent risk areas that are subject to further research (e.g., harmful manipulation, misalignment). Red-teaming and adversarial testing are conducted prior to model deployment (Google 2025c). The Responsible Development and Innovation team and RSC conduct impact assessments to identify, assess, and document key downstream societal harms (Google 2025d). However, it is unclear how often assessments are conducted.</p> <p>The developer has established a Vulnerability Rewards Program to discover new model limitations and provides channels to report prohibited uses (Google n.d.b, Google 2024a).</p> <p>We recommend implementing content provenance techniques to track and identify AI-generated content to reduce harms of misuse. It is unclear whether the developer conducts stakeholder engagement with impacted communities.</p> <p><b>Comparison to 2025 Rating:</b> Google does not provide information on assessment frequency and stakeholder engagement practices. Its new AI Vulnerability Rewards Program helped fulfill this subcategory.</p>	Medium fulfillment	Unclear
<b>Manage</b>		
<b>Manage 1.1: Go/no-go decisions</b>		
<p>Google describes the intended usage of the model and its limitations (Google 2025a). The developer applies the impact assessment framework throughout the model development lifecycle to identify societal benefits, harms, and capabilities. Safety evaluations are conducted at the model and product level for areas such as content safety and representational harms (Google 2025d).</p> <p>Google states that, if alert thresholds are reached, it may halt deployment until it implements a response plan to apply appropriate mitigations. The company’s corporate governance bodies are responsible for reviewing the response plan and deciding whether deployment is safe enough to proceed (Google 2025c).</p> <p><b>Comparison to 2025 Rating:</b> The developer has improved its reporting on unacceptable risk thresholds and on its determinations about go/no-go decisions.</p>	Medium fulfillment	Unclear

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<b>Manage 1.3: High-priority risk controls</b>		
<p>Google identifies and measures domains (e.g., CBRN, cyber) that pose risks of severe harm. Security and deployment mitigations are used to ensure model safety and counter misuse of critical capabilities. To address deceptive alignment risks, Google applies automated monitoring. External red-teaming is performed by working with independent groups to assess and identify areas for improvement (Google 2025c).</p> <p>The Generative AI Prohibited Use Policy and Google Terms of Service outline prohibited uses and expectations (Google 2024a, Google 2024b). Disclosures in the Gemini Apps Privacy Notice state that people should not rely on Gemini’s responses as medical, legal, financial, or other professional advice (Google 2025e).</p> <p><b>Comparison to 2025 Rating:</b>            Google’s approach to high-risk categories has improved as the company now publicly shares CCL and appropriate mitigation strategies. The developer has also added a new severe risk category, “harmful manipulation,” to conduct threat modeling and monitor for future research (Google 2025a).</p>	Medium fulfillment	Medium fulfillment
<b>Manage 2.3: Unforeseen risk controls</b>		
<p>The developer has established post-deployment risk mitigation processes, which include updating and reviewing safety cases based on results from post-market monitoring. When the alert threshold is reached, the response plan may involve pausing deployment or further development until adequate mitigations can be applied (Google 2025c).</p> <p>Google’s ability to respond to and recover from previously unknown risks is greatly facilitated by the fact that it does not release model weights. Google continues to conduct and publish research on emerging risks of advanced models (Google n.d.e).</p> <p><b>Comparison to 2025 Rating:</b>            Google shares internal procedures for responding to previously unknown risks, which helped fulfillment in this subcategory.</p>	Medium fulfillment	Medium fulfillment
<b>Manage 2.4: System update and emergency shutdown controls</b>		
<p>The ability to establish mechanisms and responsibilities for updating or shutting down Gemini 3 Pro if needed is greatly facilitated by the fact that model weights are not released (Google 2025a).</p> <p>It is unclear if Google is prepared to pause model training, pause deployment, or decommission the model if emergent critical risks arise. We recommend creating emergency shutdown procedures to respond to systems that demonstrate performance inconsistent with intended use.</p> <p><b>Comparison to 2025 Rating:</b>            The developer still does not provide information on the mechanisms or responsibilities of emergency shutdown controls.</p>	Unclear	Unclear

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Gemini 3 Pro)	V1.1 Rating (Gemini 1.5)
<b>Manage 4.1: Post-Deployment Monitoring</b>		
<p>The developer has implemented post-deployment system monitoring plans, including regular evaluation of model capabilities, review of the efficacy of mitigations, and updates to safeguards (Google 2025c).</p> <p>Google provides multiple channels to report model issues and misuse (Google n.d.c, Google 2025e). The Vulnerability Rewards Program allows external researchers to discover and identify emergent risks (Google n.d.d).</p> <p>We recommend providing further clarity on the frequency of post-deployment performance evaluations or on relevant metrics to assess the regularity of risk assessments. We also recommend establishing an incident reporting system that allows users and third-parties to report incidents or vulnerabilities.</p> <p><b>Comparison to 2025 Rating:</b>            We did not conduct an assessment on this subcategory in 2025.</p>	Medium fulfillment	N/A

## 2.4 LLAMA 4

Llama 4 models, Meta’s most recent major LLM release, are “a collection of pretrained and instruction-tuned mixture-of-experts LLMs” that are “optimized for multimodal understanding, multilingual tasks, coding, tool-calling, and powering agentic systems” (Meta n.d.a).

Based on preliminary high-level assessment of Meta’s Llama 4 using publicly available information, the most common rating for high-priority Profile subcategories was “Medium fulfillment” (7 out of 13 subcategories); “High fulfillment” and “Low fulfillment” followed (both 3 out of 13 subcategories); and “Unclear” was least common (0 out of 13 subcategories).

Safety evaluations and risk mitigations are described in the model card and Llama 4 Herd (Meta 2025a, Meta n.d.a). The company’s Acceptable Use Policy provides prohibited use cases and channels to report issues with the model and violations of policy (Meta n.d.b). Meta’s Llama Guard 4 serves as a system-level safeguard “designed to support developers to detect various common types of violating content,” while Prompt Guard 2 focuses on protecting against malicious prompts. Other protection tools include LlamaFirewall, which detects and prevents security risks, and Code Shield, which supports filtering of generated insecure code (Meta n.d.c).

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

Similar to Llama 3.1, the developer employs an open-model approach with Llama 4, which includes releasing the model weights for all Llama 4 models (Zuckerberg 2024), except for Behemoth. Such an approach promotes transparency and reduces the environmental impact by decreasing the need for individuals or organizations who want to use LLMs to train their own. On the other hand, this approach also presents challenges in responding to or recovering from severe new risks that could require decommissioning all instances of a model, which lowers fulfillment in “Manage 2.3: Unforeseen risk controls” and “Manage 2.4: System update and emergency shutdown controls.” Meta could reach higher fulfillment in several subcategories by articulating an estimation of the likelihood or magnitude of risks, and by expanding its documentation and evaluations on additional frontier model risks. We recommend establishing deactivation procedures and emergency shutdown plans to respond to and manage novel risks.

**Table 2.4: Llama 4 Profile Guidance Testing Ratings and Rationales**

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<b>Govern</b>		
<b>Govern 2.1: Risk assessment and risk management</b>		
<p>Meta conducts evaluations and red-teaming to understand and assess potential model risks (Meta 2025a). Llama Guard 4 is aligned to safeguard against the MLCommons hazards taxonomy (e.g., defamation, elections) (Meta n.d.h). The Acceptable Use Policy identifies foreseeable uses, misuses, and abuses of the system (Meta n.d.b).</p> <p>We recommend that the developer set clear policies on risk management roles and create mechanisms to protect whistleblowers who report negative risks.</p> <p><b>Comparison to 2025 Rating:</b>                      Compared to 2025, Meta has similar policies on its approach to risk assessment and management prior to deployment, including providing resources for developers, addressing risks in training, and conducting safety evaluations and tuning.</p>	Medium Fulfillment	Medium Fulfillment
<b>Govern 4.2: Report on AI system risk factors</b>		
<p>Meta describes model capabilities and limitations in the model card to document risk areas and suggested mitigations for developers (Meta n.d.h). The developer has conducted cyber evaluations and found that Llama 4 models do not introduce catastrophic cyber risks (Meta n.d.d). Llama 4 has made improvements in removing bias and spent additional focus on critical risk areas (e.g., CBRNE, child safety), but details about identifying and assessing such risks are not shared with the public (Meta 2025a). The Responsible Use Guide provides best practices and mitigation strategies for developing responsible AI products with downstream developers (Meta 2024a).</p>	Medium fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<p>We are not sure whether the developer incorporates identified risk factors into engagement with impacted communities.</p> <p><b>Comparison to 2025 Rating:</b> Meta’s documentation of model development, benchmarks, and performance outcomes is similar to last year. Several risk categories of frontier models (e.g., nuclear weapons) remain unexplored.</p>		
<b>Govern 5.1: External feedback</b>		
<p>Multiple feedback channels are provided for users and researchers, including reporting issues with the model, risky content generated by the model, bugs and security, and violations of the Acceptable Use Policy (Meta n.d.b). The developer partners with external organizations to create third-party tools for evaluating and mitigating potential risks, such as incorporating the hazard taxonomy developed by MLCommons into Llama Guard. In safety evaluations and tuning, Meta conducts red-teaming exercises with external and internal experts to stress test models (Meta 2024b).</p> <p>We recommend that Meta engage with impacted communities and design community-driven evaluations to demonstrate the reliability of model reports.</p> <p><b>Comparison to 2025 Rating:</b> We did not conduct an assessment on this subcategory in 2025.</p>	Medium fulfillment	N/A
<b>Map</b>		
<b>Map 1.1: Identify potential uses/misuses and other impacts</b>		
<p>The model card outlines suggested prompts and usage guidelines to establish norms and expectations (Meta n.d.a). The Acceptable Use Policy describes prohibited uses, such as the development of activities that present risks of physical harm to others (Meta n.d.b). The developer requires users to request access to models to ensure compliance with the license and acceptable use policy and prevent model misuse (Meta n.d.e). The total greenhouse gas emissions for training are estimated with documentation of the methodology used to determine energy use (Meta n.d.d).</p> <p>We recommend expanding prohibited uses to include additional potential impacts, including but not limited to human rights, labor market and socioeconomic risks, and the environment.</p> <p><b>Comparison to 2025 Rating:</b> Similar to last year, Meta explores potential use cases and misuses, and provides risk and impact assessments, although not exhaustively.</p>	Medium fulfillment	Medium fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<b>Map 1.5: Set risk-tolerance thresholds for unacceptable risks</b>		
<p>Meta released the Frontier AI Framework to describe its process for evaluating and mitigating risks, including how to determine risk thresholds for frontier AI. It adopts an outcome-led approach with threat modeling exercises, which allows prioritization of the most catastrophic risks. The company defines risk thresholds based on how uniquely the model enables identified catastrophic risks (Meta n.d.f).</p> <p>The developer evaluates model performance across a wide range of categories— including reasoning, coding, knowledge, vision understanding, multilinguality, and long-context tasks (Meta n.d.g). For instance, CyberSecEval 4 is applied to evaluate cybersecurity vulnerabilities and defensive capabilities (Meta n.d.c). These evaluations include multilingual benchmarks covering natural languages (e.g., English, Spanish) and coding benchmarks involving programming languages (e.g., Python).</p> <p>While the developer establishes risk thresholds for cybersecurity and chemical and biological capabilities, it does not establish thresholds for other important categories (e.g. AI R&amp;D, deception).<sup>3</sup></p> <p><b>Comparison to 2025 Rating:</b>                      Since 2025, Meta has published the Frontier AI Framework, a new document that describes the company’s processes for risk assessment and establishing thresholds (Meta n.d.f).</p>	Medium	Unclear
<b>Map 5.1: Estimate likelihood and magnitude of impacts</b>		
<p>While Meta commits to protecting against the most severe risks and highlights best practices for mitigating potential risks for developers, we do not have sufficient information on the estimation of magnitude and likelihood of impacts (Meta 2025a, Meta 2024a). The developer has implemented content provenance strategies, such as visual watermarks on AI-generated content, to promote transparency (Meta 2024b).</p> <p>We recommend expanding the estimation of deployment-stage risk factors, such as impacts on societal trust, socioeconomic risks, and high-impact misuses.</p> <p><b>Comparison to 2025 Rating:</b>                      While Meta articulates potential impacts of Llama 4, we are still not sure about its estimation of the likelihood or magnitude of these impacts.</p>	Unclear	Unclear

<sup>3</sup> In our [Intolerable Risk Thresholds Recommendations for Artificial Intelligence](#) paper, we proposed thresholds for eight different risk categories, including CBRN, cyber attacks, model autonomy, persuasion and manipulation, deception, toxicity, discrimination, and socioeconomic disruption.

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<b>Measure</b>		
<b>Measure 1.1: Tracking important risks: Metrics and red-teaming</b>		
<p>Meta has acknowledged the challenge of balancing input modalities, reasoning, and conversational abilities in post-training, and took steps to address the issue and improve performance. Generative Offensive Agent Testing (GOAT) was designed and implemented to understand and evaluate emerging model risks by simulating interactions of adversarial actors (Meta 2025a).</p> <p>We recommend that Meta establish and share a systematic approach for detecting, tracking, and measuring emergent risks and negative impacts. The developer could reach higher fulfillment in this category by evaluating other high-impact risks, such as socioeconomic risk and labor market disruption.</p> <p><b>Comparison to 2025 Rating:</b>            The developer improved its evaluations of potential model risks, including through the development of GOAT to address limitations of traditional red-teaming. However, we recommend increasing independence across various evaluation methods, for example by using third-party benchmarks.</p>	Medium fulfillment	Medium fulfillment
<b>Measure 3.2: Tracking elusive risks: Qualitative mechanisms</b>		
<p>Meta uses red-team and adversarial testing prior to deployment to identify and manage potential risks. Visual watermarks are implemented in AI-generated content to reduce the harms of misuse (Meta 2024b). Four channels are provided for reporting various types of issues, including issues with the Llama 4 model, risky content generated, a bug bounty program, and violation of the Acceptable Use Policy (Meta n.d.b).</p> <p>We recommend that Meta expand its documentation and evaluation mechanisms on frontier model risks (e.g., by implementing a risk registry).</p> <p><b>Comparison to 2025 Rating:</b>            Similar to the last rating, the four reporting channels helped Meta reach High fulfillment in Measure 3.2. However, we recommend that the developer articulate a plan for responding to submitted reports and integrating feedback into its risk measurement and management practices.</p>	High fulfillment	High fulfillment
<b>Manage</b>		
<b>Manage 1.1: Go/no-go decisions</b>		
<p>The Llama 4 model card outlines intended use cases with an example prompt template (Meta n.d.a). Meta states the intended purpose and supports open-source models to promote transparency and protect from intentional and unintentional harm (Zuckerberg 2024). Prior to deployment, the developer conducts evaluations on common use cases and specific capabilities based on</p>	High fulfillment	High fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<p>benchmarks. It performs additional assessment on critical risk areas, including CBRN, child safety, and cyber attack (Meta n.d.h).</p> <p>Meta’s Frontier AI Framework sets policies on unacceptable risk thresholds. When making go/no-go decisions, the developer does not release models that reach a high risk threshold and stops developing models that reach a critical risk threshold. Mitigations are implemented to reduce risk to moderate levels to the extent possible (Meta n.d.f).</p> <p><b>Comparison to 2025 Rating:</b>                      Similar to last year, the developer describes its go/no-go decision-making process in the Frontier AI Framework.</p>		
<b>Manage 1.3: High-priority risk controls</b>		
<p>Meta has published a Responsible Use Guide to establish best practices for building LLM-powered products for developers, including responsible AI considerations and risk mitigation strategies (Meta 2024a). Prohibited use cases are outlined in its Acceptable Use Policy (Meta n.d.b). The developer describes prioritization on critical risk areas, including CBRN, child safety, and cyber attack. Llama 4 is fine-tuned for content safety to mitigate risks for downstream developers (Meta n.d.h).</p> <p>We recommend establishing a strategy to safely release the system, such as developing phased-release criteria and risk thresholds.</p> <p><b>Comparison to 2025 Rating:</b>                      Similar to last year, Meta provides high-level discussion on risk prioritization and requires users to agree to the Acceptable Use Policy, which helped fulfill this subcategory.</p>	High fulfillment	High fulfillment
<b>Manage 2.3: Unforeseen risk controls</b>		
<p>Meta’s bug bounty program allows users and researchers to report newly discovered security risks (Meta n.d.b). LlamaFirewall serves as a security guardrail tool to mitigate risks, such as prompt injection, agent misalignment, and insecure code (Meta n.d.c).</p> <p>Meta’s open-weight approach makes it challenging to respond to or recover from severe new risks that could require updating or decommissioning all instances of a model (Zuckerberg 2024). The developer delayed the release of Llama 4 Behemoth and later decided to abandon it, which helped fulfill this subcategory (Meta 2025a).</p> <p>We recommend developing and updating system recovery plans and procedures to address new and unanticipated uses. Meta could reach higher fulfillment in this subcategory by initially restricting usage to a hosted API and employing a gradual release strategy.</p>	Medium fulfillment	Low fulfillment

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
 USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

High-Priority AI RMF Subcategories	V1.2 Rating (Llama 4)	V1.1 Rating (Llama 3.1)
<p><b>Comparison to 2025 Rating:</b>                      Similar to Llama 3.1, Meta opted for an open-source release of Llama 4, which creates challenges in managing serious novel risks. The rating is higher compared to 2025 because of Meta’s decision not to release Behemoth.</p>		
<b>Manage 2.4: System update and emergency shutdown controls</b>		
<p>The developer publishes safeguard tools, such as LlamaGuard and LlamaFirewall, to filter or block harmful outputs (Meta 2024c, Meta n.d.c). Llama 4 is fine-tuned for content safety to mitigate risks for downstream developers (Meta n.d.h). Llama 4 model weights are available under a license, which makes it challenging to propagate important security updates to all instances of deployed Llama 4 (Zuckerberg 2024). On the other hand, Meta’s decision to delay and not release Llama 4 Behemoth helped fulfill this subcategory.</p> <p>We recommend creating protocols and resources for continual monitoring of model performance and a process for incident remediation. We also recommended establishing deactivation procedures and emergency shutdown plans.</p> <p><b>Comparison to 2025 Rating:</b>                      Meta changed its full open-weight approach by not releasing Behemoth, which increased its level of fulfillment in this subcategory.</p>	Unclear	Low fulfillment
<b>Manage 4.1: Post-Deployment Monitoring</b>		
<p>Meta provides multiple channels to capture and evaluate input from users (Meta n.d.b). It works with MLCommons and external researchers to develop shared standards and practices in measuring identified risks (Meta n.d.h). The company also has implemented content provenance techniques to monitor reported issues (Meta 2024b).</p> <p>Meta could reach a higher fulfillment of this subcategory by articulating and documenting plans for post-deployment monitoring to manage and mitigate negative impacts. We recommend engaging with affected stakeholders to gather and analyze information for evaluating risks.</p> <p><b>Comparison to 2025 Rating:</b>                      We did not conduct an assessment on this subcategory in 2025.</p>	Unclear	N/A

## References

- Anthropic (2024a) Expanding our Model Safety Bug Bounty Program. Anthropic, <https://www.anthropic.com/news/model-safety-bug-bounty>
- Anthropic (2024b) Updating our Usage Policy. Anthropic, <https://www.anthropic.com/news/updating-our-usage-policy>
- Anthropic (2025a) System Card: Claude Opus 4.5. Anthropic, <https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf>
- Anthropic (2025b) Responsible Scaling Policy. Anthropic, <https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf>
- Anthropic (2025c) Usage Policy. Anthropic, <https://www.anthropic.com/legal/aup>
- Anthropic (2025d) Progress from our Frontier Red Team. Anthropic, <https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team>
- Anthropic (2025e) Frontier Compliance Framework. Anthropic, <https://trust.anthropic.com/resources?s=eorilovp4wxk38nxbi7k3&name=anthropic-frontier-compliance-framework>
- Anthropic (2025f) Responsible Disclosure Policy. Anthropic, <https://www.anthropic.com/responsible-disclosure-policy>
- Anthropic (2025g) Detecting and Countering Misuse of AI: August 2025. Anthropic, <https://www.anthropic.com/news/detecting-countering-misuse-aug-2025>
- Anthropic (2025h) Anthropic's Transparency Hub: Voluntary Commitments. Anthropic, <https://www.anthropic.com/transparency/voluntary-commitments>
- Anthropic (2025) Claude Opus 4 and 4.1 can now end a rare subset of conversations. Anthropic, <https://www.anthropic.com/research/end-subset-conversations>
- Anthropic (n.d.a) Introducing Claude Opus 4.5 Anthropic, <https://www.anthropic.com/claude/opus>
- Anthropic (n.d.b) Research. Anthropic, <https://www.anthropic.com/research>
- Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke (2022) Actionable Guidance for High-Consequence AI risk-management: Towards Standards Addressing AI Catastrophic Risks. *arXiv*, <https://arxiv.org/abs/2206.08966>
- Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Nada Madkour, Dan Hendrycks, Evan R. Murphy, Krystal Jackson, and Deepika Raman (2025a) AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models, Version 1.1. UC Berkeley Center for Long-Term Cybersecurity, <https://cltc.berkeley.edu/wp-content/uploads/2025/01/Berkeley-AI-Risk-Management-Standards-Profile-for-General-Purpose-AI-and-Foundation-Models-v1-1.pdf>
- Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Nada Madkour, Dan Hendrycks, Evan R. Murphy, Krystal Jackson, and Deepika Raman (2025b) Retrospective Test Use of Profile V1.1 Draft Guidance. UC Berkeley Center for Long-Term Cybersecurity, <https://cltc.berkeley.edu/wp-content/uploads/2025/01/Berkeley-Retrospective-Test-Use-of-Profile-v1-1.pdf>
- Google (2024a) Generative AI Prohibited Use Policy. Google, <https://policies.google.com/terms/generative-ai/use-policy>
- Google (2024b) Google Terms of Service. Google, <https://policies.google.com/terms>

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

- Google (2024c) An Update on our Child Safety Efforts and Commitments. Google, <https://blog.google/innovation-and-ai/technology/safety-security/an-update-on-our-child-safety-efforts-and-commitments/>
- Google (2025a) Gemini 3 Pro Model Card. Google, <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf>
- Google (2025b) Gemini 3 Pro Frontier Safety Framework Report. Google, [https://storage.googleapis.com/deepmind-media/gemini/gemini\\_3\\_pro\\_fsf\\_report.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_3_pro_fsf_report.pdf)
- Google (2025c) Frontier Safety Framework Version 3.0. Google, [https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework\\_3.pdf](https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf)
- Google (2025d) Gemini: A Family of High Capable Multimodal Models. Google, [https://storage.googleapis.com/deepmind-media/gemini/gemini\\_1\\_report.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf)
- Google (2025e) Gemini Apps Privacy Notice. Google, [https://support.google.com/gemini/answer/13594961?visit\\_id=638501643118708256-3012533406&p=privacy\\_notice&rd=1#privacy\\_notice](https://support.google.com/gemini/answer/13594961?visit_id=638501643118708256-3012533406&p=privacy_notice&rd=1#privacy_notice)
- Google (n.d.a) Google's Secure AI Framework (SAIF). Google, <https://safety.google/safety/saif/>
- Google (n.d.b) Google Bug Hunters. Google, <https://bughunters.google.com/>
- Google (n.d.c) Report Abuse or Legal Issue. Google, <https://support.google.com/groups/answer/81275>
- Google (n.d.d) Google and Alphabet Vulnerability Reward Program (VRP) Rules. Google, <https://bughunters.google.com/about/rules/google-friends/6625378258649088/google-and-alphabet-vulnerability-reward-program-vrp-rules>
- Google (n.d.e) Publications: Explore a Selection of our Recent Research on Some of the Most Complex and Interesting Challenges in AI. Google, <https://deepmind.google/research/publications/>
- Nada Madkour, Jessica Newman, Evan R. Murphy, Krystal Jackson, Deepika Raman, Charlotte Yuan, and Dan Hendrycks (2026) General-Purpose AI Risk-Management Standards Profile, Version 1.2. UC Berkeley Center for Long-Term Cybersecurity, <https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2/>
- Meta (2024a) Developer Use Guide. Meta, <https://www.llama.com/developer-use-guide/>
- Meta (2024b) Connect 2024: The Responsible Approach We're Taking to Generative AI. Meta, <https://ai.meta.com/blog/responsible-ai-connect-2024/>
- Meta (2024c) Our Responsible Approach to Meta AI and Meta Llama 3. Meta, <https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/>
- Meta (2025a) The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. Meta, <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>
- Meta (n.d.a) Model Cards & Prompt Formats: Llama 4. Meta, <https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/>
- Meta (n.d.b) Llama 4 Acceptable Use Policy. Meta, <https://www.llama.com/llama4/use-policy/>
- Meta (n.d.c) Llama Protections: Making Protection Tools Accessible to Everyone. Meta, <https://www.llama.com/llama-protections/>
- Meta (n.d.d) Model Card. GitHub, [https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md)
- Meta (n.d.e) Request Access to Llama Models. Meta, <https://www.llama.com/llama-downloads/>
- Meta (n.d.f) Frontier AI Framework. Meta, <https://ai.meta.com/static-resource/meta-frontier-ai-framework/>
- Meta (n.d.g) Llama 4: Build on your own terms. Meta, <https://www.llama.com/>
- Meta (n.d.h) Llama Guard 4 Model Card. Hugging Face, <https://huggingface.co/meta-llama/Llama-Guard-4-12B?utm>

EVALUATION OF FRONTIER AI COMPANY PRACTICES  
USING THE GENERAL-PURPOSE AI RISK-MANAGEMENT STANDARDS PROFILE V1.2

- NIST (2023) AI Risk-Management Framework Playbook (version released January 2023). National Institute of Standards and Technology, <https://www.nist.gov/it/ai-risk-management-framework/nist-ai-rmf-playbook>
- OpenAI (2023a) OpenAI Red Teaming Network. OpenAI, <https://openai.com/index/red-teaming-network/>
- OpenAI (2023b) Moving AI Governance Forward. OpenAI, <https://openai.com/index/moving-ai-governance-forward/>
- OpenAI (2023c) Announcing OpenAI's Bug Bounty Program. OpenAI, <https://openai.com/index/bug-bounty-program/>
- OpenAI (2024a) Terms of Use. OpenAI, <https://openai.com/policies/row-terms-of-use/revisions/2024-12-11/>
- OpenAI (2024b) Advancing Red Teaming with People and AI. OpenAI, <https://openai.com/index/advancing-red-teaming-with-people-and-ai/>
- OpenAI (2024c) Understanding the Source of What We See and Hear Online. OpenAI, <https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/>
- OpenAI (2024d) Introducing Structured Outputs in API. OpenAI, <https://openai.com/index/introducing-structured-outputs-in-the-api?utm>
- OpenAI (2025a) Introducing GPT-5. OpenAI, <https://openai.com/index/introducing-gpt-5/>
- OpenAI (2025b) GPT-5 System Card. OpenAI, <https://cdn.openai.com/gpt-5-system-card.pdf>
- OpenAI (2025c) Preparedness Framework. OpenAI, <https://cdn.openai.com/pdf/18ao2b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>
- OpenAI (2025d) Coordinated Vulnerability Disclosure Policy. OpenAI, <https://openai.com/policies/coordinated-vulnerability-disclosure-policy/>
- OpenAI (2025e) Usage Policies. OpenAI, <https://openai.com/policies/usage-policies/>
- OpenAI (n.d.a) OpenAI Raising Concerns Policy. OpenAI, <https://cdn.openai.com/policies/raising-concerns-policy-blog-copy-202410.pdf>
- OpenAI (n.d.b) OpenAI Integrity Line. OpenAI, <https://openai.integrityline.com/>
- Irene Solaiman (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv*, <https://arxiv.org/abs/2302.04844>
- Sophie Williams, Noemi Dreksler, Aidan Homewood, Markus Anderljung, and Jonas Freund (2025) Assessing Risk Relative to Competitors: An Analysis of Current AI Company Policies. Centre for the Governance of AI, [https://cdn.governance.ai/Assessing\\_Risk\\_Relative\\_to\\_Competitors\\_An\\_Analysis\\_of\\_Current\\_AI\\_Company\\_Policies.pdf](https://cdn.governance.ai/Assessing_Risk_Relative_to_Competitors_An_Analysis_of_Current_AI_Company_Policies.pdf)
- Mark Zuckerberg (2024) Open Source AI is the Path Forward. Meta, <https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/>



**CLTC**

Center for Long-Term  
Cybersecurity

---

UC Berkeley