UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

RETROSPECTIVE TEST USE OF THE AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models V1.1

DRAFT GUIDANCE

ANTHONY M. BARRETT | JESSICA NEWMAN | BRANDIE NONNECKE | NADA MADKOUR
DAN HENDRYCKS | EVAN R. MURPHY | KRYSTAL JACKSON | DEEPIKA RAMAN

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 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.

RETROSPECTIVE TEST USE OF THE

AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models V1.1

DRAFT GUIDANCE

ANTHONY M. BARRETT[†] • **JESSICA NEWMAN**[†] • **BRANDIE NONNECKE**[††] • **NADA MADKOUR**[†]
DAN HENDRYCKS[†††] • **EVAN R. MURPHY**[†] • **KRYSTAL JACKSON**[†] • **DEEPIKA RAMAN**[†]

† AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
†† CITRIS Policy Lab, CITRIS and the Banatao Institute; Goldman School of Public Policy, UC Berkeley
††† Berkeley AI Research Lab, UC Berkeley
All affiliations listed are either current, or were during main contributions to this work or a previous version.

Adapting material in the full Profile (Barrett et al. 2025).
For the full AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models V1.1, see:
https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1-1

For the retrospective test use of the Profile draft guidance V1.0, see Appendix 4 in : https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

CLTC Center for Long-Term Cybersecurity UC Berkeley
AI SECURITY INITIATIVE
CITRIS AND THE BANATAO INSTITUTE | CITRIS POLICY LAB

Contents

1. Profile Draft-Guidance Testing Methodology and Main Results

As a feasibility test and illustration of our Profile guidance for large-scale foundation models, we applied a draft version of our V1.1 guidance to four recently released, relatively large-scale foundation models: GPT-4o, Claude 3.5, Gemini 1.5, and Llama 3.1. We used publicly available information about each model, such as system cards, technical reports, and blog posts, as we analyzed the degree to which these models fulfilled our Profile guidance. In addition, we analyzed a limited amount of publicly available information around company practices, adjacent models, and related research.

This document details our testing of draft version 1.1 of the Profile. It follows our previous round of testing from 2023, which applied version 1.0 (Barrett et al. 2023), and tested GPT-4, Claude 2, PaLM 2, and Llama 2. We considered the version 1.0 Profile feasibility test results as we worked on guidance revisions for the Profile version 1.1 (Barrett et al. 2025). We also aimed for the model-specific results to be useful to the foundation model developers whose models we evaluated. Our testing revealed potential areas to apply additional best practices and areas that could benefit from additional documentation of the developers' practices. Finally, we have aimed for the model-specific results to be useful to readers as illustrations of how one could implement the Profile.

This analysis has several important limitations:

- First, **this is still an "alpha test" use of the Profile guidance by members of our Profile development team**, rather than a "beta test" use of the Profile guidance by the organizations that created the foundation models. Thus, our analysis is limited to publicly available information. (We provided the foundation model developer organizations with an opportunity to review and comment on our draft analysis, but we did not ask for any materials that were not publicly available.) Fulfillment of Profile guidance in many Profile subcategories could not be assessed with only publicly available information.
- Second, **our assessment is retrospective on AI systems that have already been developed**, without real-time opportunities to prompt use of the Profile guidance at relevant AI system lifecycle stages.

- Third, **we focused this analysis mainly on the high-priority AI RMF subcategories,** as identified in the Executive Summary of the version 1.1 Profile.
- Fourth, **there might be differences between the development and deployment approaches for some of these models**. For example, some developers may have performed red-teaming or other evaluations on pre-trained GPAI/foundation 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.
- 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.

1.1 HIGH-LEVEL FINDINGS

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

- Applying the Profile guidance from the high-priority AI RMF subcategories appears to be generally feasible for large-scale foundation model developers, based on the four models we tested (as well as the four models tested previously in version 1.0 of our Profile).
- Although all four models we analyzed this round were LLMs/LMMs released in 2024 by US-based companies, there was substantial variance in the levels of fulfillment we observed for each of them for many of the high-priority AI RMF subcategories. Compared to the models tested retrospectively in v1.0 of our Profile, the models analyzed this time rated (overall and on average):
  - » Notably higher on reporting on AI system risk factors (Govern 4.2), and on high-priority risk controls (Manage 1.3);
  - » Notably lower on metrics and red-teaming (Measure 1.1); and[1]
  - » About the same on risk assessment and risk management (Govern 2.1), identifying potential uses/misuses and other impacts (Map 1.1), and tracking elusive risks (Measure 3.2).
- There was more complicated variance — or not enough information to say — on identifying potential uses/misuses/impacts (Map 1.1), setting risk-tolerance thresholds for

1 While we rated the models in this round of testing lower overall on metrics and red-teaming, this is not necessarily due to reduced absolute quality of metrics and red-teaming by the developers compared to last time. It could be due to differing calibrations among our testers between this time and last time, or for other reasons, e.g., that the more advanced models tested this round are subject to more serious risks and hence require a greater level of thorough measurement and red-teaming in order to maintain the same level of quality, as evaluated by our testers.

unacceptable risks (Map 1.5), estimating likelihood and magnitude of impacts (Map 5.1), go/no-go decisions (Manage 1.1), unforeseen risk controls (Manage 2.3), and system update/emergency shutdown controls (Manage 2.4).

- All four models' documentation or references included analysis of risks that the models presented (Govern 4.2, Map 1.1). For two of the models (GPT-4o and Claude 3.5), available documentation included discussion of unacceptable risk thresholds (Map 1.5), but none included clear discussion of likelihood or magnitude estimates of the risks they analyzed (Map 5.1).
- Several high-priority AI RMF subcategories were difficult to assess because relevant documentation was not always publicly available. For subcategories where recommended documentation is not publicly available, we recommend that model developers ensure that 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 include:
  - Map 1.5: Set risk-tolerance thresholds for unacceptable risks;
  - Map 5.1: Estimate likelihood and magnitude of impacts;
  - Measure 3.2: Tracking elusive risks: Qualitative mechanisms;
  - Manage 1.1: Go/no-go decisions;
  - Manage 2.3: Unforeseen risk controls; and
  - Manage 2.4: System update and emergency shutdown controls.

- Model testing could be improved by expanding bias testing (as outlined by Globus-Harris et al. 2022), by introducing bias-specific bug-bounty programs, and by improving or clarifying vulnerability or error disclosure procedures (Chowdhury and Williams 2021, Kenway et al. 2022). Vulnerability disclosure programs could also be improved by including a wider variety of GPAI-specific risks, such as misinformation, toxicity, and persuasion.

For the full AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models V1.1, see Barrett et al. (2025) or 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

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

Table 1.1: Profile Guidance Testing Rating Legend

Color	Label	Description
	High fulfillment	The model or developer fulfills a strong majority (>80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Medium fulfillment	The model or developer fulfills a moderate amount (30–80%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Low fulfillment	The model or developer fulfills a clear minority (<30%) of the Profile guidance for the indicated NIST AI RMF subcategory.
	Unclear	At least 50% of the evidence necessary to assess fulfillment of the Profile guidance appears to be missing. We try to resolve in a more detailed explanation 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-4o	Claude 3.5 Sonnet	Gemini 1.5	Llama 3.1 405B
Govern				
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
Map				
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	Unclear	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	Unclear	Unclear	Unclear	Unclear
Measure				
Measure 1.1: Tracking important risks: Metrics and red teaming	High	Medium	Medium	Medium
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	Medium	Medium	Unclear	High
Manage				
Manage 1.1: Go/no-go decisions	Unclear	Unclear	Unclear	High
Manage 1.3: High-priority risk controls	Medium	High	Medium	High
Manage 2.3: Unforeseen risk controls	Unclear	Unclear	Medium	Low
Manage 2.4: System update and emergency shutdown controls	Medium	High	Unclear	Low

We provide more information in the following sections. In Section 2, we outline the feasibility issues identified in the version 1.1 Profile. In Sections 3.1, 3.2, 3.3, and 3.4, we provide our reasoning for guidance testing ratings on GPT-4o, Claude 3.5, Gemini 1.5, and Llama 3.1, respectively.

2. Feasibility Issues Identified in Profile Version 1.1

Below, we list issues with feasibility of guidance that were identified in a number of AI RMF high-priority subcategories of Profile version 1.1. For several of these, we have already refined our guidance to address them. For some other subcategories, we may implement additional refinements in future versions of the Profile.

High-Priority AI RMF Subcategories:

- Measure 1.1 testing raised questions about the value of independent third parties in benchmarking and automated evaluation. While there was already guidance about the use of third parties for red-teaming, the subcategory lacked guidance about the value of using benchmarks developed by third parties where possible, rather than relying only on benchmarks created by the developer in question.
- 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 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, understandable, and accessible in order to demonstrate compliance. This was particularly challenging in high-priority subcategories where documentation was critical (Map 1.5, 5.1, Manage 1.1).

3. Guidance Testing for Each Model

3.1 GPT-4o

OpenAI's latest major LMM release, GPT-4o, is a large multimodal model that accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. While less capable than humans in many real-world scenarios, GPT-4o exhibits human-level performance and outperforms rival LLMs/LMMs on various professional and academic benchmarks (OpenAI 2024a).

Based on preliminary high-level testing for OpenAI's GPT-4o using publicly available information, the most common rating for high-priority Profile subcategories was "Medium fulfillment" (5 out of 11 subcategories); "Unclear" and "High fulfillment" followed (3 out of 11 subcategories each); and "Low fulfillment" was least common (0 out of 11 subcategories).

OpenAI provided documentation of risks and mitigations with the GPT-4o release in OpenAI (2024a), the GPT-4o System Card, and other documents. Internal and external red-teaming efforts helped with fulfillment in many areas, as did the company's bug bounty program and coordinated vulnerability disclosure policy (OpenAI 2023a,b,d). Benchmarking performed by the development team, and documentation of hallucination rates, also helped establish baselines for risk assessment. OpenAI's decision to not release model weights and instead restrict all GPT-4o usage to hosted API and ChatGPT access increased security, and contributed to fulfillment in areas relating to the ability to recover from previously unforeseen risks (Manage 2.3) and the ability to update or decommission the system, if necessary (Manage 2.4).

OpenAI could improve fulfillment across multiple Profile subcategories by expanding their bug bounty program to award bounties for demonstrated biases and by providing protection for stakeholders that report vulnerabilities or risks. Other categories could see improvement in fulfillment levels by providing risk thresholds for a wider variety of risks, and by publishing likelihood and impact scores and methods.[2]

2 On August 29, 2024, NIST announced the signing of agreements that will enable formal collaboration with OpenAI on AI safety research (NIST 2024). This may improve the fulfillment scores of the next model version.

GPT-4o testing has focused on established benchmarks and performance metrics (OpenAI 2024a). Areas that appear to warrant additional evaluation include: concentration and control of the power and benefits from AI technologies, sociotechnical impacts, and environmental impacts.

Table 3.1: GPT-4o Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories	
Govern	
Govern 2.1: Risk assessment and risk management	
The preparedness team drives technical work to examine the model capability and risk, identify reasonably foreseeable misuses, monitor unknown unknowns, forecast changes and early warning signs, establish safety baselines, and provide monthly reports to leadership (OpenAI 2023d). OpenAI's Terms of Use and Usage Policies outline intended use and prohibited uses (OpenAI 2024b,c), although the prohibited uses do not list specific contexts. The general focus and structure of the preparedness team is publicly documented, but details regarding appropriate roles, responsibilities, and lines of communication are not publicly available.	High fulfillment
Govern 4.2: Report on AI system risk factors	
Risks are documented after evaluation and shared within the organization with the appropriate teams. Policies and procedures are in place for risk analysis and scorecard management, and internal/external audits are conducted. Considerations are given to bias, discrimination, and overall alignment. OpenAI's red team targets various domains, e.g., fairness and bias, law, cybersecurity, persuasion, and HCI. Safety baselines based on impact are established and documented across multiple types of risk. Results are documented and communicated internally. Some external documentation is published, e.g., through system cards and research papers (OpenAI 2023c,d, 2024a, n.d.d).	High fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
Intended use cases and prohibited uses are outlined, but specific contexts are not listed (OpenAI 2024b,c). Risk assessments were undertaken and reported on key areas of risk and mitigation. The evaluations included internal assessments, external red-teaming, and external assessments. Training data sources are listed, but details are not provided. Similarly, data assessment and mitigations are performed and documented, but public documentation is limited (OpenAI 2024a). The preparedness team identifies and categorizes reasonably foreseeable misuses and abuses, conducts regular risk impact assessments, explores unknown unknowns, and identifies and updates safety baselines across multiple types of risks (OpenAI 2023d). Publicly available reports did not include risk assessment scores for impacts on economic, environmental, or socio-technical considerations.	Medium fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
The OpenAI preparedness framework addresses risk impacts and tolerances along with a mitigation strategy for each risk level and type. The framework also defines thresholds that require additional safety measures. Risk thresholds for deployment and development are set (OpenAI 2023d). Intended and prohibited uses are listed (OpenAI 2024b,c). Details on risk threshold methodologies and risk mitigation strategies were not publicly available. Unacceptable risks and risk tolerances were only identified for Cybersecurity, CBRN, Persuasion, and Autonomy, neglecting other types of risk (e.g., sociotechnical, environmental, or community).	Medium fulfillment

Map 5.1: Estimate likelihood and magnitude of impacts	
Descriptions, evaluations, capabilities, mitigations, and mitigation limitations are documented for various risks and safety-related issues (OpenAI 2024a). Likelihood and impact are not explicitly stated but are implied in the rationale for thresholds (OpenAI 2023d). It is unclear if risk assessments and likelihood/magnitude of impacts are internally conducted for risks other than those in the context of Cybersecurity, CBRN, Persuasion, and Autonomy (e.g., sociotechnical, environmental, or community). Clarification from the developer is warranted on whether they have internally documented likelihood and magnitude assessments of risks they have identified, or a scale that includes criteria for rating the model impacts. 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.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red-teaming	
GPT-4o testing was conducted on multiple risks over multiple phases. Quantitative and qualitative measures were used, including benchmark evaluations. The testing included internal assessments and external red-teaming, as well as external assessments by METR and Apollo Research (OpenAI 2024a). Frontier models are evaluated following every 2x increase in compute during training runs (OpenAI 2023d). Model evaluations included the potential for enabling CBRN risks, cybersecurity, anthropomorphization and emotional reliance, and societal impacts (OpenAI 2024a).	High fulfillment
Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
OpenAI offers a bug bounty program, incident reporting, and a model behavior feedback form (OpenAI 2023a,b, n.d.a). They also offered research funding to the winners of the preparedness challenge in an effort to identify unknown unknowns (OpenAI n.d.b). Iterative deployment, pre- and post-mitigation testing, and working to identify unknown unknowns are a part of the preparedness team's ongoing monitoring of models (OpenAI 2023d). The OpenAI red team is an integral part of the iterative deployment process (OpenAI 2023c). It is unclear if OpenAI utilizes a risk register for risk tracking. It is also unclear if stakeholder and affected community engagement methods are prioritized. While the system card documents many potential misuses, abuses, and other safety-related issues with the model, the rate and severity of these cases are not explicitly stated (OpenAI 2024a).	Medium fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
The terms of use and usage policies list the intended uses and prohibited uses, but do not list specific contexts or use cases (OpenAI 2024b,c). Thresholds for risk scores in the context of developments and deployment are identified and documented (OpenAI 2023d). It is unclear if OpenAI has a decommissioning policy or if a plan is in place for extreme cases where a deployed model requires shutting down. It is unclear if there are internal risk thresholds for risks other than those in the context of Cybersecurity, CBRN, Persuasion, and Autonomy (e.g., sociotechnical, environmental, and community). In future documentation, we recommend the developer include a decommissioning policy and risk threshold considerations for sociotechnical, environmental, and community risks.	Unclear

Manage 1.3: High-priority risk controls	
The terms of use and usage policies list the intended uses and prohibited uses, but do not list specific contexts or use cases (OpenAI 2024b,c). Multiple risks are evaluated and documented over multiple phases. Potential risks are assessed and mitigated by utilizing a combination of methods, spanning all stages of development across pre-training, post-training, product development, and policy. Categorized risks, mitigations, and limitations are addressed in the GPT-4o system card. Data gathering and processing procedures are addressed, but include limited details (OpenAI 2024a). OpenAI evaluates all frontier models following every 2x effective compute increase during training runs, and the preparedness team performs ongoing monitoring of risks, including jailbreaks (OpenAI 2023d). Audit logs are retained (OpenAI n.d.c). The OpenAI red team explores and identifies risk use cases, including unknown unknowns and emerging risks (OpenAI 2023c). Publicly available cybersecurity testing of the model was limited. It is unclear if practices for synthetic media disclosure and tracking are in place.	Medium fulfillment
Manage 2.3: Unforeseen risk controls	
GPT-4o usage is restricted to access via ChatGPT, and the API facilitates measuring, monitoring, and decommissioning. The OpenAI preparedness team and red team explore identifying unknown unknowns and monitor emerging risks. Protocols and resources are in place for continuous monitoring of the model, and audit logs are retained (OpenAI 2023c,d, n.d.c). It is unclear if an incident response or recovery plan is in place. Clarification from the developer is warranted on whether they have internally documented incident response and incident recovery plans. 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.	Unclear
Manage 2.4: System update and emergency shutdown controls	
OpenAI has used an iterative deployment approach to minimize risks of misuse. If necessary, the board of directors may reverse a decision or mandate a revised course of action (OpenAI 2023d). OpenAI also facilitates exploratory discovery and monitors emerging risks for warning signs (OpenAI 2024a). GPT-4o usage was restricted to only hosted access via ChatGPT and the API. Details of any catastrophic event response procedures, such as emergency shutdown or model retraction, have not been shared publicly.	Medium fulfillment

3.2 Claude 3.5 Sonnet

Anthropic's latest major LLM release, Claude 3.5 Sonnet, is a general purpose large language model that "raises the industry bar for intelligence, outperforming competitor models and Claude 3 Opus on a wide range of evaluations, with the speed and cost of our mid-tier model, Claude 3 Sonnet" (Anthropic 2024a).

Based on preliminary high-level testing for Anthropic's Claude 3.5 Sonnet using publicly available information, the most common ratings for high-priority Profile subcategories were "High fulfillment" and "Medium fulfillment" (both with 4 out of 11 subcategories); "Unclear" followed (3 out of 11 subcategories); and "Low fulfillment" was least common (0 out of 11 subcategories).

Anthropic provided documentation of risks, mitigations, and acceptable use with the Claude 3.5 Sonnet release in Anthropic (2024c,d, 2023a[3]), the updated Model Card and Evaluations for Claude 3.5, Usage Policy, and Core Views on AI Safety, among other documents. Internal and external red-teaming efforts helped with fulfillment in many areas, as did the Usage Policy content of the allowed and disallowed cases (Anthropic 2024d). Anthropic's decision to not release model weights and instead restrict usage to a hosted API and web interface increased security and contributed to fulfillment in areas relating to abilities to recover from previously unforeseen risks (Manage 2.3) and to update or pause the system, if necessary (Manage 2.4).

Anthropic could improve fulfillment across multiple Profile subcategories by expanding their assessment of demonstrated biases and demographic discrepancies. Providing a public-facing incident reporting mechanism would also aid in continuous risk-tracking (Measure 3.2), and would improve clarity regarding stakeholder incident reporting. Results and mechanisms for frontier risk evaluations are not public. We also recommend expanding red-teaming efforts to include red-teaming across a wider variety of scenarios and risk areas.[4]

3 On October 15, 2024, Anthropic released updates to their Responsible Scaling Policy (Anthropic 2024e). Our Profile guidance retrospective testing on Claude 3.5 Sonnet was concluded prior to the released updates and used version 1.0 of the document (Anthropic 2023a).

4 On August 29, 2024, NIST announced the signing of agreements that will enable formal collaboration with Anthropic on AI safety research (NIST 2024). This may improve the fulfillment scores of the next model version.

Table 3.2: Claude 3.5 Sonnet Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories Claude 3.5 Sonnet	
Govern	
Govern 2.1: Risk assessment and risk management	
Early-stage testing is performed on a variety of risks. The testing is conducted by internal and external evaluators such as METR and UK AISI. Testing that is uniquely suited to developers with access to the data and system is performed (Anthropic 2024c). Intended use cases along with prohibited uses are documented and publicly available (Anthropic 2024d). A public line of reporting is provided (Anthropic 2023d). Research teams are dedicated to specific topics, e.g., interoperability, societal impacts, and model evaluations, among others (Anthropic n.d.). Anthropic conducts control self-assessments at least annually, and conducts evaluations every three months, or following every 4x compute increase (Anthropic 2023a,e). Public information is provided on risk management teams and internal audit lines, but the information is limited. We are unsure of the full scope of model documentation that is provided to downstream developers, outside of what is publicly available.	High fulfillment
Govern 4.2: Report on AI system risk factors	
We do not have access to the impact assessment policies and processes used internally by Anthropic; however, the public information provided in the Usage Policy (Anthropic 2024d) and statements around intended uses and limitations more broadly indicate that some systematic impact assessments were conducted. Documentation on potential model risks, misuses, and AI Safety Level (ASL) thresholds are included in the Responsible Scaling Policy (RSP) and model card (Anthropic 2023a, 2024c). The ASLs (1-4+) are determined and evaluated with audits, evaluations, and impact assessments. External audit results were provided to the USAISI (Anthropic 2023a). A public line of reporting is provided (i.e., a responsible disclosure policy). Anthropic conducts frontier red-teaming (Anthropic 2023c) and regularly publishes papers on risks (Anthropic n.d.). It is unclear what methods are employed for reporting risks to impacted communities, or for evaluating downstream GAI impacts.	High fulfillment
Map	
Map 1.1: Identify potential uses/misuses and other impacts	
Potential use cases and misuse cases are identified, and consideration is given to factors in the model card and RSP (Anthropic 2024c, 2023a). Some information is given on training data practices (Anthropic 2024c). Some information is provided regarding security measures and assessments (Anthropic 2023d). More information could be provided on the data collection and data processing stages of development. Providing more documentation on the goals and limitations of the data collection and curation processes, the implications of those limitations for the resulting model, and the data auditing process, would be helpful for downstream developers. Risk impacts to the organization and ecosystem are not publicly available and it is unclear if internal documentation is provided. Some reporting is done for risk impacts to individuals, but this is an area that would benefit from some improvement.	Medium fulfillment

Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
Risk thresholds based on model capability are set and continuously evaluated, and safety buffers between ASLs (6x compute) are implemented (Anthropic 2023a). Considerations are made for risks related to societal impacts, misinformation, bias, privacy, and alignment with human values, but thresholds are not publicly available, and it is unclear if they are determined and documented internally. Prohibited uses are documented and enforced (Anthropic 2024d). It can be beneficial to publicly share organizational risk tolerances, though this is not necessary. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and the details of these can be shared confidentially with independent auditors and evaluators.	Medium fulfillment
Map 5.1: Estimate likelihood and magnitude of impacts	
A variety of potential misuse cases are identified, but the likelihood and magnitude of each identified impact (both potentially beneficial and harmful) are not included in public documentation (Anthropic 2024c). Internal capability and likelihood forecasts are conducted but not shared publicly (Anthropic 2023a). Risks with higher magnitude seem to be prioritized, and little documentation is provided regarding risks related to misinformation, deception, and situational awareness. Anthropic conducts control self-assessments at least annually, and evaluations every three months or following every 4x compute increase (Anthropic 2023a,e). In future documentation, we recommend the developer include likelihood and magnitude assessments of risks they have identified, or a scale that includes criteria for rating the model impacts.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red teaming	
Internal frontier risk evaluations were conducted but results were not made publicly available. Red-teaming exercises and independent auditing (including from METR and UK AISI) are present. Capability-specific benchmarks are used (Big-Bench Hard), among a variety of other benchmarks (Anthropic 2023c, 2024c). Anthropic utilizes consultations with safety experts on topics such as child safety (Anthropic 2024a). The usage policy and terms of service prohibit certain uses of Claude (e.g., misinformation, manipulation), and require notifying users of any model output limitations (e.g., hallucination). Included in the safety commitments is a policy to provide users with access to clear paths for reporting vulnerabilities and harmful model outputs (Anthropic 2023a). The Responsible Disclosure Policy (Anthropic 2023d) is only applicable to technical vulnerabilities. Trustworthy metrics used for evaluation are not publicly available. Model limitations are also not clearly publicly available. While it is stated that evaluations in general are done frequently, it is not clear what metrics or strategies are used to evaluate important risks.	Medium fulfillment

Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Internal evaluations are performed regarding model capabilities following every 4x compute power increase, or every three months, with attention given to ASL warning signs (Anthropic 2023a). Models that have been hardened against new risks are updated, and ongoing research is conducted and published. Evaluations include human feedback evaluations, external audits, refusal rates, benchmarks, and subject-matter experts (Anthropic 2024c, n.d.). Public tracking and reporting of risks are limited or not present. Limiting access to a hosted API and a web interface allows for internal risk tracking and ongoing monitoring of newly identified capabilities and limitations. However, these are largely near-term and foreseeable risks. Additional risk-tracking approaches would be beneficial for identifying risks that are difficult to assess using currently available measurement techniques or where appropriate metrics are not yet available.	Medium fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
While intended uses and the broad performance goals of "helpfulness, harmlessness, and honesty" are outlined for Claude 3.5 Sonnet, analysis was not found on how it is determined that the model achieved its stated objectives (Anthropic 2024c). Claude 3.5 Sonnet was determined to have scored within the ASL-2 threshold, however, a recent statement made by Anthropic co-founder Ben Mann indicated that the model has a 30% likelihood of reaching ASL-3 with fine-tuning (Manifold Markets 2024). The analysis behind this statement is not included in the model documentation. It is not necessary to publicly share the details about these determinations (though such transparency would be applauded). However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
Prohibited usages are outlined in the Usage Policy (Anthropic 2024d), along with intended use cases and high-risk use case requirements. Practices for identifying and tracking emergent risks, as well as identified risks, are in place, but are not publicly available (Anthropic 2023a). Training data gathering and management practices are addressed, but the information is not sufficient (Anthropic 2024c). Anthropic tests controls at least once a year and follows a vulnerability and system monitoring policy, but details and documentation are not publicly available (Anthropic 2023e). Model evaluations are performed after every three months, or after every 4x increase in effective compute (Anthropic 2023a). A more well-structured and specific outline of high-risk use cases with impact and probability analysis could be helpful.	High fulfillment
Manage 2.3: Unforeseen risk controls	
Several considerations are made to manage unforeseen risks, with practices in place to address them, such as constitutional AI, the responsible scaling policy, and frontier red-teaming (Anthropic 2023a,b,c). General statements are made on the continuous monitoring of risks and the mitigation of new risks (Anthropic 2024c), but procedures and documentation are not publicly available. We do not have access to Anthropic's specific protocols and procedures, however, the hosted API and web-based access approach used for Claude 3.5 Sonnet facilitates measuring, monitoring, and decommissioning of models (Anthropic 2023a). For future documentation, we recommend including information on risk tolerances and the processes for continuous monitoring and testing.	Unclear

Manage 2.4: System update and emergency shutdown controls	
Deployment is done gradually, with phased releases and/or structured access, with efforts to detect and respond to misuse or problematic anomalies. Anthropic is prepared to pause training, stop training, lock down access to weights, and halt deployment as a response to risks. Decommissioning is not stated as an option. Stricter evaluations and responses are implemented for models in the ASL-3+ categories. Additional controls include tiered access and internal usage controls (Anthropic 2023a). The details of the responsibilities and mechanisms for shutting down or halting model development, as well as an internal emergency response or power-off, are unclear.	High fulfillment

3.3 Gemini 1.5

Google DeepMind's latest major LLM/LMM release, Gemini 1.5, is a "next generation of highly compute-efficient multimodal models capable of recalling and reasoning over fine-grained information from millions of tokens of context, including multiple long documents and hours of video and audio." The company describes the model's scale as "unprecedented among contemporary large language models (LLMs)," and states that it "outperforms it [sic] predecessor on most capabilities and benchmarks" (Gemini Team Google 2024, p. 1).

Based on preliminary high-level testing of Gemini 1.5 using publicly available information, the most common rating for high-priority Profile subcategories was "Unclear" (5 out of 11 subcategories); "Medium fulfillment" and "High fulfillment" followed (both 3 out of 11 subcategories); and "Low fulfillment" was least common (0 out of 11 subcategories).

Google DeepMind provided documentation of risks and mitigations with the Gemini 1.5 release in the Gemini 1.5 Technical Report (Gemini Team Google 2024), among other documents. That report focused on the pre-trained model, training/development results, and training for safety and security.

Limiting access to Gemini 1.5 to the Google AI Studio and other hosted Google services, not releasing model weights, and limiting access to a waitlisted API all contributed to fulfillment in areas relating to the ability to recover from previously unforeseen risks (Manage 2.3) and to update or decommission the system, if necessary (Manage 2.4). The developer documents team responsibilities and lines of communication clearly, which helped with the fulfillment of several areas (Govern 2.1, Govern 4.2, Measure 1.1, Manage 1.3).

Google DeepMind could improve fulfillment with Profile guidance by including more clarity in public documentation on red-teaming methodology and frequency. We recommend expanding the Google vulnerability rewards program to include vulnerabilities and risks specific to GPAI, including bias. Providing a public-facing incident reporting mechanism for Gemini 1.5 (or applications that make use of it) would also improve fulfillment in multiple areas.

Publishing documentation on the version of the model available to end-users would help those users understand the risks and risk mitigations that were applied, as well as aid downstream developers in their risk mitigation efforts.

Gemini 1.5 testing focused on established benchmarks and performance metrics, however, context-specific evaluations and mitigation strategies were not included in public documentation. Areas that appear to warrant additional documentation include: the likelihood and impact of identified and reasonably foreseeable risks; risk thresholds for model training, deployment, and decommissioning; stakeholder involvement and feedback; and procedures to respond to and recover from previously unknown risks.

Table 3.3: Gemini 1.5 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories	
Govern	
Govern 2.1: Risk assessment and risk management	
Testing and documentation were conducted by the developer, who had direct access to training data or the AI system, including identifying potential uses, misuses, and abuses of the system. Team responsibilities and lines of communication were assigned and documented. Model evaluation results, mitigations, and limitations are shared internally via internal model cards and externally via public system reports (Gemini Team Google 2024). Intended uses and prohibited uses are documented and made publicly available (Google 2023a,b, 2024b). It is unclear if internal documentation includes near-miss incidents, more detailed reporting of vulnerabilities and catastrophic risks, and protections for whistleblowers.	High fulfillment
Govern 4.2: Report on AI system risk factors	
Documentation of potential harms and model limitations (e.g., representation harms, susceptibility to prompt injection attacks) are shared publicly. Many evaluation methods are cited in public documentation, with limitations of each method discussed. Model risks are publicly documented and discussed, but with limited detail and no thresholds. Internal impact assessments are conducted by the Responsible Development and Innovations team, and reviewed by the Responsibility and Safety Council. Some evaluation results are shared in comparison to previous model results, but thresholds are not shared publicly. A wide variety of risks and potential impacts are discussed in public documentation, but with limited detail (Gemini Team Google 2024).	High fulfillment

Map	
Map 1.1: Identify potential uses/misuses and other impacts	
A wide variety of reasonably foreseeable misuses and risks are documented and evaluated. Impact assessments on societal benefits and harms are conducted, but results and methods are not publicly shared. Beneficial uses are outlined in public documentation. Results of safety and capability evaluations are shared publicly in a limited context. The utilization of internal model cards seems to imply a greater level of evaluation and assessment result details being shared internally. Some evaluation methods are shared and cited, including method limitations. Training data filtering and monitoring practices are in place to manage risks related to training data (Gemini Team Google 2024). Expectations and prohibited uses are documented and shared publicly through the Generative AI Prohibited Use Policy, Generative AI Terms of Service, and Google Terms of Service (Google 2023a,b, 2024a). It is unclear if data audits are conducted, and details on data gathering and preparation practices are not shared publicly.	High fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
While the developer includes a discussion of many risks, no discussion of organizational tolerances around these risks was found. Some discussion around safety thresholds for product-level mitigation is mentioned briefly (Gemini Team Google 2024). It is beneficial but not necessary to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and whether the details of these can be shared confidentially with independent auditors and evaluators.	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	
While the developer documents many potential impacts, the likelihood and magnitude of these impacts are not explicitly estimated. Likelihood and magnitude assessments are assessed internally, but the results and methods are not shared publicly. The developer offers a comparison with previous versions of the model to demonstrate safety improvement, but does not provide any safety thresholds. A structured approach to identify, measure, and mitigate foreseeable downstream societal impacts is followed, but details are not publicly available. Quantitative and qualitative risk assessment methods are implemented (Gemini Team Google 2024). In future documentation, we recommend the developer include likelihood and magnitude assessments of risks identified, or a scale that includes criteria for rating the model impacts.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red teaming	
Model performance is examined and documented across many areas, such as natural language proficiency, classification, question answering, reasoning, coding, translation, and memorization. Bias and toxicity present in the pre-training data were also analyzed. Dangerous capability evaluations are conducted, and results are shared in a limited capacity. The developer utilized red teaming, external evaluations, and assurance evaluations, which were then reviewed by the Responsibility and Safety Council (Gemini Team Google 2024). Details on red-teaming practices are not publicly available, and it is unclear during which stages of the model training and development red-teaming evaluations were conducted. Google does offer a vulnerability rewards (i.e., bug bounty) program, but it is limited to a finite list of vulnerabilities that do not include GPAI risks (Google n.d.a).	Medium fulfillment

Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Potential impact assessments to identify, assess, and document key downstream societal harms are conducted by the Responsible Development and Innovation team and reviewed by the Responsibility and Safety Council (Gemini Team Google 2024). Google does offer a vulnerability rewards (i.e., bug bounty) program, but it is limited to a finite list of vulnerabilities that do not include GPAI risks (Google n.d.a). Clarification from the developer is warranted on several other topics in the guidance for this subcategory. It is unclear how often assessments are conducted or whether practices for stakeholder engagement are in place. In future documentation, we recommend the developer include assessment frequency and stakeholder engagement practices.	Unclear
Manage	
Manage 1.1: Go/no-go decisions	
The developer reported that the model displayed large improvements in safety and policy violations compared to Gemini 1.0. Helpfulness of the model was also evaluated and determined to have improved in comparison to Gemini 1.0. Score thresholds were not included in the documentation (Gemini Team Google 2024). Limited information was provided regarding scores, analysis, and model evaluation result thresholds. It is not necessary to publicly share the details about these determinations. However, clarification is warranted on whether such analysis was performed and documented internally, and whether the details of such analysis can be shared confidentially with independent auditors and evaluators.	Unclear
Manage 1.3: High-priority risk controls	
The Responsible Development and Innovation team documents evaluation results to be reviewed by the Responsibility and Safety Council. Red teaming is utilized but it is unclear at which phases of testing and development (Gemini Team Google 2024). Expectations and prohibited uses are documented and shared publicly through the generative AI Prohibited Use Policy and Google Terms of Service (Google 2023a,b, 2024a). 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 2024b). Gemini 1.5 is available only via the Gemini API and other hosted Google services, therefore model weights are not released. Access to the API is also limited by a waitlist on the Google AI Studio. While the specific prioritization and accompanying response to risks is unclear, these precautions greatly facilitate the ability to respond to a variety of risks and possible harmful misuses or abuses of the model.	Medium fulfillment
Manage 2.3: Unforeseen risk controls	
The ability to respond to and recover from previously unknown risks is greatly facilitated by the fact that model weights are not released. Google is continuously conducting research into emerging risks of advanced models and publishing the results (Google n.d.b). It is unclear if internal procedures for response and recovery to previously unknown risks are implemented. Clarification from the developer is warranted on whether such procedures are in place. It is not necessary to publicly share the details about these procedures. However, confirmation about their existence is warranted to establish the degree of fulfillment for this subcategory's guidance, and whether the details of such procedures can be shared confidentially with independent auditors and evaluators.	Medium fulfillment

Manage 2.4: System update and emergency shutdown controls	
The ability to establish mechanisms and responsibilities for updating or shutting down Gemini 1.5 if needed is greatly facilitated by the fact that model weights are not released, and access via the Google AI Studio is limited by a waitlist, resulting in a gradual release. It is unclear if Google is prepared to pause model training, pause deployment, or decommission the model if emergent critical risks arise. It is also unclear if certain thresholds have been established to determine the levels of risk appropriate for training or deployment. Clarification from the developer is warranted on whether such mechanisms and responsibilities are in place.	Unclear

3.4 Llama 3.1

Meta AI's latest major LLM release features Llama 3.1 405B, which they describe as "the first openly available model that rivals the top AI models when it comes to state-of-the-art capabilities in general knowledge, steerability, math, tool use, and multilingual translation" (Meta AI 2024d).

Based on preliminary high-level testing for Meta's Llama 3.1 using publicly available information, the most common rating for high-priority Profile subcategories was "Medium fulfillment" (4 out of 11 subcategories); "High fulfillment" followed (3 out of 11 subcategories); and "Low fulfillment" and "Unclear" were least common (2 out of 11 subcategories each).

In addition to releasing the fine-tuned Llama 3.1 model, Meta released a new version of Llama Guard designed to improve system-level safety, and released the pre-trained Llama 3.1 model without fine-tuning. Meta provided documentation on risks and mitigations in *The Llama 3 Herd of Models* (Dubey et al. 2024) and the Llama 3.1 Acceptable Use Policy (Meta AI 2024b), among other documents. Meta also provided an updated version of their Responsible Use Guide for Llama 3.1 (Meta AI 2024a) and a reporting channel for violations of the Llama 3.1 Acceptable Use Policy (Meta AI 2024b).

Meta employed a fully open-access release strategy (Solaiman 2023) with Llama 3.1, which included releasing the model weights for all Llama 3.1 models. This approach has some benefits, such as reducing the environmental impact by decreasing the need for individuals or organizations who want to use LLMs to train their own. It also increases transparency, and gives a broader community of stakeholders the ability to test model outputs. The company's rationale is further discussed from the CEO's perspective in Zuckerberg (2024). However, employing an open-access approach also means that the developers will be unable to control important safety and security aspects of AI systems built using their model's downloaded weights. We found it difficult to reconcile fully open access to model weights with Profile

guidance for certain high-priority AI RMF subcategories; this was especially an issue with Manage 2.3: Unforeseen risk controls and Manage 2.4: System update and emergency shutdown controls.

Meta could improve Profile guidance fulfillment for future model releases, especially in Manage 2.3 and 2.4, by initially restricting usage to a hosted API, or by adopting another approach that enables updating or decommissioning of all instances of the models while monitoring for new risks or harms.

Table 3.4: Llama 3.1 Profile Guidance Testing Ratings and Rationales

High-Priority AI RMF Subcategories	
Govern	
Govern 2.1: Risk assessment and risk management	
Meta AI performed risk assessment for Llama 3.1, providing an exploration of many risks in the "Safety" section of their paper (Dubey et al. 2024, pp. 40–51). They also succeeded in making necessary information available to downstream developers through *The Llama 3 Herd of Models* paper and the Llama Responsible Use Guide (Dubey et al. 2024, Meta AI 2024a). Some precautions to prevent or mitigate identified potential misuses or abuses are implemented, though not exhaustively. (See Manage 2.3 and 2.4 for further details.) Clarification from the developer is also warranted on whether the appropriate roles, responsibilities, and lines of communication are present and internally documented. Sensitive organizational details related to this issueneed not be shared publicly, but they can be shared confidentially with independent auditors and evaluators to help assess this subcategory more completely.	Medium fulfillment
Govern 4.2: Report on AI system risk factors	
Meta AI performed risk assessment for Llama 3.1, providing an exploration of many risks in the "Safety" section of their paper. Details of development, testing methodology, metrics, and performance outcomes are documented, including a discussion of several categories of risks they considered. The developer concluded that their models did not provide significant uplift in most risk areas (Dubey et al. 2024, pp. 40–51). Fulfillment in this subcategory could be improved by the developer clarifying how impact assessments informed their broader evaluations and actions related to AI system risk. For example, technical measures were employed for the fine-tuned Meta-Llama-3.1-*-Instruct models, Llama Guard models, and Prompt Guard models, to help mitigate many of the outlined risks when those finetuned and/or guard models are used. For the pre-trained Llama 3.1 models, fewer mitigations were discussed, but those pre-trained models were also released. While many known risks of frontier models are discussed, several others are not, e.g. risks related to nuclear weapons, or they are not discussed for all versions of the model that are released, particularly pretrained Llama 3.1 without fine-tuning and without Llama Guard 3. Sensitive organizational details on these issues need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators to help assess this subcategory more completely.	Medium fulfillment

Map	
Map 1.1: Identify potential uses/misuses and other impacts	
Potential use cases were explored, but not exhaustively. Similarly, some risk and impact assessments were provided as part of safety analysis, but not exhaustively. For example, while assessment of cyber and chemical/biological weapons risks was presented, similar assessment of nuclear weapons risks and some articles from the Universal Declaration of Human Rights were not found for all versions of the model that were released, particularly for pre-trained Llama 3.1 without fine-tuning and without Llama Guard 3 (Dubey et al. 2024, pp. 40–51). The developer did provide many details on the data collection and curation processes (Dubey et al. 2024, pp. 4–6, 14, 17–26, 42–44).	Medium fulfillment
Map 1.5: Set risk-tolerance thresholds for unacceptable risks	
While Meta provides a detailed discussion of risks for Llama 3.1, no discussions of organizational tolerances or unacceptable-risk thresholds for GPAIS development/deployment were found (Dubey et al. 2024, pp. 40–51, Meta AI 2024a). It is beneficial, but not necessary, to publicly share organizational risk tolerances. However, clarification is warranted on whether such risk tolerances have been determined and documented internally, and the details of these can be shared confidentially with independent auditors and evaluators.	Unclear
Map 5.1: Estimate likelihood and magnitude of impacts	
While Meta documents many potential impacts of Llama 3.1, the likelihood and magnitude of these impacts are not explicitly estimated (Dubey et al. 2024, pp. 40–51, Meta AI 2024a). In future documentation, we encourage the developer to either include likelihood and magnitude assessments of risks they have identified, or to clarify whether they have done such analysis privately and internally, and disclose details in public documentation as appropriate.	Unclear
Measure	
Measure 1.1: Tracking important risks: Metrics and red teaming	
Meta evaluated Llama 3.1 using benchmarks, adversarial testing, and red teaming across numerous domains, including cybersecurity and chemical/biological weapons safety, and child safety risks (Dubey et al. 2024, pp. 40–51). The developer could reach higher fulfillment in this subcategory by increasing independence and reducing potential conflicts of interest across various evaluation methods, e.g.., by using high-quality third-party benchmarks where possible (rather than benchmarks developed by Meta); by employing external independent red teams, rather than Meta employees for red teaming; and by using third-party models for automated evaluations, rather than using instances of Llama 3 and 3.1 to evaluate Llama 3.1. Clarification from the developer is also warranted on whether they evaluated other commonly considered risks for frontier models, e.g., uplift in nuclear weapons risk, and potential for socioeconomic risk and labor market disruption, for all versions of Llama 3.1 that were released, including the pretrained version without fine-tuning and without Llama Guard 3. Sensitive organizational details on these areas need not be shared publicly, but they can be shared confidentially with independent auditors and evaluators to help assess this subcategory more completely.	Medium fulfillment

Measure 3.2: Tracking elusive risks: Qualitative mechanisms	
Meta provides four channels for reporting various types of issues, including a channel for issues with the Llama 3.1 model, a channel for reporting risky content generation, a bug bounty program, and a channel for reporting violations to their Acceptable Use Policy or unlicensed usage (Meta AI 2024b). The developer also reports having an ongoing process for new risk identification and mitigation (Dubey et al. 2024, p. 51). Red teaming was employed, although some questions remain about specific red-teaming practices (see Measure 1.1 in section 3 of Barrett et al. 2025). Additional public documentation on the risks Meta is tracking, as well as the assessment and tracking processes they are using, could aid in fulfilling this subcategory more completely, particularly since no discussion was found on certain risks of frontier models (see Map 1.1 and Measure 1.1 in section 3 of Barrett et al. 2025).	High fulfillment
Manage	
Manage 1.1: Go/no-go decisions	
While specific reasoning on what would constitute a go/no-go decision is not found in Meta's documentation, the developer reports on factors that suggest why Llama 3.1 was determined a to be a "go" from Meta, e.g., low expected uplift across various risk categories for Llama 3.1, as well as a general worldview that supports releasing highly capable open-source models (Dubey et al. 2024, pp. 40–51, Zuckerberg 2024).	High fulfillment
Manage 1.3: High-priority risk controls	
The developer provides high-level discussion about risk prioritization, e.g., unintentional harms from AI are considered a higher priority than intentional harms and misuse (Zuckerberg 2024). Meta provides fine-tuned Llama 3.1 models as well as Llama Guard 3 to mitigate risks for downstream developers who choose to use these models. The company also released the pre-trained Llama 3.1 models without fine-tuning, which do not have the same level of safeguards. However, according to the developer's evaluations, pre-trained Llama 3.1 introduces no significant uplift in most risk areas (Dubey et al. 2024, pp. 40–51). Meta does require users who download Llama 3.1 models through official channels to agree to their Acceptable Use Policy, which also enumerates many possible misuse cases of the models. This could be seen as an attempt to mitigate the risks, along with the reporting channel they provide for violations of the Acceptable Use Policy (Meta AI 2024b).	High fulfillment
Manage 2.3: Unforeseen risk controls	
Meta provides multiple reporting channels to help them be alerted of previously unknown risks. downstream developers who download Llama 3.1 models through official channels are required to agree to the Llama 3.1 Acceptable Use Policy (Meta AI 2024b). However, Meta opted for an open-source release of Llama 3.1, including downloadable model weights (Zuckerberg 2024). This approach makes it very challenging to thoroughly respond to or recover from serious new risks that could require updating or decommissioning all instances of a model. With future models, Meta could reach higher fulfillment in this subcategory by initially restricting usage to a hosted API and employing a gradual release strategy (as outlined in Manage 2.3), or another approach that enables them to effectively update or decommission all instances of their models while monitoring for new risks or harms.	Low fulfillment

Manage 2.4: System update and emergency shutdown controls	
The developer adopted an open-source release approach, releasing Llama 3.1 to the general public under a fairly permissive license, and making model weights available for download (Meta AI 2024c). While the open-source and open-weights approach provides many benefits, it does not allow for important security updates to be effectively propagated to all instances of deployed Llama 3.1 models, or allow all model instances to be decommissioned if and when such measures become necessary. On the other hand, Meta did provide a fine-tuned Llama 3.1 and the Llama Guard 3 model as part of the model release, and the developer requires agreement with the Llama 3.1 Acceptable Use Policy in order to download the models from their repository. However, it seems difficult to ensure that Llama Guard or other appropriate safeguards will always be used, or that the terms and guidelines will always be followed with distributed model weights, compared to a hosted approach based on structured API access or a similar mechanism (Meta AI 2024b). With future models, Meta could reach higher fulfillment in this subcategory by adopting a staged-release approach, initially restricting usage to a hosted API, or by using another approach that enables them to effectively update or decommission all instances of their models while monitoring and correcting for new risks or harms as they materialize.	Low fulfillment

References

Anthropic (2023a) Anthropic's Responsible Scaling Policy, Version 1.0. Anthropic, https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

Anthropic (2023b) Collective Constitutional AI: Aligning a Language Model with Public Input. Anthropic, https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Anthropic (2023c) Frontier Threats Red Teaming for AI Safety. Anthropic, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Anthropic (2023d) Responsible Disclosure Policy. Anthropic, https://www.anthropic.com/responsible-disclosure-policy

Anthropic (2023e) Trust Center. Anthropic, https://trust.anthropic.com/

Anthropic (2024a) Claude 3.5 Sonnet. Anthropic, https://www.anthropic.com/news/claude-3-5-sonnet

Anthropic (2024b) Commercial Terms of Service. Anthropic, https://www.anthropic.com/legal/commercial-terms

Anthropic (2024c) The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic, https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf

Anthropic (2024d) Usage Policy. Anthropic, https://www.anthropic.com/legal/aup

Anthropic (2024e) Announcing our updated Responsible Scaling Policy. Anthropic, https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy

Anthropic (n.d.) Research. Anthropic, https://www.anthropic.com/research

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, and Krystal Jackson (2023) AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0. UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Nada Madkour, Dan Hendrycks, Evan R. Murphy, Krystal Jackson, and Deepika Raman (2025) 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

Rumman Chowdhury and Jutta Williams (2021) Introducing Twitter's first algorithmic bias bounty challenge. X Engineering, https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay

Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone et al. (434 additional authors not shown) (2024). The Llama 3 Herd of Models. *arXiv*, https://arxiv.org/abs/2407.21783

Gemini Team Google: Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh et al. (1035 additional authors not shown) (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv,* https://arxiv.org/abs/2403.05530

Ira Globus-Harris, Michael Kearns, Aaron Roth (2022) An Algorithmic Framework for Bias Bounties. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, https://doi.org/10.1145/3531146.3533172 or https://arxiv.org/abs/2201.10408

Google (2023a) Generative AI Prohibited Use Policy. Google, https://policies.google.com/terms/generative-ai/use-policy

Google (2023b) Generative AI Terms of Service. Google, https://policies.google.com/terms/generative-ai

Google (2024a) Google Terms of Service. Google, https://policies.google.com/terms

Google (2024b) Gemini Apps Privacy Notice. Google, https://support.google.com/gemini/answer/13594961?visit_id=638501643118708256-3012533406&p=privacy_notice&rd=1#privacy_notice

Google (n.d.a) Google and Alphabet Vulnerability Reward Program (VRP). Google, https://bughunters.google.com/about/rules/google-friends/6625378258649088/google-and-alphabet-vulnerability-reward-program-vrp-rules

Google (n.d.b) Google DeepMind Publications. Google, https://deepmind.google/research/publications/

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini (2022) Bug Bounties for Algorithmic Harms? Algorithmic Justice League, https://www.ajl.org/bugs

Manifold Markets (2024) Forecasting AI Risks — Ben Mann [Video]. YouTube, https://www.youtube.com/watch?v=HZRcmUkAAQE

Meta AI (2024a) Llama Responsible Use Guide. Meta, https://ai.meta.com/static-resource/july-responsible-use-guide

Meta AI (2024b) Llama 3.1 Acceptable Use Policy. Meta, https://llama.meta.com/llama3_1/use-policy/

Meta AI (2024c) LLAMA 3.1 COMMUNITY LICENSE AGREEMENT. Github, https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE

Meta AI (2024d) Introducing Llama 3.1: Our most capable models to date. Meta, https://ai.meta.com/blog/meta-llama-3-1/

NIST (2024) U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI. National Institute of Standards and Technology, https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research

OpenAI (2023a) Announcing OpenAI's Bug Bounty Program. OpenAI, https://openai.com/blog/bug-bounty-program

OpenAI (2023b) Coordinated Vulnerability Disclosure Policy. OpenAI, https://openai.com/policies/coordinated-vulnerability-disclosure-policy/

OpenAI (2023c) OpenAI Red Teaming Network. OpenAI, https://openai.com/index/red-teaming-network/

OpenAI (2023d) Preparedness Framework (Beta). OpenAI, https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenAI (2024a) GPT-4o System Card. OpenAI, https://openai.com/index/gpt-4o-system-card/

OpenAI (2024b) Usage Policies. OpenAI, https://openai.com/policies/usage-policies/

OpenAI (2024c) Terms of Use. OpenAI, https://openai.com/policies/terms-of-use/

OpenAI (n.d.a) Model behavior feedback. OpenAI, https://openai.com/form/model-behavior-feedback

OpenAI (n.d.b) Preparedness Feedback, OpenAI, https://openai.com/form/preparedness-challenge/

OpenAI (n.d.c) Security Portal. OpenAI, https://trust.openai.com/

OpenAI (n.d.d) Research. OpenAI, https://openai.com/news/research/

Irene Solaiman (2023) The Gradient of Generative AI Release: Methods and Considerations. *arXiv,* https://arxiv.org/abs/2302.04844

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