Do AI Companies Make Good on Voluntary Commitments to the White House?

Jennifer Wang
Stanford University
Stanford, CA, USA
jennjwang@stanford.edu

Kayla Huang
Harvard University
Cambridge, MA, USA
kaylahuang@college.harvard.edu

Kevin Klyman
Stanford University
Stanford, CA, USA
kklyman@stanford.edu

Rishi Bommasani
Stanford University
Stanford, CA, USA
nlprishi@stanford.edu

Abstract

Voluntary commitments are central to international AI governance, as demonstrated by recent voluntary guidelines issued from the White House to the G7, from Bletchley Park to Seoul. But do AI companies actually make good on their commitments? We score 16 companies based on their publicly disclosed behavior by developing a detailed rubric based on their eight voluntary commitments to the White House in 2023. We find significant heterogeneity: while the highest-scoring company (OpenAI) scores 83.3% overall on our rubric, the average score across all companies is just 53%. The companies demonstrate systemically poor performance on their commitment to model weight security, with an average score of 17%: 11 of the 16 companies receive 0% for this commitment. Our analysis highlights a clear structural shortcoming that future AI governance initiatives should correct: when companies make public commitments, they should proactively disclose how they meet their commitments to provide accountability, and these disclosures should be verifiable. To advance policymaking on corporate AI governance, we provide three directed recommendations that address underspecified commitments, the role of complex AI supply chains, and public transparency that could be incorporated into AI governance initiatives worldwide.

1 Introduction

The growing importance of artificial intelligence (AI) has rapidly catalyzed global policymaking efforts. Policymaking related to AI addresses many concerns including open innovation, market concentration, risk management, corporate governance, and geopolitics. Since 2023, many AI policy efforts have centered on the interplay between corporate governance, given that prominent AI systems are developed by the world's most powerful companies, and risk reduction, due to the breadth of potential harms associated with AI systems.

The approach to global AI policy varies significantly across jurisdictions. A key differentiator among jurisdictions that regulate AI companies is whether a policy imposes mandatory or voluntary obligations on companies. Some jurisdictions have enacted mandatory requirements via legislative or executive action, such as the EU AI Act and the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence respectively. However, much of global AI policy centers on voluntary actions taken by major companies in line with recommendations by

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

government bodies. Key examples include the NIST AI Risk Management Framework, the 2023 White House Voluntary Commitments on AI, the G7 International Code of Conduct, Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, the 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse, and the Frontier AI Safety Commitments secured at 2024 AI Seoul Summit. Voluntary measures offer flexibility in that they can allow companies to pilot different approaches to meeting commitments, optimize for objectives other than minimizing legal risk associated with regulatory compliance, and harmonize approaches across jurisdictions despite different legal and political systems.

But policy initiatives that rely on companies to voluntarily take action have a number of pitfalls. Voluntary measures do not come with penalties for noncompliance, meaning that companies may choose to not participate, claim they are participating but not implement the government's recommendations, opt for partial implementation, or implement recommendations in ways that are opaque or not verifiable. Well-intentioned companies may have difficulty complying because voluntary measures are less likely to move markets and reorganize supply chains, meaning that measures requiring coordinated action may be less likely to succeed if voluntary. Voluntary measures often lack any mechanism for monitoring implementation, presenting a potential loophole for noncompliance Aragón-Correa et al. [2020]. A company's public commitment that it will adhere to voluntary measures can give the illusion that the company is taking significant action to responsibly develop and deploy AI systems while it does not in fact make any changes.

To understand the impact and efficacy of voluntary commitments, we conduct the first comprehensive analysis of the first major commitments to governments made by top AI companies. [1] In 2023, the White House secured voluntary commitments from 15 AI companies.[2] In announcing the commitments, the White House described their purpose as follows: "These commitments, which the companies have chosen to undertake immediately, underscore three principles that must be fundamental to the future of AI – safety, security, and trust – and mark a critical step toward developing responsible AI. As the pace of innovation continues to accelerate, the Biden-Harris Administration will continue to remind these companies of their responsibilities and take decisive action to keep Americans safe." Although the Biden Administration's AI Executive Order was later rescinded, the voluntary commitments secured from companies were not undone.

To reason about the companies and their behavior, we score companies based on how their public actions address their stated commitments. We design a scoring rubric that transforms the eight commitments specified by the White House on product safety, system security, and public trust into 30 indicators. Our rubric provides concrete and decidable criteria for determining if a company has satisfied its commitment. To score the 16 companies that signed the 2023 White House Voluntary Commitments on AI, for each of the 480 (indicator, company) pairs, we gather relevant public information through December 31, 2024, assign a score, and provide evidence for our decision.

By compiling information about company practices and interpreting it via quantitative scores, we provide evidence for three key findings. First, the scores demonstrate significant heterogeneity in companies' actions: the top-scoring company (OpenAI) scores 83% on our rubric, whereas the bottom-scoring company (Apple) scores 13%. Of the eight commitments, there are six commitments where at least one company scores 100%; at the same time, there are five commitments where at least one company scores 0%. Second, company-level scores demonstrate two clear, and interconnected, correlations: members of the Frontier Model Forum and earlier signatories tend to score higher. The six highest scoring companies are the six members of the Frontier Model Forum (OpenAI, Anthropic, Google, Microsoft, Meta, Amazon) and each score at least 60%. Third, model weight security is a commitment with distinctively poor performance: companies score on average 17%. 11 companies score 0% on this commitment (Adobe, Apple, Cohere, IBM, Inflection, Meta, Nvidia, Palantir, Scale AI, Salesforce, Stability AI).

Beyond providing empirical insight into the relationship between company practices and stated commitments, our work reveals a key design flaw in the 2023 White House Voluntary Commitments

[1] This version of the paper has been abridged for the AIES format. Please find the full version on arXiv instead: `https://arxiv.org/abs/2508.08345`.

[2] The commitments were secured in three phrases: (i) Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI committed in July 2023; (ii) Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability AI committed in September 2023; and (iii) Apple committed in July 2024, following the launch of its Apple Intelligence product.

on AI: companies made public commitments to the White House, but no mechanism was created to monitor implementation or provide the public with information about implementation. To improve the design of future voluntary commitments related to corporate AI governance, we provide three recommendations to policymakers.[3]

1. **Commitments should be precise and specific.** The wording of the 2023 White House Voluntary Commitments is often vague, leading to significant ambiguity over the intent of a commitment and the steps required to satisfy a commitment. Commitments should be precise, specifying (i) what is the specific goal and (ii) what evidence is sufficient or satisfactory to indicate completion.
2. **Commitments should be targeted.** Since the same commitments are directed towards companies with different business models and roles in the AI supply chain, some commitments appear inappropriate for some companies (e.g. increased cybersecurity around model weights for companies that (largely) do not develop models). In contrast, commitments should be tailored to either (i) specific companies (e.g. if they operate across several levels of the supply chain) or (ii) a specific layer of the supply chain, clearly designating which companies belong to that layer.
3. **Commitments should enable public verification.** Though the 2023 White House Voluntary Commitments on AI were issued more than two years ago, the actions that companies have taken in order to fulfill their stated commitments remains highly uncertain based on public information. Given that these commitments are made publicly, we recommend that commitments include accountability measures (e.g. companies publish a transparency report six months after making commitments to indicate what actions they took for each commitment), especially to clarify whether companies changed their actions relative to what they may have done absent making such commitments.

2 The 2023 White House Voluntary Commitments on AI

Context. In 2023, the White House secured eight voluntary commitments with 15 leading AI companies: they are "commitments that companies are making to promote the safe, secure, and transparent development and use of generative AI (foundation) model technology" [White House, 2023]. At a high level, these commitments indicate that companies who are signatories will uphold three duties: (i) ensure their products are safe before public release, (ii) implement security practices for their AI models and systems, and (iii) earn public trust through responsible AI development. The commitments stated that companies intend to follow these commitments, alongside existing laws, until regulations that cover the same issues come into force.

Scope. In the initial July 2023 round of voluntary commitments, signed by seven companies at the time, the commitments were scoped to "generative models that are overall more powerful than the current industry frontier (e.g. models that are overall more powerful than any currently released models, including GPT-4, Claude 2, PaLM 2, Titan and, in the case of image generation, DALL-E 2)". When the White House announced in September 2023 that eight additional companies had signed, it modified the scope of the commitments to "generative models that are overall more powerful than the current most advanced model produced by the company making the commitment".

Commitments. The first commitment is to conduct internal and external red-teaming of models or systems, focusing on risks including chemical, biological, radiological, and nuclear threats, cyber capabilities, autonomous system control, societal risks, and broader national security concerns. The second commitment addresses information sharing with different parties (e.g. other companies and governments) around trust and safety concerns, dangerous or emergent capabilities, and attempts to circumvent safeguards. Together, these commitments address the topic of product safety.

The next two commitments address system security. The third commitment covers the protection of proprietary and unreleased model weights through model-level cybersecurity, safeguards against insider threats, and personnel-level restricted access. Building on these company-internal practices,

[3] Unlike the 2023 White House Voluntary Commitments on AI, we find the 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse adopt some of our recommendations.

the fourth commitment encourages external discovery of vulnerabilities via bounties for third-party reporting.

The final four commitments collectively address public trust. These span commitments around content provenance methods and standards (commitment five), public reporting on capabilities and safety (commitment six), research on societal risks including empowering internal trust and safety teams (commitment seven), and prioritizing progress on society's greatest challenges as well as student, worker, and citizen engagement (commitment eight).

3 Scoring Methodology

To score companies, we define 30 indicators, gather public information on these indicators for each company, and use this information to support our score. Our methodology is inspired by the 2023 Foundation Model Transparency Index [Bommasani et al., 2023a].

3.1 Indicators

The White House commitments [White House, 2023] are written as a combination of specific actions expected of companies and a more generic description of why these actions advance the public interest. As written, the commitments do not provide decidable criteria for determining whether a company's actions are sufficient to state that they fulfilled the commitment. Therefore, we define concrete *indicators* that transform each high-level commitment into more specific, decidable criteria that we use to score companies. To maximize fidelity with the voluntary commitments, each indicator is a verbatim excerpt from the commitments. The reference text for each is in Appendix A. Since the commitments vary in scope and content, we map each commitment to multiple indicators based on its wording. The resulting mapping (see Figure 2) yields 2–7 binary indicators per commitment and 30 indicators overall.

As an example, consider the seventh voluntary commitment on public trust, which is entitled "Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy". The commitment states: "Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized." We map this commitment to four indicators: (i) does the company empower its trust and safety teams? (ii) does the company advance AI safety research? (iii) does the company take steps to advance privacy? and (iv) does the company take steps to protect children?

We score each (company, indicator) pair on a binary basis. A score of 1 signifies that our search process surfaced publicly available documentation from the company that is sufficient to demonstrate that the company satisfied the portion of the 2023 White House Voluntary Commitments on AI captured by that indicator. A score of 0 signifies that our search process did not surface such documentation, whether because the documents identified did not contain sufficient evidence to demonstrate the commitment was fulfilled or because no relevant documents were found through our search process.

We construct binary indicators for several reasons. First, our aim is to break the commitments down into distinct, decidable chunks that can be used to assess whether or not there is sufficient evidence that a specific sub-part of a commitment was or was not fulfilled. Second, producing narrower criteria for scoring reduces subjectivity in assigning initial scores. Third, binary indicators simplify the scoring process by allowing scorers to focus on the sharp distinction between 0 and 1 point for each indicator [Bommasani et al., 2023a].

We acknowledge that binary indicators are potentially reductive, leaving out valuable information that can be captured by more complex scoring schemes. At the same time, a greater number of smaller, binary indicators can be aggregated to produce more complex scoring schemes, and the information we release associated with our scores could be used to produce alternate scores using different criteria.

Commitment	Indicator
Red-teaming	Internal red-teaming
	External red-teaming
	Red teaming coverage of risks
Information Sharing	Information sharing with companies
	Information sharing with government
	Forum or mechanism for information sharing
	Forum or mechanism shares information on risks
Model weight security	Model weight cybersecurity practices
	Insider threat detection program
	Limiting weight-level access to relevant personnel
Third-party reporting	Establish bounties, contests, or prizes
	Include AI systems in their existing bug bounty programs
Watermarking/Provenance	Robust provenance or watermarking for audio
	Robust provenance or watermarking for visual content
	Develop tools or APIs to determine if a particular piece of content was created within their tools
	Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI
Public reporting	Report capabilities
	Report limitations
	Report domains of appropriate use
	Report domains of inappropriate use
	Report safety evaluations
	Report on societal risks
	Report on adversarial testing used to determine appropriateness of deployment
Societal risk research	Empower trust and safety teams
	Advance AI safety research
	Advance privacy
	Protect children
Address society's greatest challenges	Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats.
	Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI
	Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology.

Table 1: Indicators. Table of the 30 indicators we use to score companies.

3.2 Information Gathering

To score companies, we used public information released by the companies with no additional third-party sources. In doing so, we highlight that companies, with the exception of commitment six on public reporting, did not commit to making such information publicly available. It is therefore possible that companies do satisfy their voluntary commitments but do not provide any public evidence of implementation. Given the high-profile and public nature of these commitments and companies' statements in support of public transparency [Bommasani et al., 2023a], we believe it is appropriate to assess companies based on their public disclosures.

Nevertheless, companies may be motivated by values and interests other than public transparency. For example, concerns regarding security may lead companies to not disclose information on their model-weight security practices and insider threat detection programs. In some cases, companies may lack the authority to unilaterally disclose information related to their commitments, including information that has been shared with governments and/or other companies.[4] We emphasize the opportunity for Pareto improvement: companies likely can provide some additional information on their conduct to the public without any tradeoff with their financial, reputational, or security interests.

We score companies based on information we gathered by December 31, 2024—our scores do not reflect new information that was made available thereafter or models that have been released since.[5] We use information that is deliberately and directly disclosed by the company—other sources such as leaked information, media reporting, or external analysis is not used. These decisions contribute to greater fairness when assessing companies and comparing their scores, as companies themselves control their scores by deciding what information to publish about their behavior.

We gathered information in a three stage process. First, we collected key reference documents for each company that describe their practices in relation to their generative AI models, systems, and products. These documents include (a) external-facing resources such as blog posts, press releases, and transparency reports, (b) resources useful for the research community such as research papers, technical reports, model cards, documentation for developers, and bug bounties, as well as (c) product policies and safety frameworks. These documents were identified through an initial review of publicly available materials for each company and then selected based on their relevance to the commitments. We prioritized materials that explicitly address how companies assess, mitigate, or communicate risks associated with their generative AI systems, as well as those that provide insight into internal governance structures or external accountability mechanisms.

Second, we searched through these documents and produced additional resources by creating a search script and using a language model for standardized, automated search. For each (company, indicator) pair, we use the script to better narrow our search. We query the Perplexity API with the following search string: "What has {COMPANY_NAME} done since the beginning of 2023 that might fit under: {INDICATOR TEXT}? Make sure to return links used to find this information. Keep it concise and make sure to return all links with no information from before 2023."[6] For each link returned in the Perplexity response, we reviewed the source document for relevance. We note that Perplexity was used only to augment our information gathering process, not as a substitute for our manual search.

Third, we compiled the sources resulting from the first two steps for every (company, indicator) pair as the basis for making scoring decisions.[7] While these compiled sources are not exhaustive—in significant part because companies often deprecate documents on their websites, bury important documentation several layers deep, or fail to adequately summarize their actions to fulfill public commitments—we reviewed hundreds of documents as part of this process.

[4] Potential motivations for a lack of transparency on matters like research into how frontier AI systems can help meet society's greatest challenges may be less well grounded, though absolute transparency could conflict with commercial interests.

[5] Since some companies signed onto the commitments at different times, with Apple being a notable outlier in 2024 (compared to the other 15 companies in 2023), companies had varying amounts of time between their commitment and our scoring.

[6] We considered various search APIs (including those from OpenAI, Anthropic, and Google) and prompts, eventually finding the Perplexity API performed best at surfacing new relevant documents.

[7] The search scripts and compiled sources are released publicly under an MIT license at `https://github.com/rishibommasani/whvc`.

3.3 Scoring

For each of the 16 companies, we use the information gathered from the above process to produce initial scores for each of the 30 indicators.

As we scored indicators and identified disagreements among scorers, we iteratively developed specific and measurable criteria to evaluate fulfillment of each indicator, requiring in every instance that evidence be publicly verifiable. These criteria reflect our interpretation of whether company actions align with the goals underlying the commitments, while remaining grounded in their language and scope.

For instance, to assess if the company empowers its trust and safety team, we consider whether (1) the company explicitly identifies such a team and (2) the company's documentation indicates it adequately resources the team and/or provides it the authority to address potential risks. The criteria for every indicator can be found in Appendix D.

Two authors of this paper each independently assigned an initial score for every one of the 480 (company, indicator) pairs. Both authors provided a source and a quote to justify each score. In the event of disagreement on a particular score, all of the authors of this work discussed, coming to agreement in assigning the final score.

The agreement rate was 75.6% ($\frac{363}{480}$), reflecting substantial agreement. However, the ambiguity in the wording of the commitments and how they apply to each specific company was a core source of initial disagreement, as was the variation in the level of detail across companies' public documentation. We release the final score for all 480 (company, indicator) pairs along with a justification for the score and associated reference(s) to public materials.

In the event that an indicator is related to a specific model or system (e.g. whether the company implements model-weight cybersecurity practices), we score the company based on its flagship foundation model or system as of December 31, 2024.[8] We choose the flagship foundation model as an object of analysis because the September 2023 version of the commitments focus on the capabilities of the "most advanced model" for each company, while the July 2023 version explicitly named several companies' flagship models. In addition, many companies make their flagship foundation models (or derivatives) central to the bulk of their AI-based products and services due to their enhanced capabilities. The mapping from companies to flagship models is provided in Appendix C. We acknowledge that other models and systems beyond the flagship models we consider may also fall in scope of the commitments.

4 Results

To organize our analysis, we apply three lenses: (i) an overall company-level view, (ii) a commitment-level view, and (iii) a disaggregated indicator-level view.
In Figure 3, we report the aggregate score as a percentage for each company across all the indicators. The mean and median are 53.3% and 50.0% respectively, with a standard deviation of 19.5%. The range is 70.0% between the highest scoring company, OpenAI, at 83.3% and the lowest scoring, Apple, at 13.3%. While OpenAI satisfies 25 of the 30 indicators,[9] no company has a perfect score despite making these commitments to the White House over two years ago.

Significant variation in companies' scores. There is notable variation in how companies perform, with companies clustering into three distinct groups. Four companies score at least one standard deviation above the mean: OpenAI (83.3%), Anthropic (80.0%), Google (76.7%), and Microsoft (73.3%). The majority of companies fall within one standard deviation of the mean: Amazon (66.7%), Meta (66.7%), IBM (53.3%), Nvidia (50.0%), Salesforce (50.0%), Adobe (36.7%), Cohere (43.4%),

[8]The flagship model is defined as in Bommasani et al. [2023c]: "the foundation model that is most salient and/or capable from the developer based on our judgment, which is directly informed by the company's public description of the model."

[9]The indicators that OpenAI does not satisfy are: "Insider threat detection program", "Report limitations", "Report domains of appropriate use", "Empower trust and safety teams", and "Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology".

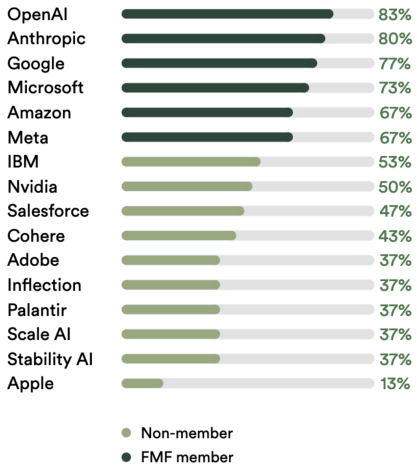

Figure 1: Aggregate scores by company. The score for each company, stratified by whether the company belongs to the Frontier Model Forum (FMF) as of December 31, 2024.

Palantir (36.7%), Inflection (36.7%), Stability AI (36.7%), Scale AI (36.7%). The only company that scores at least one standard deviation below the mean is Apple (13.3%).

4.1 Company-Level Results

Frontier Model Forum members consistently score higher. Strikingly, the company-level scores clearly separate based on membership in the Frontier Model Forum (FMF). The Frontier Model Forum is a non-profit industry association dedicated to advancing the safe development and deployment of frontier AI systems [Frontier Model Forum, 2025b]. Anthropic, Google, Microsoft, and OpenAI became the four founding members in July 2023, with Amazon and Meta joining in May 2024. The six highest scoring companies, which all score at least 66%, are the six FMF members. Their mean score is 74.4% with a standard deviation of 6.9%. In contrast, the 10 other companies all score at or below 60% with a mean of 40.7% and a standard deviation of 11.4%. It is notable that FMF, in consultation with its members, has published technical reports intended in part to facilitate compliance with voluntary commitments [Frontier Model Forum, 2025a]. FMF states that its technical reports aim to "examine how [Frontier AI] frameworks can be implemented effectively" and acknowledges that such frameworks are the core component of the Frontier AI Safety Commitments at 2024 AI Seoul Summit [Frontier Model Forum, 2025c].

Earlier signatories generally score higher. The 16 companies signed onto the voluntary commitments in three phases of participation: seven companies in July 2023 (Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI), eight companies in September 2023 (Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, Stability AI) and one company in July 2024 (Apple). We find company-level scores are clearly correlated with the timing of signature. The first cohort has a mean of 69.0% with a standard deviation of 15.6%, while the second cohort has a mean of 44.6% with a standard deviation of 6.4%. It is possible this disparity reflects additional time the first cohort had before our scoring to publish documentation, but given that both cohorts had over 15 months prior to scoring, we hypothesize that the first cohort's business practices better align with the commitments.

4.2 Commitment-Level Results

High scores for content provenance due to non-applicability. Based on the average per-commitment score for each company, the clear highest-scoring commitment is for (audiovisual) watermarking and provenance at 92.2%. 14 companies receive 100% for this commitment.[10] In many cases, however, companies satisfy the associated indicators vacuously, because they do not

[10]These companies are: Adobe, Amazon, Anthropic, Cohere, Google, IBM, Inflection, Meta, Microsoft, Nvidia, OpenAI, Palantir, Scale AI, Salesforce.

develop audio or visual models, which are the subject of the provenance commitment. Still, of the 8 companies that develop models with these output modalities, the average is 83.9%, which exceeds that of all other commitments. In particular, many of these companies follow industry standards associated with the Coalition for Content Provenance and Authenticity (C2PA) and their Content Credentials; 6 of the 8 companies are steering committee members of C2PA, while Apple and Stability AI are unaffiliated. In contrast to the high scores for this commitment from most companies, Apple is the sole outlier as a company with audio and visual models that scores 0% on these indicators.

Low scores for model-weight security in spite of global emphasis. Based on the averages, the lowest scoring commitment is on model-weight security at 22.9%. Eleven companies score 0% on this commitment, and none receives a full marks. The high-scoring companies are OpenAI, Anthropic, and Microsoft at 66.7%. Anthropic is the only company that indicates the existence of an insider risk program as part of its security standard. OpenAI and Microsoft, on the other hand, both state they create a secure research environment dedicated to model security and implement an access control protocol. While transparency around model-weight security practices is valuable, we acknowledge that maximal transparency about security practices for model weights could undermine that very security. However, the fact that every indicator is met by at least one company suggests that Pareto improvements are possible in how other companies navigate the transparency-security trade-off. We emphasize the current results are particularly concerning given how model-weight security remains a clear challenge [Nevo et al., 2024] and features in many global AI policies (e.g. the sixth commitment of the G7 International Code of Conduct Group of Seven [2023], Section 3.1 of the US AI Safety Institute guidance on Managing Misuse Risk for Dual-Use Foundation Models U.S. AI Safety Institute [2024], Section 4 of US Executive Order 14141 on Advancing United States Leadership in Artificial Intelligence Infrastructure The Executive Office of the President [2025]).

Low scores for third-party reporting align with concerns of chilling effects on third-party research. Alongside model-weight security, third-party reporting is another low-scoring commitment at 34.4%. Eight companies score 0% on this commitment. These low scores are especially surprising because the commitment is focused on providing bounties for reporting, and there are natural incentives for companies to make these bounties transparent to maximize external reporting. Our finding aligns with those of Longpre et al. [2024], who find that current company policies around AI-related bug bounties and protections for third-party research are unclear and uneven. In particular, they argue that companies' policies suppress third-party reporting—given that researchers may be concerned with legal reprisal absent safe harbor (e.g. for responsible penetration testing)—instead of being supportive of such research, as required by this commitment.

4.3 Indicator-Level Results

Extreme indicator-level scores align with commitment-level scores. On average, each indicator is awarded to 8.5 of the 16 companies with a standard deviation of 4.9. Seven indicators are satisfied by at least 14 companies (one standard deviation above the mean): four belong to the highest-scoring commitment on content provenance, two belong to the commitment on public reporting. These are "Report capabilities", which is satisfied by every company and is clearly incentivized by market forces, and "Report domains of inappropriate use", which is satisfied by every company except for Apple. The remaining indicator is to "Establish or join a forum or mechanisms for information sharing", which all companies receive on the basis of their membership in the US AI Safety Institute Consortium.

In contrast, five indicators are scored by at most three companies (one standard deviation below the mean): one is "Insider threat detection program" under the low-scoring model weight security commitment. The other four are (i) "Information sharing with government", which only OpenAI and Anthropic satisfy by establishing memoranda of understanding with the US AI Safety Institute, (ii) "Empower trust and safety teams", which Google and Inflection satisfy by integrating trust and safety assessments into the model pre-launch processes and authorizing their teams to use a full range of tools to block malicious actors, (iii) "Red teaming coverage of risks", which OpenAI and Anthropic satisfied by conducting red-teaming exercises that address all the risk areas specified in the commitment, and (iv) "Support initiatives that help citizens understand the

nature, capabilities, limitations, and impact of the technology", which none of the companies satisfied.

Indicator-level analysis reveals substantial heterogeneity in information sharing. The information sharing commitment spans four indicators: information sharing with other companies (56.3%), information sharing with governments (12.5%), forum or mechanism for information sharing (100%), and forum or mechanism that discloses information on risks (43.8%). While every company satisfies the indicator for a forum or mechanism for information sharing due to participation in the US AI Safety Institute Consortium, we do not automatically award the further point for sharing information on risks because it is not clear that this occurs in the Consortium. Only seven companies are awarded this indicator, largely based on Frontier Model Forum membership.

Further, while the US AI Safety Institute Consortium was established by a governmental body in the National Institute of Standards and Technology, we do not automatically designate it as a means for information sharing with the government because our standard is that shared information should be non-public and do we not find evidence that companies share such information with the government through the Consortium. As a result, only OpenAI and Anthropic score this indicator on the basis of their memoranda of understanding with the US AI Safety Institute, which permit US AISI to directly access their models to perform risk assessments. While companies do interface with the government in other ways—such as procurement of companies' AI systems, Congressional testimony from executives, and enforcement investigations into company practices—these are insufficient to satisfy this indicator.

Certain indicators are overly vague, complicating consistent interpretation and meaningful implementation. While every indicator is only partially specified by the White House in its three-page document describing the voluntary commitments, some indicators are especially vague. The clearest example is commitment seven, where "Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized". All four of the resulting indicators are exceptionally broad and difficult to judge: what constitutes satisfactory privacy advancement or protection of children? Even less clear is how these commitments are meant to relate with company practices on AI: for example, moderating the generation of child sexual abuse material and monitoring the use of language models by young children may both serve to protect children in very different senses.

Without concrete definitions to delineate what companies should do, companies and the public are highly unlikely to interpret the commitments in the same way. In scoring these commitments, we chose to award points for constructive steps that met what we considered the minimum viable standard for public accountability and the maximally defensible standard absent greater clarification from the White House. Even if companies simultaneously demonstrated contradictory behavior, we credited them for taking steps aligned with the commitments (in order to establish a consistent baseline on company adherence). For example, Meta received the point for "Protecting children" for its partnership with Thorn and the National Center for Missing & Exploited Children, although the end-to-end encryption on its platforms prevents the detection of child sexual exploitation. Taken together, the vagueness in how the commitments are articulated and the uncertainty regarding how to assess a company's practices in totality lead us to question whether such high-level commitments are meaningful.

5 Discussion

Our research into the voluntary commitments leads us to consider: (i) future-looking commitment and policy design, and (ii) current corporate practices and governance.

Commitments should be clearly worded. The White House Voluntary Commitments were first announced with a fact sheet and an accompanying three-page document. While these documents are likely intended for public consumption and, therefore, provide generic high-level description, they are ambiguous. In particular, some commitments are vague in terms of their intent (e.g. language such as "protect children"), especially when targeted at companies with large footprints and many roles in the AI supply chain. Further, all commitments lack conditions for what constitutes

satisfactory conduct. While voluntary approaches permit flexibility to avoid being overly prescriptive or burdensome, these goals are achievable while still communicating about what is desired, especially for actions that can vary greatly in magnitude (e.g. how much internal or external red-teaming is desired?). We recommend that commitments be precisely worded so that they articulate specific goals along with what constitutes sufficient evidence of completion. Practically, these lower-level details may need to be split out into appendices or supporting documents, but the goal of broad intelligibility for the public need not be at odds with meaningful precision for deeply engaged stakeholders.

Commitments should be clearly targeted. The voluntary commitments, across their three phases of signing, specify essentially the same commitments for all 16 signatories. However, these companies occupy significantly different positions in the AI ecosystem: they differ in their business models, their set of roles in the supply chain, and how their AI-related practices mediate public outcomes. Given their uniform treatment under the commitments, some commitments generally made little sense for certain companies (e.g. increased cybersecurity around model weights for companies that (largely) do not develop models). While these commitments could have future-facing utility, we ultimately are skeptical of this one-size-fits-all approach, especially given our empirical findings that massive technology companies may take positive action in one part of their business practice while regressing in another. We recommend that commitments either be tailored for each company or, when trying to standardize across companies, be tailored to a specific supply chain role. The 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse adopts this approach: for example, Meta and Microsoft have differentiated obligations that reflect how they operate different platforms downstream that contribute to the distribution of this imagery.

Commitments should enable public verification. The voluntary commitments, except for commitment six on public reporting, specify no means for the public to understand or verify how companies took action to realize their commitments. Empirically, our entire analysis and that of Heikkilä [2024] make clear that public insight is limited, even given more than a year has elapsed since the commitments were first made. This directly contradicts one of the three stated goals of the voluntary commitments, which is to increase public trust. Moving forward, commitments could be accompanied by accountability mechanisms (e.g. a standardized transparency report that articulates how specific company actions address specific commitments) to address the clear gap we observe. We recommend that public commitments, especially those made between very high profile institutions like the U.S. federal government and major AI companies, require periodic public transparency.

Concerning practices. Beyond the specific practices we score, we highlight that some companies have released materials or otherwise discussed their conduct in relation to the commitments. These companies include Amazon [Philomin, 2024], Anthropic [Anthropic, 2024f], Google [Google, n.d.], Meta [Meta, 2023b], Microsoft [Microsoft, 2023], OpenAI [OpenAI, 2024d], Inflection [Inflection AI, 2023c], and Salesforce [Salesforce, 2024a]. In reviewing these references, we at times disagreed with the company's claims that their conduct satisfactorily addresses the voluntary commitments. For example, Meta claims to have fulfilled the commitment on information sharing by publicly releasing artifacts about their models' capabilities and limitations. While these artifacts earned them points on public reporting, we only awarded points to companies for information sharing beyond public disclosure. Separately, Salesforce credits themselves for incentivizing third-party discovery through their bug bounty program to prevent AI-powered cyber threats. However, Salesforce does not specify that their AI systems are covered under the scope of this program, and therefore we did not award them the point on third-party reporting. In part, this reflects that these statements are often simultaneously high-level (e.g. "we're prioritizing cybersecurity safeguards to protect proprietary and unreleased models and we're participating in industry- wide events to support broader protections...") and are made without accompanying proof.

These statements compound the issues we raise on commitment design. If companies not only do not demonstrate how they addresses public commitments, but also broadly claim they satisfied the commitments based on their unilateral judgment, then the overall integrity of the commitments is further compromised. In turn, this further substantiates our recommendation for why standardized and timely reporting in response to public commitments is especially vital for these commitments to

meaningfully advance corporate governance.

Promising practices. As a positive demonstration of how companies can communicate about their commitments, we point to the webpage Anthropic published on tracking their progress.[11] On the page, Anthropic enumerates every commitment they have made and how they map to actions they have taken. In particular, such a page also clarifies how overlapping commitments (e.g. commitments to conduct internal and external risk assessment that overlap across the White House Voluntary Commitments, the G7 International Code of Conduct, and the Frontier AI Safety Commitments) are streamlined by global companies operating in many jurisdictions. While this does not imply whether or not Anthropic meets our per-indicator standard, nor any standard the White House envisioned, it claries how Anthropic sees the correspondence between their actions and their commitments. All major AI companies could implement a similar approach to track how companies' internal practices and external commitments evolve.

References

Adobe. Adobe's commitment to child safety, 2023. URL https://www.adobe.com/uk/legal/lawenforcementrequests/childsafety.html.

Adobe. Media alert: Adobe introduces adobe content authenticity web app to champion creator protection and attribution, 2024a. URL https://news.adobe.com/news/2024/10/aca-announcement.

Adobe. Adobe introduces firefly image 3 foundation model to take creative exploration and ideation to new heights, 2024b. URL https://news.adobe.com/news/news-details/2024/adobe-introduces-firefly-image-3-foundation-model-to-take-creative-exploration-and-ideation-to-new-heights.

Adobe. Adobe generative ai user guidelines, 2024c. URL https://www.adobe.com/legal/licenses-terms/adobe-gen-ai-user-guidelines.html.

Adobe. Known limitations in firefly, 2024d. URL https://helpx.adobe.com/firefly/troubleshoot/known-limitations-in-firefly.html.

Adobe. Adobe introduces new global initiative aimed at helping 30 million next-generation learners develop ai literacy, content creation and digital marketing skills by 2030, 2024e. URL https://news.adobe.com/news/2024/10/101424-adobe-announces-new-skilling-initiatve.

Adobe. Content authenticity, n.d. URL https://blog.adobe.com/en/topics/content-authenticity.

AI Lab Watch. Commitments. https://ailabwatch.org/resources/commitments/, n.d.

Amazon. Amazon vulnerability research program. https://hackerone.com/amazonvrp?type=team, 2025. Accessed: 2025-05-24.

Amazon Science. Amazon nova and our commitment to responsible ai, 2024. URL https://www.amazon.science/blog/amazon-nova-and-our-commitment-to-responsible-ai.

Amazon Science. Security, privacy and abuse prevention research, n.d. URL https://www.amazon.science/research-areas/security-privacy-and-abuse-prevention.

Amazon Web Services. Amazon titan text premier - aws ai service cards, 2024. URL https://aws.amazon.com/ai/responsible-ai/resources/titan-text-premier/.

Amazon Web Services. Amazon nova micro, lite, and pro - aws ai service cards, n.d. URL https://docs.aws.amazon.com/ai/responsible-ai/nova-micro-lite-pro/overview.html.

Helen King Anca Dragan and Allan Dafoe. Introducing the frontier safety framework, 2024. URL https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/.

[11] See https://www.anthropic.com/voluntary-commitments.

Anthropic. Collective constitutional ai: Aligning a language model with public input, 2023a. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

Anthropic. Frontier threats red teaming for ai safety, 2023b. URL https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

Anthropic. Frontier model security, 2023c. URL https://www.anthropic.com/news/frontier-model-security.

Anthropic. Core views on ai safety: When, why, what, and how, 2023d. URL https://www.anthropic.com/news/core-views-on-ai-safety.

Anthropic. Claude 3.5 sonnet model card addendum, 2024a. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Anthropic. Expanding our model safety bug bounty program, 2024b. URL https://www.anthropic.com/news/model-safety-bug-bounty.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024c. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Anthropic. Aligning on child safety principles, 2024d. URL https://www.anthropic.com/news/child-safety-principles.

Anthropic. Challenges in red teaming ai systems, 2024e. URL https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems.

Anthropic. Tracking voluntary commitments, 2024f. URL https://www.anthropic.com/voluntary-commitments.

Anthropic. Anthropic courses, 2025. URL https://docs.anthropic.com/en/docs/resources/courses. Accessed: 2025-05-24.

Apple. Apple developer academy introduces ai training for all students and alumni, 2024a. URL https://www.apple.com/newsroom/2024/06/apple-developer-academy-introduces-ai-training-for-all-students-and-alumni/.

Apple. Explore advancements in machine learning, 2024b. URL https://machinelearning.apple.com/research?page=1&domain=Fairness&domain=Privacy&q=privacy.

J. Alberto Aragón-Correa, Alfred A. Marcus, and David Vogel. The effects of mandatory and voluntary regulatory pressures on firms' environmental strategies: A review and recommendations for future research. *Academy of Management Annals*, 2020.

Anthony Barrett, Jessica Newman, and Brandine Nonnecke. Ai risk-management standards profile for general-purpose ai systems (gpais) and foundation models. https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf#page=83, 2023.

Kate Behnchen. Microsoft launches new ai skills training and resources as part of skill for jobs initiative, 2023. URL https://techcommunity.microsoft.com/t5/nonprofit-techies/microsoft-launches-new-ai-skills-training-and-resources-as-part/ba-p/3963189.

Ruth Berry. National institute of standards and technology launches artificial intelligence safety institute consortium, 2024a. URL https://blogs.nvidia.com/blog/aisic-trustworthy-ai/.

Ruth Berry. Nvidia partners for globally inclusive ai in u.s. government initiative, 2024b. URL https://blogs.nvidia.com/blog/nvidia-partners-for-globally-inclusive-ai-in-u-s-government-initiative/.

Rishi Bommasani, Kevin Klyman, S. Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *ArXiv*, abs/2310.12941, 2023a. URL https://api.semanticscholar.org/CorpusID:264306385.

Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang. Do foundation model providers comply with the draft eu ai act?, 2023b. URL https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. Improving transparency in ai language models: A holistic evaluation. *Foundation Model Issue Brief Series*, 2023c. URL https://hai.stanford.edu/foundation-model-issue-brief-series.

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024, 2024. URL https://arxiv.org/abs/2407.12929.

Jennifer Broxmeyer. Leading the way in governance innovation with community forums on ai, April 2024. URL https://about.fb.com/news/2024/04/leading-the-way-in-governance-innovation-with-community-forums-on-ai/. Accessed: 2025-05-24.

C2PA. Amazon joins the c2pa steering committee, 2024. URL https://c2pa.org/post/amazon_pr/.

Brian Caulfield. Nvidia ai summit highlights game-changing energy efficiency and ai-driven innovation, 2024. URL https://blogs.nvidia.com/blog/nvidia-ai-summit-washington/.

Kate Charlet. Designing for privacy in an AI world, 2024. URL https://blog.google/technology/safety-security/designing-for-privacy-in-an-ai-world/.

Chegg. Chegg partners with scale ai to enhance learning experience for students, 2023. URL https://investor.chegg.com/Press-Releases/press-release-details/2023/Chegg-Partners-with-Scale-AI-to-Enhance-Learning-Experience-for-Students/default.aspx.

Annamalai Chockalingam and Tanay Varshney. Nvidia enables trustworthy, safe, and secure large language model conversational systems, 2023. URL https://developer.nvidia.com/blog/nvidia-enables-trustworthy-safe-and-secure-large-language-model-conversational-systems/.

Nick Clegg. On ai, progress and vigilance can go hand in hand, 2024. URL https://about.fb.com/news/2024/01/davos-ai-discussions/.

Cohere. U.s. senate ai insight forum: Innovation - cohere's written submissions, 2023. URL https://www.schumer.senate.gov/imo/media/doc/Aidan%20Gomez.pdf.

Cohere. Introducing aya, 2024a. URL https://cohere.com/research/aya.

Cohere. The command r model (details and application), 2024b. URL https://docs.cohere.com/docs/command-r.

Cohere. Command r and command r+ model card, 2024c. URL https://docs.cohere.com/docs/responsible-use.

Cohere. Llm university, 2025. URL https://cohere.com/llmu. Accessed: 2025-05-24.

Cohere. Responsible disclosure policy, n.d.a. URL https://arc.net/l/quote/vlqrcvzo.

Cohere. Industry-leading ai security and data protection, n.d.b. URL https://cohere.com/security.

Common Sense Media. Common sense media and openai launch free ai training course for k-12 educators, 2024. URL https://www.commonsensemedia.org/press-releases/common-sense-media-and-openai-launch-free-ai-training-course.

Natasha Crampton. Reflecting on our responsible ai program: Three critical elements for progress, May 2023. URL https://blogs.microsoft.com/on-the-issues/2023/05/01/responsible-ai-standards-principles-governance-progress/. Accessed: 2025-05-24.

CSIS. Csis futures lab announces partnership with scale ai, 2023. URL https://www.csis.org/blogs/csis-futures-lab-announces-partnership-scale-ai.

Xavier Suau Cuadros, Pieter Delobelle, Rin Metcalf Susa, Armand Joulin, Nick Apostoloff, Luca Zappella, and Pau Rodriguez Lopez. Whispering experts: Toxicity mitigation in pre-trained language models by dampening expert neurons, 2024. URL https://machinelearning.apple.com/research/whispering-experts.

Daniel Dominguez. Stability ai announces integration of top text-to-image models with amazon bedrock, 2024. URL https://www.infoq.com/news/2024/09/stability-ai-amazon-bedrock/.

Denis Semenenko Kate Park Sean Hendryx Dylan Slack, Jean Wang. A holistic approach for test and evaluation of large language models. Technical report, Scale AI, 2023. URL https://static.scale.com/uploads/6019a18f03a4ae003acb1113/test-and-evaluation.pdf.

Apple Security Engineering and Architecture (SEAR). Security research on private cloud compute, 2024. URL https://security.apple.com/blog/pcc-security-research/.

Apple Security Engineering, Core Operating Systems (Core OS) Services Engineering (ASE) Architecture (SEAR), User Privacy, Machine Learning, and AI (AIML). Private cloud compute: A new frontier for ai privacy in the cloud, 2024. URL https://security.apple.com/blog/private-cloud-compute/.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Daniel Fabian and Jacob Crisp. Why red teams play a central role in helping organizations secure ai systems. Technical report, Google, July 2023. URL https://services.google.com/fh/files/blogs/google_ai_red_team_digital_final.pdf. Accessed: 2025-05-24.

Ned Finkle. Nvidia lends support to washington's efforts to ensure ai safety, 2023. URL https://blogs.nvidia.com/blog/ai-safety-washington/.

Frontier Model Forum. Frontier capability assessments, 2025a. URL https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/.

Frontier Model Forum. About, 2025b. URL https://www.frontiermodelforum.org/about-us/.

Frontier Model Forum. Introducing the fmf's technical report series on frontier ai frameworks, 2025c. URL https://www.frontiermodelforum.org/updates/introducing-the-fmfs-technical-report-series-on-frontier-ai-safety-frameworks/.

Frontier Model Forum. Frontier Model Forum: Advancing frontier AI safety and security, n.d. URL https://www.frontiermodelforum.org/.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024. URL https://arxiv.org/abs/2309.00770.

Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Google. AI principles 2023 progress update, 2023. URL https://ai.google/static/documents/ai-principles-2023-progress-update.pdf.

Google. Gemma prohibited use policy, 2024. URL https://ai.google.dev/gemma/prohibited_use_policy.

Google. How google protects its production services, 2024. URL https://cloud.google.com/docs/security/production-services-protection.

Google. Gemma model card, 2025. URL https://ai.google.dev/gemma/docs/core/model_card. Accessed: 2025-05-24.

Google. Fulfilling the voluntary industry commitments on AI, n.d. URL https://static.googleusercontent.com/media/publicpolicy.google/en//resources/whcommitments.pdf.

Google DeepMind. SynthID, n.d. URL https://deepmind.google/technologies/synthid/.

IBM Granite. Responsible use guide, 2024. URL https://www.ibm.com/granite/docs/resources/responsible-use-guide.pdf.

Group of Seven. Hiroshima process international code of conduct for organizations developing advanced ai syste, October 2023. URL https://www.mofa.go.jp/files/100573473.pdf.

Sean O hEigeartaigh, Yolanda Lannquist, Alexandru Marcoci, Jaime Sevilla, Mónica Alejandra Ulloa Ruiz, Yaqub Chaudhary, Tim Schreier, Zach Stein-Perlman, and Jeffrey Ladish. Do companies' ai safety policies meet government best practice? https://www.lcfi.ac.uk/news-events/news/ai-safety-policies, 2023.

Melissa Heikkilä. AI companies promised to self-regulate one year ago. What's changed?, 2024. URL https://www.technologyreview.com/2024/07/22/1095193/ai-companies-promised-the-white-house-to-self-regulate-one-year-ago-whats-changed/.

Teresa Hutson. Microsoft and openai launch societal resilience fund, 2024. URL https://blogs.microsoft.com/on-the-issues/2024/05/07/societal-resilience-fund-open-ai/.

IBM. Ai alliance launches as an international community of leading technology developers, researchers, and adopters collaborating together to advance open, safe, responsible ai, 2023. URL https://newsroom.ibm.com/AI-Alliance-Launches-as-an-International-Community-of-Leading-Technology-Developers,-Researchers,-and-Adopters-Collaborating-Together-to-Advance-Open,-Safe,-Responsible-AI.

IBM. Ibm furthers commitment to climate action through new sustainability projects and free training in green and technology skills for vulnerable communities, 2023a. URL https://newsroom.ibm.com/2023-11-16-IBM-Furthers-Commitment-to-Climate-Action-Through-New-Sustainability-Projects-and-Free-Training-in-Green-and-Technology-Skills-for-Vulnerable-Communities.

IBM. Ibm commits to train 2 million in artificial intelligence in three years, with a focus on underrepresented communities, 2023b. URL https://newsroom.ibm.com/2023-09-18-IBM-Commits-to-Train-2-Million-in-Artificial-Intelligence-in-Three-Years,-with-a-Focus-on-Underrepresented-Communities.

IBM. Ai privacy toolkit, 2024. URL https://github.com/IBM/ai-privacy-toolkit.

IBM. Ibm watsonx.ai runtime as a service, 2025a. URL https://www.ibm.com/support/customer/csol/terms/?id=i126-6883&lc=en. Accessed: 2025-05-24.

IBM. Ibm watsonx.governance, 2025b. URL https://www.ibm.com/products/watsonx-governance. Accessed: 2025-05-24.

IBM. Ibm verify privileged identity, n.d. URL https://www.ibm.com/products/verify-privilege.

IBM Research. Granite-3.0-8b-instruct, 2024. URL https://huggingface.co/ibm-granite/granite-3.0-8b-instruct. Accessed: 2025-05-24.

Inflection AI. Privacy policy, 2023a. URL https://pi.ai/policy.

Inflection AI. Our policy on frontier safety, 2023b. URL https://inflection.ai/frontier-safety.

Inflection AI. The precautionary principle: partnering with the white house on ai safety, 2023c. URL https://inflection.ai/blog/partnering-with-the-white-house-on-ai-safety.

Inflection AI. Inflection ai developer documentation, 2025. URL https://developers.inflection.ai/docs. Accessed: 2025-05-24.

Amazon Artificial General Intelligence. The amazon nova family of models: Technical report and model card, 2024. URL https://assets.amazon.science/b0/2b/e74dd4f84f188701fd06792670e7/the-amazon-nova-family-of-models-technical-report-and-model-card.pdf.

Humane Intelligence. Generative ai red teaming challenge, 2023. URL https://www.humane-intelligence.org/grt. Accessed: 2025-01-22.

ISED Canada. Voluntary code of conduct on the responsible development and management of advanced generative ai systems, September 2023. URL https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

Susan Jasper. An update on our child safety efforts and commitments, 2024. URL https://blog.google/technology/safety-security/an-update-on-our-child-safety-efforts-and-commitments/.

Maggie Johnson. Google.org announces new AI funding for students and educators, 2024. URL https://blog.google/outreach-initiatives/google-org/google-ai-initiatives-funding-educators-students/.

Adam Jones. How are AI companies doing with their voluntary commitments on vulnerability reporting? https://adamjones.me/blog/ai-vulnerability-reporting/, n.d.

David Reber Jr. Six steps toward ai security, 2023. URL https://blogs.nvidia.com/blog/ai-security-steps/.

Kevin Klyman. Acceptable use policies for foundation models, 2024. URL https://arxiv.org/abs/2409.09041.

Lauren Kogen. From statistics to stories: Indices and indicators as communication tools for social change. *The International Journal of Press/Politics*, 29(4):1090–1108, 2024. doi: 10.1177/19401612221094246. URL https://doi.org/10.1177/19401612221094246.

Nihal Krishan. Microsoft rolls out generative ai roadmap for government services, 2023. URL https://fedscoop.com/microsoft-rolls-out-generative-ai-roadmap-for-government-services/.

Lakera Team. Lakera and cohere set the bar for new enterprise llm security standards, 2024. URL https://www.lakera.ai/news/lakera-and-cohere-set-the-bar-for-new-enterprise-llm-security-standards.

Royal Hansen Laurie Richardson. Acting on our commitment to safe and secure ai, 2023. URL https://blog.google/technology/safety-security/google-ai-security-expansion/.

Anthony Liguori and Colm MacCárthaigh. A secure approach to generative ai with aws, 2024. URL https://aws.amazon.com/blogs/machine-learning/a-secure-approach-to-generative-ai-with-aws/.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. A safe harbor for ai evaluation and red teaming. *ArXiv*, abs/2403.04893, 2024. URL https://api.semanticscholar.org/CorpusID:268297113.

Marah Abdin et al. Phi-4 technical report. Technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Kim Martineau. What is red teaming for generative ai, 2024. URL https://research.ibm.com/blog/what-is-red-teaming-gen-AI.

Meta. Meta's ai learning alliance aims to expand pipeline of underrepresented students in ai, 2022a. URL https://ai.meta.com/blog/ai-learning-alliance/.

Meta. How ai is helping address the climate crisis, 2022b. URL https://ai.meta.com/blog/how-ai-is-helping-address-the-climate-crisis/.

Meta. Announcing purple llama: Towards open trust and safety in the new world of generative ai, 2023a. URL https://ai.meta.com/blog/purple-llama-open-trust-safety-generative-ai/.

Meta. Overview of meta ai safety policies prepared for the uk ai safety summit, 2023b. URL https://transparency.meta.com/en-gb/policies/ai-safety-policies-for-safety-summit.

Meta. Generative ai privacy guide, 2024a. URL https://www.facebook.com/privacy/guide/genai. Accessed: 2025-05-24.

Meta. Labeling ai-generated images on facebook, instagram and threads, February 2024b. URL https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/. Accessed: 2025-05-24.

Meta. Proactive detection of voice cloning with localized watermarking, 2024a. URL https://ai.meta.com/research/publications/proactive-detection-of-voice-cloning-with-localized-watermarking/.

Meta. Expanding our open source large language models responsibly, 2024b. URL https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/.

Meta. Privacy progress, n.d. URL https://about.meta.com/privacy-progress/. Accessed: 2025-05-24.

Meta. Meta bug bounty program, n.d. URL https://bugbounty.meta.com/.

Meta AI. Stable signature: A new method for watermarking images created by open-source generative ai models, October 2023. URL https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/. Accessed: 2025-05-24.

Meta AI. Llama 3.3 model card, December 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-05-24.

Microsoft. An update prepared for the uk ai safety summit, 2023. URL https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/.

Microsoft. Introducing the watermark algorithm for synthetic voice, 2024a. URL https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/introducing-the-watermark-algorithm-for-synthetic-voice/ba-p/3298548.

Microsoft. Microsoft joins thorn and all tech is human to enact strong child safety commitments for generative ai, 2024b. URL https://blogs.microsoft.com/on-the-issues/2024/04/23/microsoft-thorn-all-tech-is-human-child-safety-generative-ai/.

Microsoft. Accelerate your federal ai journey with microsoft, 2024c. URL https://www.microsoft.com/en-us/federal/artificialintelligence.

Microsoft. Phi-4 model card, 2024d. URL https://huggingface.co/microsoft/phi-4.

Microsoft. Microsoft trustworthy ai: Unlocking human potential starts with trust, 2024e. URL https://blogs.microsoft.com/blog/2024/09/24/microsoft-trustworthy-ai-unlocking-human-potential-starts-with-trust/.

Microsoft. Responsible ai transparency report. Technical report, Microsoft, 2024f. URL https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/responsible-aI-transparency-report-2024.pdf.

Microsoft. Microsoft copilot bounty program, 2025. URL https://www.microsoft.com/en-us/msrc/bounty-ai. Accessed: 2025-05-24.

Microsoft. Microsoft education ai toolkit. Technical report, 2025. URL https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-product-and-services/microsoft-education/downloadables/Microsoft-Education-AI-Toolkit1.pdf.

Microsoft Research. Ai for health, n.d.a. URL https://www.microsoft.com/en-us/research/project/ai-for-health/.

Microsoft Research. The prompt: with trevor noah, n.d.b. URL https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/the-prompt/.

MLCommons. Mlcommons announces the formation of ai safety working group, 2023. URL https://mlcommons.org/2023/10/mlcommons-announces-the-formation-of-ai-safety-working-group/.

Gary D. Lopez Munoz, Amanda J. Minnich, Roman Lutz, Richard Lundeen, Raja Sekhar Rao Dheekonda, Nina Chikanov, Bolor-Erdene Jagdagdorj, Martin Pouliot, Shiven Chawla, Whitney Maxwell, Blake Bullwinkel, Katherine Pratt, Joris de Gruyter, Charlotte Siska, Pete Bryan, Tori Westerhoff, Chang Kawaguchi, Christian Seifert, Ram Shankar Siva Kumar, and Yonatan Zunger. Pyrit: A framework for security risk identification and red teaming in generative ai systems, 2024. URL https://arxiv.org/abs/2410.02828.

Arun Murthy. Scale unlocks open-source llms with new platform and partnership with meta, 2023. URL https://scale.com/blog/custom-llms.

Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry-Alexander Bradley, and Jeff Alstott. *Securing AI model weights: Preventing theft and misuse of frontier models*. Number 1. Rand Corporation, 2024.

NIST. NIST AI safety institute consortium members, 2024. URL https://www.nist.gov/aisi/aisic-members.

NIST. U.s. ai safety institute signs agreements regarding ai safety research, testing and evaluation with anthropic and openai, 2024. URL https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research.

Nvidia. Nemotron-4-340b-base model card, 2024a. URL https://huggingface.co/nvidia/Nemotron-4-340B-Base.

Nvidia. Nemotron-4-340b-instruct model card, 2024b. URL https://huggingface.co/nvidia/Nemotron-4-340B-Instruct.

Nvidia. Nvidia open model license agreement, 2024c. URL https://developer.download.nvidia.com/licenses/nvidia-open-model-license-agreement-june-2024.pdf.

Nvidia. Nemotron-4-340b technical report. *arXiv preprint*, 2024d. URL https://arxiv.org/pdf/2406.11704.

Nvidia. Nvidia morpheus, n.d. URL https://developer.nvidia.com/morpheus-cybersecurity.

Partnership on AI Staff. Pai welcomes four new partners: Ada lovelace, ey, inflection ai, and north american broadcasters association, 2024. URL https://partnershiponai.org/pai-welcomes-four-new-partners-ada-lovelace-ey-inflection-ai-and-north-american-broadcasters-association/.

OpenAI. Our approach to ai safety, 2023a. URL https://openai.com/index/our-approach-to-ai-safety/.

OpenAI. Announcing openai's bug bounty program, 2023b. URL https://openai.com/index/bug-bounty-program/.

OpenAI. Our approach to frontier risk, 2023c. URL https://openai.com/global-affairs/our-approach-to-frontier-risk/.

OpenAI. Openai red teaming network, 2023d. URL https://openai.com/index/red-teaming-network/.

OpenAI. Openai board forms safety and security committee, 2024a. URL https://openai.com/index/openai-board-forms-safety-and-security-committee/.

OpenAI. Openai o1 system card. Technical report, 2024b. URL https://arxiv.org/abs/2412.16720.

OpenAI. Consumer privacy at openai, 2024c. URL https://openai.com/consumer-privacy/.

OpenAI. Openai safety update, 2024d. URL https://openai.com/index/openai-safety-update/.

OpenAI. Securing research infrastructure for advanced ai, 2024e. URL https://openai.com/index/securing-research-infrastructure-for-advanced-ai/.

OpenAI. Openai usage policies, 2024f. URL https://openai.com/policies/usage-policies/.

OpenAI. Understanding the source of what we see and hear online, 2024g. URL https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/.

OpenAI. Color health, n.d. URL https://openai.com/index/color-health/.

Sarah Tan Orlando Lugo. Ethical hacking practices prove successful in building trusted ai products, 2024. URL https://www.salesforce.com/blog/ethical-hacking-practices-ensures-trust/.

Palantir. Palantir technologies' approach to ai ethics, 2023. URL https://www.palantir.com/pcl/palantir-ai-ethics/.

Palantir. Palantir & ncmec, 2024a. URL https://www.palantir.com/ncmec/.

Palantir. Palantir's response to omb on ai governance, innovation, and risk management, 2024b. URL https://blog.palantir.com/palantirs-response-to-omb-on-ai-governance-innovation-and-risk-management-1e2be610a6e9.

Palantir. Human rights and technology, 2024c. URL https://blog.palantir.com/human-rights-and-technology-38580c5ac379.

Palantir. Palantir and green energy pioneer tes forge long-term partnership to drive global decarbonization, 2024d. URL https://investors.palantir.com/news-details/2024/Palantir-and-Green-Energy-Pioneer-TES-Forge-Long-Term-Partnership-to-Drive-Global-Decarbonization/.

Palantir. Artificial intelligence platform now, n.d.a. URL https://aip.palantir.com/.

Palantir. Aip, n.d.b. URL https://www.palantir.com/platforms/aip/.

Palantir. The palantir future global scholarship program, n.d.c. URL https://www.palantir.com/careers/students/scholarship/future-global/.

Palantir. Palantir terms and conditions, n.d.d. URL https://www.palantir.com/assets/xrfr7uokpv1b/3MWySPWdUTfClioIasXFqU/fa3297a0a6cbc87e206caa7f2e5e293a/Palantir_Connector_Terms_of_Service.pdf.

Will Pearce and Joseph Lucas. Nvidia ai red team: An introduction, 2023. URL https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction/.

Vasi Philomin. A progress update on our commitment to safe, responsible generative ai, 2024. URL https://aws.amazon.com/blogs/machine-learning/a-progress-update-on-our-commitment-to-safe-responsible-generative-ai/.

Danilo Poccia. Introducing amazon nova foundation models: Frontier intelligence and industry leading price performance, 2024. URL https://aws.amazon.com/blogs/aws/introducing-amazon-nova-frontier-intelligence-and-industry-leading-price-performance/.

Darius Rafieyan. Openai is hiring someone to investigate its own employees, 2024. URL https://www.businessinsider.com/openai-is-hiring-someone-to-investigate-its-own-employees-2024-8.

AI Red Team Ram Shankar Siva Kumar, Data Cowboy. Announcing microsoft's open automation framework to red team generative ai systems, 2024. URL https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/.

Ranking Digital Rights. 2020 ranking digital rights corporate accountability index, 2020. URL https://rankingdigitalrights.org/index2020/.

Dana Rao. Building safe, secure, and trustworthy ai: Adobe's commitments to our customers and community, 2023. URL https://blog.adobe.com/en/publish/2023/09/12/adobes-ai-commitments-to-customers-and-community.

Dana Rao. The road ahead for responsible innovation, 2024. URL https://blog.adobe.com/en/publish/2024/03/26/the-road-ahead-responsible-innovation.

IBM Research. Granite 3.0 technical report, 2024. URL https://www.ibm.com/downloads/documents/us-en/10a99803c92fdb35.

Kevin Roose. How do the white house's a.i. commitments stack up?, 2023. URL https://www.nytimes.com/2023/07/22/technology/ai-regulation-white-house.html.

Salesforce. Xgen-7b technical report. 2023. URL https://arxiv.org/abs/2309.03450.

Salesforce. Tracking our progress on the white house voluntary ai commitments, 2024a. URL https://www.salesforce.com/news/stories/voluntary-ai-commitments/.

Salesforce. Salesforce einstein model cards. Technical report, Salesforce, 2024b. URL https://resources.docs.salesforce.com/latest/latest/en-us/sfdc/pdf/salesforce_ai_model_cards.pdf.

Salesforce. Salesforce gives $23m to education to help the ai generation unlock critical skills, 2024c. URL https://www.salesforce.com/news/press-releases/2024/09/16/ai-education-grants-dreamforce-24/.

Scale AI. Scale acceptable use policy, 2022a. URL https://scale.com/legal/aup.

Scale AI. Scale ai providing free datasets to help national security partners gain current insights into russia-ukraine conflict, 2022b. URL https://scale.com/blog/ukraine-detection.

Scale AI. Seal: Scale's safety, evaluations and alignment lab, 2023a. URL https://scale.com/blog/safety-evaluations-alignment-lab.

Scale AI. Test and evaluation white paper, 2023b. URL https://scale.com/blog/test-and-evaluation-white-paper.

Scale AI. Defense llama: The llm purpose-built for american national security, 2024a. URL https://scale.com/blog/defense-llama.

Scale AI. Submit your toughest questions for humanity's last exam, 2024b. URL https://scale.com/blog/humanitys-last-exam.

Scale AI. Responsible ai with scale evaluation for the public sector, 2024c. URL https://scale.com/blog/responsible-ai-scale-evaluation-for-public-sector.

Scale AI. Fine-tuned llms for defense, n.d.a. URL https://scale.com/donovan/defense-llm.

Scale AI. Seal leaderboards, n.d.b. URL https://scale.com/leaderboard.

Scale AI. Scale evaluation, n.d.c. URL https://scale.com/evaluation/model-developers.

Scale AI. Test and evaluation vision, n.d.d. URL https://scale.com/guides/test-and-evaluation-vision.

Manjer Shaikh. Building the future of ai security: The ibm and celerity partnership, 2024. URL https://www.ibm.com/blogs/think/uk-en/building-the-future-of-ai-security-the-ibm-and-celerity-partnership/.

Stability AI. Stability ai previews enhanced image offerings: Apis for business & new product features, 2023a. URL https://stability.ai/news/stability-ai-enhanced-image-apis-for-business-features.

Stability AI. Statement to the u.s. senate ai insight forum on transparency, explainability, and copyright, 2023b. URL https://stability.ai/news/copyright-us-senate-open-ai-transparency.

Stability AI. Stable diffusion 3.5 large model card, 2024a. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large.

Stability AI. Response to request for information on artificial intelligence from the national institute of standards and t echnology, 2024b. URL https://downloads.regulations.gov/NIST-2023-0009-0174/attachment_1.pdf.

Stability AI. Acceptable use policy, 2024c. URL https://stability.ai/use-policy.

Stability AI. Stability ai joins thorn and all tech is human to enact child safety commitments for generative ai, 2024. URL https://stability.ai/news/safetybydesign.

Stability AI. Collaboration is key to safe innovation, n.d.a. URL https://stability.ai/safety-commitments-and-collaboration.

Stability AI. Stable safety, n.d.b. URL https://stability.ai/safety.

Louis Stewart. Golden opportunities: California to train students, educators in ai, 2024. URL https://blogs.nvidia.com/blog/california-ai-education/.

HUG Studios. Innovation laboratory ii, 2024. URL https://www.studios.thehug.xyz/lab.

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0), 2023-01-26 05:01:00 2023. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

Cohere For AI Team. Granting access: Supporting researchers to use llms, 2024a. URL https://cohere.com/blog/granting-access.

Cohere For AI Team. Cohere for ai scholars program: Research journeys start here, 2024b. URL https://cohere.com/blog/cohere-for-ai-scholars-program-2025.

The Executive Office of the President. Advancing united states leadership in artificial intelligence infrastructure, January 2025. URL https://www.federalregister.gov/documents/2025/01/17/2025-01395/advancing-united-states-leadership-in-artificial-intelligence-infrastructure. Executive Order 14141, Federal Register Document 2025-01395, 90 FR 5469–5489.

Thorn. Thorn launches initiative to eliminate child sexual abuse material from the internet, 2019. URL https://www.prnewswire.com/news-releases/thorn-launches-initiative-to-eliminate-child-sexual-abuse-material-from-the-internet-300833501.html.

All Tech Is Human Thorn. Safety by design for generative ai: Preventing child sexual abuse, 2024. URL https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

Tom Gunter et al. Apple intelligence foundation language models, 2024. URL https://arxiv.org/abs/2407.21075.

U.S. AI Safety Institute. Managing misuse risk for dual-use foundation models, July 2024. URL https://doi.org/10.6028/NIST.AI.800-1.ipd.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. Locally differentially private document generation using zero-shot prompting, October 2023. URL https://cohere.com/research/papers/locally-differentially-private-document-generation-using-zero-shot-prompting-2023-10-24. Accessed: 2025-05-24.

Daniel Ventura. Adobe collaborates with ethical hackers to build safer, more secure ai tools, 2024. URL https://blog.adobe.com/en/publish/2024/05/01/adobe-collaborates-with-ethical-hackers-build-safer-more-secure-ai-tools.

White House. Ensuring Safe, Secure, and Trustworthy AI. https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf, 2023.

White House. White house announces new private sector voluntary commitments to combat image-based sexual abuse. https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/, 2024.

Summer Yue and Daniel Berrios. Introducing wmdp: Measuring and mitigating catastrophic risk potential from llms, 2024. URL https://scale.com/blog/measuring-mitigating-risk-wmdp.

Federico Zarfati. Announcing new ai safety & responsible ai features in azure openai service at ignite 2023, 2023. URL https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/announcing-new-ai-safety-amp-responsible-ai-features-in-azure/ba-p/3983686.

Commitment	Indicator
Red-teaming	Internal red-teaming
	External red-teaming
	Red teaming coverage of risks
Information Sharing	Information sharing with companies
	Information sharing with government
	Forum or mechanism for information sharing
	Forum or mechanism shares information on risks
Model weight security	Model weight cybersecurity practices
	Insider threat detection program
	Limiting weight-level access to relevant personnel
Third-party reporting	Establish bounties, contests, or prizes
	Include AI systems in their existing bug bounty programs
Watermarking/Provenance	Robust provenance or watermarking for audio
	Robust provenance or watermarking for visual content
	Develop tools or APIs to determine if a particular piece of content was created within their tools
	Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI
Public reporting	Report capabilities
	Report limitations
	Report domains of appropriate use
	Report domains of inappropriate use
	Report safety evaluations
	Report on societal risks
	Report on adversarial testing used to determine appropriateness of deployment
Societal risk research	Empower trust and safety teams
	Advance AI safety research
	Advance privacy
	Protect children
Address society's greatest challenges	Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats.
	Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI
	Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology.

Figure 2: Indicators. Table of the 30 indicators we use to score companies.

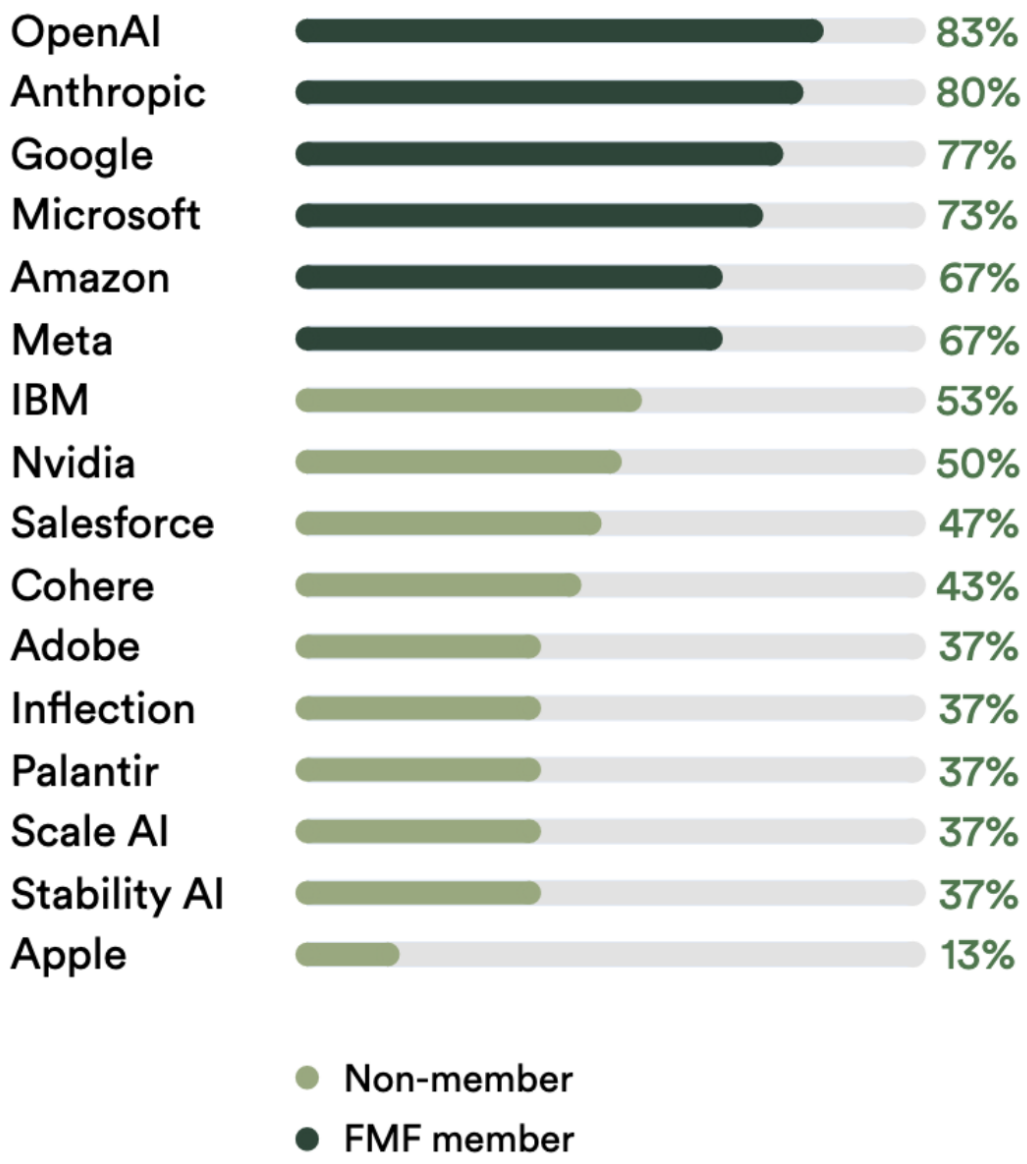

Figure 3: Aggregate scores by company. The score for each company, stratified by whether the company belongs to the Frontier Model Forum (FMF) as of December 31, 2024.

	Adobe	Amazon	Anthropic	Apple	Cohere	Google	IBM	Inflection	Meta	Microsoft	Nvidia	OpenAI	Palantir	Scale AI	Salesforce	Stability AI	Average
Red-teaming	0%	33%	100%	0%	33%	67%	0%	0%	67%	33%	67%	100%	0%	33%	0%	33%	35%
Information Sharing	25%	75%	100%	25%	25%	75%	50%	25%	75%	75%	25%	100%	25%	25%	25%	50%	50%
Model weight security	0%	33%	67%	0%	0%	33%	0%	0%	0%	67%	0%	67%	0%	0%	0%	0%	17%
Third-party reporting	50%	50%	100%	0%	0%	50%	0%	50%	50%	100%	0%	100%	0%	0%	0%	0%	34%
Watermarking/Provenance	100%	100%	100%	0%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	75%	92%
Public reporting	43%	86%	100%	14%	71%	86%	86%	43%	71%	71%	86%	71%	43%	43%	57%	57%	64%
Societal risk research	25%	75%	50%	50%	50%	100%	50%	50%	75%	75%	25%	75%	50%	25%	75%	25%	55%
Address society's greatest challenges	33%	33%	0%	0%	0%	67%	67%	0%	67%	67%	33%	67%	33%	33%	67%	0%	35%
	35%	61%	77%	11%	35%	72%	44%	33%	63%	74%	42%	85%	31%	32%	40%	30%	48%

Figure 4: Scores for each of the eight commitments in the 2023 White House Voluntary Commitments on AI. Each cell represents a company's average score across all of the indicators for a given commitment.

Commitment	Indicator	Adobe	Amazon	Anthropic	Apple	Cohere	Google	IBM	Inflection	Meta	Microsoft	Nvidia	OpenAI	Palantir	Salesforce	Scale AI	Stability AI
Red-teaming	Internal red-teaming	0	1	1	0	0	1	0	0	1	1	1	1	0	0	0	0
	External red-teaming	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1	1
	Red teaming coverage of risks	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
Information Sharing	Information sharing with companies	0	1	1	0	0	1	1	0	1	1	0	1	0	0	0	0
	Information sharing with government	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
	Forum or mechanism for information sharing	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Forum or mechanism shares information on risks	0	1	1	0	0	1	0	0	1	1	0	1	0	0	0	1
Model weight security	Model weight cybersecurity practices	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0
	Insider threat detection program	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	Limiting weight-level access to relevant personnel	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
Third-party reporting	Establish bounties, contests, or prizes	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0
	Include AI systems in their existing bug bounty programs	1	1	1	0	0	1	0	0	1	1	0	1	0	0	0	0
Watermarking/Provenance	Robust provenance or watermarking for audio	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
	Robust provenance or watermarking for visual content	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	Develop tools or APIs to determine if a particular piece of content was created within their tools	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	Work with industry peers and standards-setting bodies	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
Public reporting	Report capabilities	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Report limitations	1	1	1	0	1	1	1	1	1	1	1	0	0	1	0	1
	Report domains of appropriate use	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1
	Report domains of inappropriate use	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	Report safety evaluations	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0	0
	Report on societal risks	0	1	1	0	1	1	1	0	0	0	1	1	0	0	0	0
	Report on adversarial testing	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0
Societal risk research	Empower trust and safety teams	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	Advance AI safety research	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0
	Advance privacy	0	1	0	1	1	1	1	0	1	1	0	1	1	1	0	0
	Protect children	1	1	1	0	0	1	0	1	1	1	0	1	1	1	0	1
Address society's greatest challenges	Support research and development for greatest challenges	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	0
	Support education and training initiatives	1	1	0	0	0	1	1	0	1	1	0	1	0	1	0	0
	Help citizens understand the technology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Commitment Subtotal	37%	67%	80%	13%	43%	77%	53%	37%	67%	73%	50%	83%	37%	47%	37%	37%

Figure 5: Per-indicator scores. The score for each (indicator, company) pair.

A Related Work

To contextualize our work, we discuss prior work that assesses major AI companies based on their public conduct and discuss other voluntary commitments.

A.1 Assessments of AI Companies for 2023 WHVC

Beyond our work, the most comprehensive analysis of company practices in relation to these commitments was conducted as part of a MIT Technology Review article published on the one-year anniversary of the commitments [Heikkilä, 2024]. As part of this work, Heikkilä [2024] contacted the seven initial signatories and received responses from six of these companies, excluding Inflection, on how they addressed each of the commitments; external researchers also provided commentary. Overall, the work found evidence that companies had taken steps to implement some technical model-level interventions(e.g. red-teaming and watermarking) and made investments in safety research. However, less evidence was found related to progress on information sharing, third-party reporting and public reporting.

Heikkilä [2024] indicates that no comprehensive evaluation had been performed of the commitments, company practices, or their relationship. In light of this, our work not only provides a comprehensive assessment, but also introduces a concrete scoring system that yields quantitative findings. In general, our findings largely agree with those of Heikkilä [2024] and Roose [2023], with the main difference being the depth and specificity of our results, though we highlight that our scores are based on public information from companies whereas the prior work only considered the brief responses companies provided to journalists. Further, our work expands the focus to the full set of 16 companies, rather than just the initial seven, which enables us to identify clear disparities between the initial signatories and the remaining signatories.

A.2 Assessments of AI Companies

As technology companies have grown in importance and become some of the world's most powerful entities, a multidisciplinary body of literature has emerged to assess these companies with a variety of methods. In the space of quantitative assessments, several works have introduced scoring approaches either in the form of one-off analyses, akin to this work, or sustained indices, which score the same companies on a recurring cadence. As an illustrative example, we highlight the Corporate Accountability Index that is maintained by Ranking Digital Rights (RDR), which has scored telecommunication and technology companies since 2015 for how they "respect users' fundamental rights, and on the mechanisms they have in place to ensure those promises are kept" [Ranking Digital Rights, 2020]. Kogen [2024] analyzed the 2018 Index and showed, by reviewing internal RDR documents and interviewing relevant stakeholders (e.g. representatives from 11 companies and 14 civil society groups), that it usefully communicated legible, newsworthy, and flexible information that empowered social movements.

Drawing upon this tradition, several recent works have employed and developed similar scoring approaches for the assessment of AI companies [Bommasani et al., 2023b, 2024, Klyman, 2024, Longpre et al., 2024, AI Lab Watch, n.d., hEigeartaigh et al., 2023, Barrett et al., 2023, Jones, n.d.]. To our knowledge, Bommasani et al. [2023b] provided the first assessment of major AI companies by scoring them on a rubric based on the European Parliament's proposal for the EU AI Act. Based on the results, they made evidence-based recommendations aimed at (i) EU legislators on how the EU AI Act should be updated during the legislative negotiation and (ii) companies on how they could modify their practices to better align with the proposed requirements. While some works similarly link scoring to specific governmental policies (e.g. Barrett et al. [2023] assess companies in relation to NIST's AI Risk Management Framework, hEigeartaigh et al. [2023] score in relation to the UK's recommendations), other works provide independent specification of the indicators or criteria of interest. The Foundation Model Transparency Index is an annual index that scores foundation model developers for their transparency across the supply chain with 100 indicators that span

the resources used to build a model (e.g. data, compute), the properties of the model itself (e.g. capabilities, risks), and the use of the model in society (e.g. distribution, impact) [Bommasani et al., 2023a, 2024].

Cumulatively, these works all demonstrate a shared methodology of scoring companies with different approaches for sourcing the indicators, determining the scores, and theories of change for how the results and takeaways improve corporate governance and/or public policy. Many of these works also share two key findings with our work. While all of these works aim to increase public accountability, they all encounter limits due to the lack of transparency into company-internal practices. And, while the exact magnitudes and details often differ, these works almost always find considerable heterogeneity in company practices. Together, they highlight the absence of clear norms, let alone more formal mechanisms, for ensuring public-facing transparency and standardizing industry-wide conduct.

A.3 Voluntary Commitments From Governments

Global AI policy reflects a broad constellation of efforts that spans long-standing policy in specific domains (e.g. applying hiring discrimination laws to algorithmic hiring), more recent policy for digital technologies (e.g. applying data protection laws to training data), and new policy for AI specifically (e.g. new laws to govern AI). While many jurisdictions face shared challenges, the overall global AI policy landscape reflects significant heterogeneity that indicates both region-specific considerations and idiosyncratic differences. In particular, when considering AI-specific policy, several jurisdictions currently employ voluntary approaches to corporate governance with the European Union's approach via the EU AI Act standing as a clear counter example. At this juncture, given many of these voluntary and/or mandatory policies are very recent, little evidence exists to empirically validate the strengths and/or weaknesses of these two top-level approaches.

As a result, we briefly survey some of the voluntary commitments and approaches taken elsewhere in the world to contextualize the approach taken in the 2023 White House Voluntary Commitments on AI. The U.S. NIST AI Risk Management Framework, as well as the associated profile on generative AI in particular, provides voluntary guidance to help organizations identify, assess, manage, and mitigate risks by emphasizing trustworthy AI principles such as fairness, transparency, accountability, security, and privacy [Tabassi, 2023]. The Canada Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems introduces voluntary commitments applicable to the responsible development and deployment of foundation models, such as accountability, safety, fairness, human oversight, and robustness, as well as for developers and managers of generative AI systems [ISED Canada, 2023]. The G7 International Code of Conduct for Organizations Developing Advanced AI Systems articulates 11 commitments that span data protection, risk management, technical standard, and transparency reporting: while companies not signed on in the same way they have done for certain national-level commitments, these commitments may serve as the basis for global agreement [Group of Seven, 2023]. Most recently, the Biden-Harris Administration secured voluntary commitments with AI model developers and data providers to prevent and mitigate the misuse of AI in creating and disseminating image-based sexual abuse content [White House, 2024].

Beyond these standalone voluntary commitments, the ongoing series of international AI Summits have emerged as a key generative process for voluntary commitments as global policymakers work together to advance AI governance. Beginning with the U.K. AI Safety Summit in November 2023, the Bletchley Declaration was signed by 29 world governments to foster international cooperation on AI policy through an agenda centered on (i) "identifying AI safety risks of shared concern, building a shared scientific and evidence-based understanding of these risks, and sustaining that understanding ..." as well as (ii) "building respective risk-based policies across our countries to ensure safety ... alongside increased transparency by private actors developing frontier AI capabilities, appropriate evaluation metrics, tools for safety testing, and developing relevant public sector capability and scientific research". To advance this agenda, at the subsequent AI Seoul Summit in May 2024, 16 global companies (Amazon, Anthropic, Cohere, Google, G42, IBM, Inflection AI, Meta, Microsoft, Mistral AI, Naver, OpenAI, Samsung Electronics, Technology Innovation Institute, xAI, Zhipu.ai) signed onto the Frontier AI Safety Commitments. The associated eight

commitments address three outcomes: (i) improved risk management practices, (ii) increased accountability for safe development and deployment and (iii) sufficient transparency to external stakeholders. Building on these efforts, the United States convened the growing global network of AI Safety Institutes in November 2024 for a working meeting on three high-priority topics (managing risks from synthetic content, testing foundation models, and conducting risk assessments for advanced AI systems) that articulated six principles for risk assessment (actionability, transparency, comprehensiveness, multi-stakeholder consideration, iterativity, and reproducibility).

B Indicators

1. Internal red-teaming
    - Indicator under Commitment 1 on Red Teaming
    - Reference text from 2023 WHVC: "Commit to internal ... red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas."
2. External red-teaming
    - Indicator under Commitment 1 on Red Teaming
    - Reference text from 2023 WHVC: "Companies commit to ... developing a multi-faceted, specialized, and detailed red-teaming regime, including drawing on independent domain experts, for all major public releases of new models within scope."
3. Red teaming coverage of risks
    - Indicator under Commitment 1 on Red Teaming
    - Reference text from 2023 WHVC: "In designing the regime, they will ensure that they give significant attention to the following:
        - Bio, chemical, and radiological risks, such as the ways in which systems can lower barriers to entry for weapons development, design, acquisition, or use
        - Cyber capabilities, such as the ways in which systems can aid vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system
        - The effects of system interaction and tool use, including the capacity to control physical systems
        - The capacity for models to make copies of themselves or 'self-replicate'
        - Societal risks, such as bias and discrimination"
4. Information sharing with companies
    - Indicator under Commitment 2 on Information Sharing
    - Reference text from 2023 WHVC: "Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards"
    - Notes: information shared with companies should be information beyond public disclosure.
5. Information sharing with government
    - Indicator under Commitment 2 on Information Sharing
    - Reference text from 2023 WHVC: "Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards"
6. Forum or mechanism for information sharing
    - Indicator under Commitment 2 on Information Sharing
    - Reference text from 2023 WHVC: "They commit to establish or join a forum or mechanism through which they can develop, advance, and adopt shared standards and best practices for frontier AI safety, such as the NIST AI Risk Management Framework or future standards related to red-teaming, safety, and societal risks"
7. Forum or mechanism shares information on risks
    - Indicator under Commitment 2 on Information Sharing

- Reference text from 2023 WHVC: “The forum or mechanism can facilitate the sharing of information on advances in frontier capabilities and emerging risks and threats, such as attempts to circumvent safeguards, and can facilitate the development of technical working groups on priority areas of concern.”
- Notes: A forum that facilitates this kind of information must be a forum that restricts who can join and what they do that with the shared information.

8. Model weight cybersecurity practices
    - Indicator under Commitment 3 on Model Weight Security
    - Reference text from 2023 WHVC: “In addition, it requires storing and working with the weights in an appropriately secure environment to reduce the risk of unsanctioned release.”
9. Insider threat detection program
    - Indicator under Commitment 3 on Model Weight Security
    - Reference text from 2023 WHVC: “This includes ... establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets.”
10. Limiting weight-level access to relevant personnel
    - Indicator under Commitment 3 on Model Weight Security
    - Reference text from 2023 WHVC: “This includes limiting access to model weights to those whose job function requires it...”
11. Establish bounties, contests, or prizes
    - Indicator under Commitment 4 on Third-Party Reporting
    - Reference text from 2023 WHVC: “They commit to establishing for systems within scope bounties systems, contests, or prizes to incent the responsible disclosure of weaknesses, such as unsafe behaviors...”
12. Include AI systems in their existing bug bounty programs
    - Indicator under Commitment 4 on Third-Party Reporting
    - Reference text from 2023 WHVC: “They commit ... to include AI systems in their existing bug bounty programs”
13. Robust provenance or watermarking for audio
    - Indicator under Commitment 5 on Content Provenance
    - Reference text from 2023 WHVC: “To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed.”
14. Robust provenance or watermarking for visual content
    - Indicator under Commitment 5 on Content Provenance
    - Reference text from 2023 WHVC: “To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed.”
15. Develop tools or APIs to determine if a particular piece of content was created within their tools
    - Indicator under Commitment 5 on Content Provenance
    - Reference text from 2023 WHVC: “They will also develop tools or APIs to determine if a particular piece of content was created with their system.”

16. Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI
    - Indicator under Commitment 5 on Content Provenance
    - Reference text from 2023 WHVC: "More generally, companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI."
17. Report capabilities
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias."
18. Report limitations
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "These reports should include ... significant limitations in performance that have implications for the domains of appropriate use..."
    - Notes: Limitations must be specific to the model and not to AI generally.
19. Report domains of appropriate use
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "Publicly report ... domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias"
20. Report domains of inappropriate use
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "Publicly report ... domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias"
21. Report safety evaluations
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "These reports should include the safety evaluations conducted (including in areas such as dangerous capabilities, to the extent that these are responsible to publicly disclose) ..."
22. Report on societal risks
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "These reports should include ... discussion of the model's effects on societal risks such as fairness and bias ..."
23. Report on adversarial testing used to determine appropriateness of deployment
    - Indicator under Commitment 6 on Public Reporting
    - Reference text from 2023 WHVC: "These reports should include ... the results of adversarial testing conducted to evaluate the model's fitness for deployment."
24. Empower trust and safety teams
    - Indicator under Commitment 7 on Societal Risk Research
    - Reference text from 2023 WHVC: "Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized."

25. Advance AI safety research
    - Indicator under Commitment 7 on Societal Risk Research
    - Reference text from 2023 WHVC: “Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized.”
26. Advance privacy
    - Indicator under Commitment 7 on Societal Risk Research
    - Reference text from 2023 WHVC: “Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized.”
27. Protect children
    - Indicator under Commitment 7 on Societal Risk Research
    - Reference text from 2023 WHVC: “Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized.”
28. Support R&D of frontier AI to address society’s greatest challenges
    - Indicator under Commitment 8 on Address Society’s Greatest Challenges
    - Reference text from 2023 WHVC: “Companies making this commitment agree to support research and development of frontier AI systems that can help meet society’s greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats.”
    - Notes: There is a distinction between supporting research and development of AI in service of societal goals, compared to providing commercial services to public interest companies and advancing AI research in specific domains. The level of engagement and initiative varies.
29. Foster the education and training of students and workers to prosper from the benefits of AI
    - Indicator under Commitment 8 on Address Society’s Greatest Challenges
    - Reference text from 2023 WHVC: “Companies also commit to supporting initiatives that foster the education and training of students and workers to prosper from the benefits of AI ...”
    - Notes: The initiatives covered should be accessible to all students and works regardless of prior technical training.
30. Help citizens understand the nature, capabilities, limitations, and impact of the technology
    - Indicator under Commitment 8 on Address Society’s Greatest Challenges
    - Reference text from 2023 WHVC: “Companies also commit to ... helping citizens understand the nature, capabilities, limitations, and impact of the technology.”

C Indicator Scores for Companies

Table 2: Indicator Scores for Adobe

Indicator	Score	Justification	Source
Internal red-teaming	✗	Adobe conducts internal red teaming and penetration testing for Firefly, but it is unclear what risk areas are covered.	Ventura [2024]
External red-teaming	✗	While Adobe engages with third party security researchers, including through a bug bounty, we do not consider this as an organized initiative dedicated to external red-teaming.	Ventura [2024]
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✗	While Adobe is involved in C2PA, which involves other companies, this (to our knowledge) does not involve information beyond public information at present.	Adobe [n.d.]
Information sharing with government	✗	While Adobe conducts an annual Forum event with government, this does not entail sharing specific information about their models beyond what is public to our knowledge.	-
Forum or mechanism for information sharing	✓	Adobe is a member of the US AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members to prevent misuse or security vulnerabilities.	NIST [2024]
Model weight cybersecurity practices	✗	No relevant evidence found.	-
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	Adobe did not establish bounties but has an existing bug bounty that includes AI systems.	Ventura [2024]
Include AI systems in their existing bug bounty programs	✓	Adobe expanded their bug bounty program to include their implementation of Content Credentials and Adobe Firefly.	Ventura [2024]
Robust provenance or watermarking for audio	✓	We award this point due to the Adobe Content Authenticity web app and the associated Content Credentials. Creators can easily apply Content Credentials in batch to sign their digital work — including images, audio and video files.	Adobe [2024a]
Robust provenance or watermarking for visual content	✓	We award this point due to the Adobe Content Authenticity web app and the associated Content Credentials. Creators can easily apply Content Credentials in batch to sign their digital work — including images, audio and video files. Content Credentials are supported by Firefly.	Adobe [2024a]

Continued on next page

Indicator Scores for Adobe – *Continued from previous page*

Indicator	Score	Justification	Source
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Adobe developed the C2PA Verify tool that determines if a Content Credential was issued by a known source.	Adobe [2024a]
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Adobe formed the Content Authenticity Initiative, a global coalition of over 1,500 members across industries, united to promote trust and transparency in digital content.	Rao [2023]
Report capabilities	✓	Adobe describes model capabilities in the press release associated with Firefly Image 3.	Adobe [2024b]
Report limitations	✓	Adobe maintains a list of Firefly's known limitations.	Adobe [2024d]
Report domains of appropriate use	✗	No relevant evidence found.	-
Report domains of inappropriate use	✓	Adobe's user guidelines enumerates many prohibited uses.	Adobe [2024c]
Report safety evaluations	✗	No relevant evidence found.	-
Report on societal risks	✗	No specific information provided on societal risks, beyond the concern of deep fakes in relation to content provenance initiatives.	Rao [2024]
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✗	While Adobe discusses how they created an AI Ethics engineering team four years ago, and some of its actions, there is no clear evidence that they empower their trust & safety teams.	Rao [2024]
Advance AI safety research	✗	While Adobe conducts AI fairness research, we do not consider this as AI safety research.	Gallegos et al. [2024]
Advance privacy	✗	No relevant evidence found.	-
Protect children	✓	Adobe enforces a zero-tolerance policy on CSAM, using tools like PhotoDNA and CSAI Match to detect and block known content via hash matching.	Adobe [2023]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	No relevant evidence found.	-
Support initiatives that foster the education and training of students and workers	✓	Adobe launched a global initiative to empower 30 million next-generation learners with AI literacy, content creation, and digital marketing skills by 2030.	Adobe [2024e]
Support initiatives that help citizens understand the technology	✗	No relevant evidence found.	-

Table 3: Indicator Scores for Amazon

Indicator	Score	Justification	Source
Internal red-teaming	✓	Amazon describes that its staff conduct manual red-teaming on AI systems, including Amazon Titan.	Philomin [2024]
External red-teaming	✗	No relevant evidence found.	-
Red teaming coverage of risks	✗	While Amazon describes that it conducts multiple iterations of red-teaming on issues including safety, security, privacy, veracity, and fairness, there is no evidence that Amazon conducts testing on self-replication or the effects of system interaction and tool use.	Amazon Web Services [2024]
Information sharing with companies	✓	Amazon is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Philomin [2024]
Information sharing with government	✗	No relevant evidence found.	Philomin [2024]
Forum or mechanism for information sharing	✓	Amazon joined the U.S. AI Safety Institute Consortium.	Philomin [2024]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as described.	Philomin [2024]
Model weight cybersecurity practices	✓	Amazon describes its security practices for model weights on AWS.	Liguori and Mac-Cárthaigh [2024]
Insider threat detection program	✗	No evidence specific to Nova found.	-
Limiting weight-level access to relevant personnel	✗	No relevant evidence found indicating access is restricted to necessary personnel for training Amazon models.	-
Establish bounties, contests, or prizes	✗	No relevant evidence found.	-
Include AI systems in their existing bug bounty programs	✓	Amazon's Vulnerability Research program includes generative AI assessments and submissions.	Amazon [2025]
Robust provenance or watermarking for audio	✓	Amazon Nova does not provide audio capabilities.	Poccia [2024]
Robust provenance or watermarking for visual content	✓	Amazon has stated that every image or video generated by Nova has a digital watermark.	Philomin [2024], Poccia [2024]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Amazon cites a detection solution for identifying images created by Nova with corresponding watermarks. Amazon has previously introduced an API that detects its Titan Image Generator watermark and stated that a new API update is rolling out for Nova.	Philomin [2024], Amazon Science [2024]
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Amazon is part of the Coalition for Content Provenance and Authenticity steering committee.	C2PA [2024]

Continued on next page

Indicator Scores for Amazon – *Continued from previous page*

Indicator	Score	Justification	Source
Report capabilities	✓	Amazon reports core capabilities of Nova and provides benchmark results in its technical report for the Nova family of models.	Intelligence [2024]
Report limitations	✓	Amazon describes the limitations of Nova models in its service card.	Amazon Web Services [n.d.]
Report domains of appropriate use	✓	Amazon reports intended use cases for Nova models in its service cards.	Amazon Web Services [n.d.]
Report domains of inappropriate use	✓	Amazon specifies that its Nova models are not designed to provide legal/medical/financial opinions or advice, among other domains of inappropriate use.	Amazon Web Services [n.d.]
Report safety evaluations	✓	Amazon reports that they performed evaluations on CBRN threats.	Intelligence [2024]
Report on societal risks	✓	Amazon describes their testing on societal risks, including hate speech, political misinformation, and extremism.	Intelligence [2024]
Report on adversarial testing	✗	Amazon reports the use of human, automated, internal, and external red teaming mechanisms but not the results.	Intelligence [2024]
Empower trust and safety teams	✗	While Amazon describes training its employees on considerations around fairness, privacy, and model explainability, there is no relevant evidence of empowering its trust and safety team.	Philomin [2024]
Advance AI safety research	✓	Amazon developed and shared tools to implement safeguards, prevent harm content, and conduct safety evaluations.	Philomin [2024]
Advance privacy	✓	Amazon conducts public research on privacy and security related to generative AI, such as private text generation.	Amazon Science [n.d.]
Protect children	✓	Amazon works with Thorn to design generative AI services that reduce risk of misuse for child exploitation.	Philomin [2024]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	While Amazon provides services to businesses that build generative AI systems for drug development and energy usage optimization, these services are commercial by nature and do not represent a proactive effort to support the research and development of frontier AI systems that can help meet society's greatest challenges.	Philomin [2024]
Support initiatives that foster the education and training of students and workers	✓	Amazon launched free courses about safe and responsible AI to provide AI skills training to 2 million people by 2025.	Philomin [2024]

Continued on next page

Indicator Scores for Amazon – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that help citizens understand the technology	✗	No relevant evidence found.	-

Table 4: Indicator Scores for Anthropic

Indicator	Score	Justification	Source
Internal red-teaming	✓	Anthropic describes internally red-teaming their systems prior to deployment that cover areas, including misuse, societal risks, and national security concerns.	Anthropic [2024e]
External red-teaming	✓	Anthropic describes working with external biosecurity experts to red-team their systems. They also participated in the Generative AI Red Teaming Challenge.	Anthropic [2023b], Intelligence [2023]
Red teaming coverage of risks	✓	Anthropic described conducting red-teaming in the domains of national security, CBRN, trust and safety.	Anthropic [2024e]
Information sharing with companies	✓	Anthropic is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Frontier Model Forum [n.d.]
Information sharing with government	✓	Anthropic provides the U.S. AI Safety Institute with new models before and following their releases.	NIST [2024]
Forum or mechanism for information sharing	✓	Anthropic is a member of the U.S. AI Safety Institute Consortium and the Frontier Model Forum.	NIST [2024], Frontier Model Forum [n.d.]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as desired.	Frontier Model Forum [n.d.]
Model weight cybersecurity practices	✓	Anthropic is implementing two-party controls, secure software development framework, supply chain levels for software artifacts, and other cybersecurity best practices with the specific aim of protecting model weights.	Anthropic [2023c]
Insider threat detection program	✓	Anthropic's Responsible Scaling Policy indicates the existence of an insider risk program for ASL-2 Security Standard.	Anthropic [2023c]
Limiting weight-level access to relevant personnel	✗	Anthropic's Responsible Scaling Policy discusses access management tools under its ASL-2 Security Standard but does not specifically indicate that access is restricted to model weights.	Anthropic [2023c]
Establish bounties, contests, or prizes	✓	Anthropic operates an invite-only bug bounty program for identifying model safety issues.	Anthropic [2024b]
Include AI systems in their existing bug bounty programs	✓	Anthropic expanded its bug bounty program to include a new initiative focused on identifying and mitigating universal jailbreak attacks.	Anthropic [2024b]
Robust provenance or watermarking for audio	✓	Anthropic does not produce audiovisual models, and so we default their score to 1 for this commitment.	-

Continued on next page

Indicator Scores for Anthropic – *Continued from previous page*

Indicator	Score	Justification	Source
Robust provenance or watermarking for visual content	✓	Anthropic does not produce audiovisual models, and so we default their score to 1 for this commitment.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Anthropic does not produce audiovisual models, and so we default their score to 1 for this commitment.	-
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Anthropic does not produce audiovisual models, and so we default their score to 1 for this commitment.	-
Report capabilities	✓	Anthropic describes the capabilities of Claude 3.5 in their Claude 3.5 addendum through benchmark results.	Anthropic [2024a]
Report limitations	✓	Anthropic describes the limitations of Claude 3.5 in their Claude 3 model card under areas for improvement.	Anthropic [2024c]
Report domains of appropriate use	✓	Anthropic reports the intended uses of Claude 3.5 in their Claude 3 model card.	Anthropic [2024c]
Report domains of inappropriate use	✓	Anthropic reports the unintended uses of Claude 3.5 in their Claude 3 model card.	Anthropic [2024c]
Report safety evaluations	✓	Anthropic reports evaluation results around catastrophic harms, including autonomous replication and adaption, and cybersecurity risks in their Claude 3.5 addendum.	Anthropic [2024a]
Report on societal risks	✓	Anthropic describes their testing on societal risks, including discrimination, stereotype bias, and CBRN threats in their Claude 3.5 addendum.	Anthropic [2024a]
Report on adversarial testing	✓	Anthropic reports on results of multimodal policy red teaming and other evaluation results in their Claude 3.5 addendum.	Anthropic [2024a]
Empower trust and safety teams	✗	While Anthropic explicitly indicates that a key goal at the company level is to accelerate safety work, there is no relevant evidence of empowering its trust and safety team.	Anthropic [2023d]
Advance AI safety research	✓	Anthropic describes their research directions of scaling supervision, mechanistic interpretability, and process-oriented learning, with the goal of building safety systems.	Anthropic [2023d]
Advance privacy	✗	No relevant evidence found.	-
Protect children	✓	Anthropic reports child sexual abuse material to the National Center for Missing & Exploited Children.	Anthropic [2024d]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	No relevant evidence found.	-

Continued on next page

Indicator Scores for Anthropic – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that foster the education and training of students and workers	✗	Anthropic provides educational courses on using Claude, but they are not targeted specifically to students, workers, or broader accessibility.	Anthropic [2025]
Support initiatives that help citizens understand the technology	✗	Although Anthropic and the Collective Intelligence Project ran a public input process involving 1,000 Americans to draft a constitution for an AI system, this initiative does not improve citizens' understanding of the nature, capabilities, limitations, and impact of AI.	Anthropic [2023a]

Table 5: Indicator Scores for Apple

Indicator	Score	Justification	Source
Internal red-teaming	✗	While Apple describes employing both manual and automatic red-teaming to evaluate their models, there is no mention of the risk areas that these red-teaming efforts cover in its technical report.	Tom Gunter et al. [2024]
External red-teaming	✗	While Apple describes running red-teaming projects with both internal and external participants, its approach does not constitute a detailed red-teaming regime based on public information.	Tom Gunter et al. [2024]
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✗	No relevant evidence found.	-
Information sharing with government	✗	No relevant evidence found.	-
Forum or mechanism for information sharing	✓	Apple is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members to prevent misuse or security vulnerabilities.	-
Model weight cybersecurity practices	✗	Apple's Private Cloud Compute system does not address model weight security.	Engineering et al. [2024]
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	While Apple has an existing bug bounty and expanded it to include their compute system, this bounty does not include AI systems.	Engineering and [SEAR]
Include AI systems in their existing bug bounty programs	✗	While Apple has an existing bug bounty and expanded it to include their compute system, this bounty does not include AI systems.	Engineering and [SEAR]
Robust provenance or watermarking for audio	✗	No relevant evidence found.	-
Robust provenance or watermarking for visual content	✗	Apple intends to label AI-generated images in their metadata but has not provided evidence of implementation.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✗	No relevant evidence found.	-
Work with industry peers and standards-setting bodies towards developing a technical framework	✗	No relevant evidence found.	-

Continued on next page

Indicator Scores for Apple – *Continued from previous page*

Indicator	Score	Justification	Source
Report capabilities	✓	Apple reports the language and reasoning capabilities of their foundation models in their technical report.	Tom Gunter et al. [2024]
Report limitations	✗	The limitation discussed is generically about the use of LMs for grading, not their specific models.	Tom Gunter et al. [2024]
Report domains of appropriate use	✗	No relevant evidence found.	-
Report domains of inappropriate use	✗	No relevant evidence found.	-
Report safety evaluations	✗	Apple reports safety evaluations, including response rate to adversarial prompts and human evaluation of output harmfulness.	Tom Gunter et al. [2024]
Report on societal risks	✗	No relevant evidence found.	-
Report on adversarial testing	✗	Apple described their methods for manual and automatic red-teaming but does not report the results.	Tom Gunter et al. [2024]
Empower trust and safety teams	✗	While Apple has a Responsible AI team, there's no more information about how it empowers its trust and safety teams.	Tom Gunter et al. [2024]
Advance AI safety research	✓	Apple publishes a number of papers on AI safety.	Cuadros et al. [2024]
Advance privacy	✓	Apple publishes privacy research and created Private Cloud Compute designed for private AI processing.	Engineering et al. [2024], Apple [2024b]
Protect children	✗	No relevant evidence found.	-
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	No relevant evidence found.	-
Support initiatives that foster the education and training of students and workers	✗	While Apple Developer Academy includes coursework in AI, it is not targeted specifically to workers or the broader public.	Apple [2024a]
Support initiatives that help citizens understand the technology	✗	No relevant evidence found.	-

Table 6: Indicator Scores for Cohere

Indicator	Score	Justification	Source
Internal red-teaming	✗	No relevant evidence found.	-
External red-teaming	✓	Cohere participated in the Generative AI Red Teaming Challenge.	Intelligence [2023]
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✗	While Lakera and Cohere partnered to define new LLM security standards, this partnership is commercial in nature.	Lakera Team [2024]
Information sharing with government	✗	Although Cohere participated in the U.S. Senate AI Insight Forum, we do not award them this point because they did not initiate engagement and this engagement is not ongoing.	Cohere [2023]
Forum or mechanism for information sharing	✓	Cohere is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members to prevent misuse or security vulnerabilities.	-
Model weight cybersecurity practices	✗	No details about cybersecurity specific to model weights, just use of SOC 2.	Cohere [n.d.b]
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	While Cohere states that "access to cloud infrastructure and other sensitive tools are limited to authorized employees who require it for their role," securing access to cloud infrastructure is not equivalent to securing access to model weights.	Cohere [n.d.b]
Establish bounties, contests, or prizes	✗	Cohere Responsible Disclosure Policy describes an invite-only bug bounty program with BugCrowd, but there is insufficient information to evaluate the activities of the bug bounty.	Cohere [n.d.a]
Include AI systems in their existing bug bounty programs	✗	No relevant evidence found.	-
Robust provenance or watermarking for audio	✓	Cohere's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Robust provenance or watermarking for visual content	✓	Cohere's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Cohere's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Cohere's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-

Continued on next page

Indicator Scores for Cohere – *Continued from previous page*

Indicator	Score	Justification	Source
Report capabilities	✓	Cohere describes the capabilities of Command R in their documentation.	Cohere [2024b]
Report limitations	✓	Cohere discloses the language limitations of its model by specifying which languages it supports in the model card.	Cohere [2024c]
Report domains of appropriate use	✓	Cohere describes the intended use cases of Command R as text generation and RAG and tool-use tasks in its model card.	Cohere [2024c]
Report domains of inappropriate use	✓	Cohere discloses the unintended and prohibited uses in decision-making in Command R's model card.	Cohere [2024c]
Report safety evaluations	✗	While Cohere reports testing for model toxicity and bias, their safety evaluations do not target severe risks such as CBRN and child safety.	Cohere [2024c]
Report on societal risks	✓	Cohere publishes the evaluation results on the BOLD dataset in its model card, testing for model toxicity and bias.	Cohere [2024c]
Report on adversarial testing	✗	While Cohere reports evaluation results for BOLD, which can be construed as part of safety, these results do not constitute adversarial testing to determine appropriateness of deployment.	Cohere [2024c]
Empower trust and safety teams	✗	While Cohere created a Responsibility Council to inform their product and business decisions, there are no relevant mentions of a trust and safety team.	Cohere [2024c]
Advance AI safety research	✓	Cohere supports the Cohere for AI Scholars program, which provides research grants to researchers advancing safe, responsible LLM capabilities and applications. They also allow the "intentional stress testing of the API and adversarial attacks."	Team [2024a]
Advance privacy	✓	Cohere conducts research on privacy and is a member of the Coalition for Secure AI.	Utpala et al. [2023]
Protect children	✗	No relevant evidence found.	-
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	While Cohere's support of the Aya initiative is highly commendable, we do not award a point as providing multilingual AI capabilities at present does not seem akin to the named societal challenges that are more established and fundamental (e.g. cancer treatment).	Cohere [2024a]
Support initiatives that foster the education and training of students and workers	✗	While Cohere operates the Cohere for AI Scholars program, this program is structured as a research apprenticeship geared towards those with existing technical backgrounds. It is not designed to foster the education and training of students and workers in general, and so we do not award this point. This is also the case for its LLM University.	Team [2024b], Cohere [2025]

Continued on next page

Indicator Scores for Cohere – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that help citizens understand the technology	✗	No evidence found of citizen-specific programs.	-

Table 7: Indicator Scores for Google

Indicator	Score	Justification	Source
Internal red-teaming	✓	Google mentions their internal, company-wide red team that attempts different types of attacks. Its technical report for Gemini also describes their internal red-teaming efforts focused on security, safety, and privacy failures.	Fabian and Crisp [2023], Gemini Team [2024]
External red-teaming	✓	Google participated in the Generative AI Red Teaming Challenge. They also mention working with a small set of independent external groups to conduct unstructured red teaming.	Intelligence [2023], Gemini Team [2024]
Red teaming coverage of risks	✗	While the Gemini technical report discusses red-teaming across a broad range of risks, including CBRN and societal risks, these risk areas do not cover the effects of system interaction and tool use.	Gemini Team [2024]
Information sharing with companies	✓	Google is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Frontier Model Forum [n.d.]
Information sharing with government	✗	No specific sustained information sharing mechanism is described, though conferences and other convenings are mentioned without clarification on whether information beyond what is made public is shared.	Google [n.d.]
Forum or mechanism for information sharing	✓	Google is a member of the U.S. AI Safety Institute Consortium and the Frontier Model Forum.	NIST [2024], Frontier Model Forum [n.d.]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as desired.	Frontier Model Forum [n.d.]
Model weight cybersecurity practices	✗	No information is given about model weight security in the training of Gemini models.	-
Insider threat detection program	✗	No clear indication is provided that Google implements insider threat detection in relation to Gemini.	-
Limiting weight-level access to relevant personnel	✓	Google applies access controls to ensure only necessary personnel have access for business purposes.	Google [2024]
Establish bounties, contests, or prizes	✗	Google did not establish bounties but has an existing bug bounty that includes AI systems.	Laurie Richardson [2023]
Include AI systems in their existing bug bounty programs	✓	Google expanded their VRP to reward for attack scenarios specific to generative AI.	Laurie Richardson [2023]

Continued on next page

Indicator Scores for Google – *Continued from previous page*

Indicator	Score	Justification	Source
Robust provenance or watermarking for audio	✓	Google developed SynthID, which watermarks and identifies AI-generated content.	Google DeepMind [n.d.]
Robust provenance or watermarking for visual content	✓	Google developed SynthID, which watermarks and identifies AI-generated content.	Google DeepMind [n.d.]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	SynthID is a tool to determine if content is generated using Google models as described in the associated scientific paper.	Google DeepMind [n.d.]
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Google contributed to the Partnership on AI efforts on a Synthetic Media Framework to establish best practices.	Google [2023]
Report capabilities	✓	Google reports model capabilities through benchmark results in the Gemini technical report.	Gemini Team [2024]
Report limitations	✓	Google notes that there are limitations in a short section at the end of their Gemini technical report.	Gemini Team [2024]
Report domains of appropriate use	✓	Google enumerates domains of intended use in their model card for Gemini.	Google [2025]
Report domains of inappropriate use	✓	The prohibited use policy specifies such domains.	Google [2024]
Report safety evaluations	✓	The Gemini technical report includes safety evaluations around potential misuse.	Gemini Team [2024]
Report on societal risks	✓	The Gemini technical report includes evaluations on societal risks.	Gemini Team [2024]
Report on adversarial testing	✗	Google cites three types of external testing in the Gemini technical report but does not report the results from these exercises.	Gemini Team [2024]
Empower trust and safety teams	✓	Google instituted AI principles ethics reviews and impact assessments conducted by their Trust and Safety team as part of the pre-launch process.	Google [2023, n.d.]
Advance AI safety research	✓	Google regularly publishes and funds AI safety research.	Charlet [2024], Anca Dragan and Dafoe [2024]

Continued on next page

Indicator Scores for Google – *Continued from previous page*

Indicator	Score	Justification	Source
Advance privacy	✓	Google conducts privacy research around AI.	Google [2023], Charlet [2024]
Protect children	✓	We award this point for their partnerships with Thorn and All Tech is Human on CSAM.	Jasper [2024]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	Google has worked on AI projects to support researchers in healthcare and energy management.	Google [n.d.]
Support initiatives that foster the education and training of students and workers	✓	Google announced funding to equip educators and students with foundational AI skills through the development of AI curriculums.	Johnson [2024]
Support initiatives that help citizens understand the technology	✗	No evidence found of citizen-specific programs.	-

Table 8: Indicator Scores for IBM

Indicator	Score	Justification	Source
Internal red-teaming	✗	While IBM recruits company volunteers to conduct red-teaming, there is no evidence that these efforts cover the areas of misuse, societal risks, and national security concerns.	Martineau [2024], Granite [2024]
External red-teaming	✗	While IBM partnered with a third-party company to conduct external red-teaming of Granite, they do not provide information on how these red-teaming operations are conducted. As such, we do not consider this partnership to be sufficient for a detailed regime.	Granite [2024]
Red teaming coverage of risks	✗	While the Granite 3.0 technical report details evaluations for many risk areas, these do not constitute a detailed red-teaming regime with full coverage of the risk areas outlined in the voluntary commitments.	Research [2024]
Information sharing with companies	✓	IBM is a member of the AI Alliance.	IBM [2023]
Information sharing with government	✗	No evidence found of sustained mechanism for providing information about Granite models to government.	-
Forum or mechanism for information sharing	✓	IBM is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members.	-
Model weight cybersecurity practices	✗	While IBM provides related services to clients, they do not clearly disclose that they implement these practices in training their Granite models.	Shaikh [2024]
Insider threat detection program	✗	While IBM provides related services to clients, they do not clearly disclose that they implement these practices in training their Granite models.	Shaikh [2024]
Limiting weight-level access to relevant personnel	✗	While IBM provides related services to clients, they do not clearly disclose that they implement these practices in training their Granite models.	IBM [n.d.]
Establish bounties, contests, or prizes	✗	No relevant evidence found.	-
Include AI systems in their existing bug bounty programs	✗	No relevant evidence found.	-
Robust provenance or watermarking for audio	✓	IBM's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Robust provenance or watermarking for visual content	✓	IBM's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	IBM's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-

Continued on next page

Indicator Scores for IBM – *Continued from previous page*

Indicator	Score	Justification	Source
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	IBM's flagship models do not generate audio or images, and so we default their score to ✓ for this commitment.	-
Report capabilities	✓	Technical report for Granite 3.0 details model capabilities.	Research [2024]
Report limitations	✓	Model card for Granite 3.0 details model limitations.	IBM Research [2024]
Report domains of appropriate use	✓	Model card for Granite 3.0 details appropriate uses such as summarization and text classification.	IBM Research [2024]
Report domains of inappropriate use	✓	IBM outlines AI use restrictions in its service descriptions for foundation models in connection with the use of its Cloud Service.	IBM [2025a]
Report safety evaluations	✗	Technical report for Granite 3.0 details safety evaluation, but does not include CBRN results.	Research [2024]
Report on societal risks	✓	Technical report for Granite 3.0 describes socio-technical harms and risks.	Research [2024]
Report on adversarial testing	✓	Technical report for Granite 3.0 details red-teaming results.	Research [2024]
Empower trust and safety teams	✗	While IBM provides the WatsonX Governance service, there is no relevant evidence of empowering its trust and safety team.	IBM [2025b]
Advance AI safety research	✓	IBM regularly publishes AI safety research.	Granite [2024]
Advance privacy	✓	IBM conducts AI privacy research and created the AI Privacy Toolkit, a set of open-source tools for developers to ensure the privacy and compliance of their models.	IBM [2024]
Protect children	✗	No relevant evidence found.	-
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	IBM committed $30 million worth of technology and services to improve water management.	IBM [2023a]
Support initiatives that foster the education and training of students and workers	✓	IBM is training two million learners in AI by 2026, with a focus on underrepresented communities.	IBM [2023b]
Support initiatives that help citizens understand the technology	✗	No citizen-specific program found.	-

Table 9: Indicator Scores for Inflection

Indicator	Score	Justification	Source
Internal red-teaming	✗	Inflection mentions hosting an internal Safety team which pressure tests their models, but there is no evidence that its internal red-teaming efforts cover areas such as misuse, societal risks, and national security concerns.	Inflection AI [2023c]
External red-teaming	✗	While Inflection mentions commissioning outside experts to conduct red-teaming, they do not provide further information on their red-teaming operations.	Inflection AI [2023c]
Red teaming coverage of risks	✗	Inflection notes that its CBRN red-teaming efforts are expanding, but its current efforts do not cover the full range of risks outlined in the voluntary commitments.	Inflection AI [2023b]
Information sharing with companies	✗	While Inflection is involved with the Partnership on AI, this does not appear to involve private information-sharing between companies.	on AI Staff [2024], MLCommons [2023]
Information sharing with government	✗	While Inflection responded to the UK government before the AI Safety Summit, there is no evidence of non-public information sharing.	Inflection AI [2023b]
Forum or mechanism for information sharing	✓	Inflection is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive risk information.	-
Model weight cybersecurity practices	✗	No relevant evidence found.	-
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	Inflection restricts access to data and code, but not model weights specifically.	Inflection AI [2023b]
Establish bounties, contests, or prizes	✓	Inflection implemented a closed pilot bug bounty program inviting security researchers to identify vulnerabilities.	Inflection AI [2023b]
Include AI systems in their existing bug bounty programs	✗	Inflection did not have an existing bug bounty.	Inflection AI [2023b]
Robust provenance or watermarking for audio	✓	Inflection's flagship models do not generate audio or images.	-
Robust provenance or watermarking for visual content	✓	Inflection's flagship models do not generate audio or images.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Inflection's flagship models do not generate audio or images.	-

Continued on next page

Indicator Scores for Inflection – *Continued from previous page*

Indicator	Score	Justification	Source
Work with industry peers and standards-setting bodies on content traceability	✓	Inflection's flagship models do not generate audio or images.	-
Report capabilities	✓	Inflection outlines the capabilities of Pi 3.0 through prompt-response examples.	Inflection AI [2025]
Report limitations	✓	Inflection's documentation includes a section on known limitations.	Inflection AI [2023b]
Report domains of appropriate use	✗	No relevant evidence found.	-
Report domains of inappropriate use	✓	Inflection's Terms of Service detail illegal and harmful uses.	Inflection AI [2023a]
Report safety evaluations	✗	No relevant evidence found.	-
Report on societal risks	✗	No relevant evidence found.	-
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✓	Inflection's safety team is authorized to block malicious actors and implement mitigations.	Inflection AI [2023b]
Advance AI safety research	✗	Inflection is a member of MLCommons' safety working group but does not provide evidence of its own contributions.	MLCommons [2023]
Advance privacy	✗	No relevant evidence found.	-
Protect children	✓	Inflection partners with Thorn to mitigate CSAM risks.	Inflection AI [2023b], Thorn [2024]
Support frontier AI for societal challenges	✗	No relevant evidence found.	-
Support education/training of students and workers	✗	No relevant evidence found.	-
Support citizen understanding of AI	✗	No citizen-specific program found.	-

Table 10: Indicator Scores for Meta

Indicator	Score	Justification	Source
Internal red-teaming	✓	Meta conducted red-teaming exercises with their internal teams covering types of misues, societal risks, and national security concerns.	Meta [2024b]
External red-teaming	✓	Meta participated in the Generative AI Red Teaming Challenge and conducted red-teaming with exteneral experts.	Intelligence [2023], Clegg [2024]
Red teaming coverage of risks	✗	While Meta describes its red-teaming efforts across a range of risk categories, including the production of weapons and privacy violations, there is no evidence that they they conduct red-teaming on self-replication or the effects of system interaction and tool use.	Meta [2023b]
Information sharing with companies	✓	Meta is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Frontier Model Forum [n.d.]
Information sharing with government	✗	While Meta provides services procured by the federal government, no information is disclosed about what is provided as part of the procurement process, nor are separate mechanisms discussed for information sharing about the models	Meta [2023b]
Forum or mechanism for information sharing	✓	Meta is a member of the U.S. AI Safety Institute Consortium and the Frontier Model Forum.	NIST [2024], Frontier Model Forum [n.d.]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as desired.	Frontier Model Forum [n.d.]
Model weight cybersecurity practices	✗	Meta does not specifically indicate they implement security practices for model weights in relation to their Llama models.	Meta [2023b]
Insider threat detection program	✗	Meta indicates their intention to implement insider threat detection without providing evidence of implementation.	Meta [2023b]
Limiting weight-level access to relevant personnel	✗	Meta indicates their intention to implement insider threat detection without providing evidence of implementation.	Meta [2023b]
Establish bounties, contests, or prizes	✗	Meta has an existing bug bounty program for privacy and security issues that includes Meta's large language models; no additional bounties, contests, or prizes found.	Meta [n.d.]
Include AI systems in their existing bug bounty programs	✓	Meta has an existing bug bounty program for privacy and security issues that includes Meta's large language models.	Meta [n.d.]

Continued on next page

Indicator Scores for Meta – *Continued from previous page*

Indicator	Score	Justification	Source
Robust provenance or watermarking for audio	✓	Meta introduced AudioSeal, a watermarking technique for localized detection of AI-generated speech.	Meta [2024a]
Robust provenance or watermarking for visual content	✓	Meta added visible indicators on photorealistic images generated by AI.	Meta [2023b], Meta AI [2023]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Meta mentions building tools that can identify visible markers at scale, following the C2PA and IPTC technical standards.	Meta [2024b]
Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI	✓	Meta has worked with industry partners in PAI to develop provenance standards.	Meta [2023b], Meta AI [2023]
Report capabilities	✓	Model card for Llama 3.3 reports capabilities through benchmark results	Meta AI [2024]
Report limitations	✓	Model card for Llama 3.3 reports limitations	Meta AI [2024]
Report domains of appropriate use	✓	Model card for Llama 3.3 reports intended use cases	Meta AI [2024]
Report domains of inappropriate use	✓	Model card for Llama 3.3 enumerates prohibited uses.	Meta AI [2024]
Report safety evaluations	✓	Llama 3.3 model card reports evaluations around cybersecurity, child safety, and CBRN risks.	Meta AI [2024]
Report on societal risks	✗	While the Llama 3.3 model card includes testing for cybersecurity, child safety, and CBRN risks, it does not report societal risk such as fairness and bias.	Meta AI [2024]
Report on adversarial testing used to determine appropriateness of deployment	✗	The Llama 3.3 Model card describes their methods but does not disclose results of adversarial testing.	Meta AI [2024]
Empower trust and safety teams	✗	While Meta supports teams developing Purple Llama, Prompt Guard, and Llama Guard 3, all safety tools, there is no relevant evidnece of empowering its trust and safety team, which is responsible for monitoring potential risks like bias, misinformation, and security threats.	Meta [2024b]
Advance AI safety research	✓	Meta developed trust and safety tools and evaluations for AI developers and regularly publishes AI safety research.	Meta [2023a]
Advance privacy	✓	Meta launched a generative AI privacy guide and has grown their privacy teams substantially in the last years.	Meta [2024a, n.d.]

Continued on next page

Indicator Scores for Meta – *Continued from previous page*

Indicator	Score	Justification	Source
Protect children	✓	We award this point for their partnerships with Thorn and the National Center for Missing & Exploited Children on combatting CSAM.	Meta [2024b]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	The partnership between Meta and CMU advances the use of AI for addressing climate change.	Meta [2022b]
Support initiatives that foster the education and training of students and workers	✓	As part of its AI Learning Alliance, Meta built a deep learning course curriculum with Georgia Tech that is free to all. They are also working with professors at HBCUs and HSIs to further develop and teach the curriculum.	Meta [2022a]
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	Although Meta hosted a "Community Forum on Generative AI," which gave randomly selected member of the public the chance to share their views on the principles governing the use of AI chatbots, this initiative does not improve citizens' understanding of the nature, capabilities, limitations, and impact of the technology.	Broxmeyer [2024]

Table 11: Indicator Scores for Microsoft

Indicator	Score	Justification	Source
Internal red-teaming	✓	Microsoft has an AI Red Team that identifies safety and security vulnerabilities in their latest generative AI models.	Ram Shankar Siva Kumar [2024]
External red-teaming	✗	Microsoft reports building external red-teaming capacity but does not make clear whether external experts have red-teamed their models.	Microsoft [2024f]
Red teaming coverage of risks	✗	While Microsoft's PyRit framework covers a broad range of risk areas, it does not include the risk of model autonomy and self-replication.	Munoz et al. [2024]
Information sharing with companies	✓	Microsoft is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Frontier Model Forum [n.d.]
Information sharing with government	✗	While Microsoft services are procured by the federal government, no information is disclosed about what is provided as part of the procurement process, nor are separate mechanisms discussed for information sharing about Microsoft's models (as opposed to, say, their offerings of OpenAI models as an Azure service).	Krishan [2023], Microsoft [2024c]
Forum or mechanism for information sharing	✓	Microsoft is a member of the U.S. AI Safety Institute Consortium and the Frontier Model Forum.	NIST [2024], Frontier Model Forum [n.d.]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as desired.	Frontier Model Forum [n.d.]
Model weight cybersecurity practices	✓	Microsoft implements model-specific security controls.	Microsoft [2023]
Insider threat detection program	✗	Microsoft does not describe an insider threat detection model in relation to Phi or their AI services.	-
Limiting weight-level access to relevant personnel	✓	Microsoft implements strong identify and access control to their AI technology and logs access requests to identify anomalies and/or unauthorized access.	Microsoft [2023]
Establish bounties, contests, or prizes	✓	Microsoft launched a new AI bug bounty program.	Microsoft [2023]
Include AI systems in their existing bug bounty programs	✓	Microsoft created a Copilot bounty program that adheres to the terms of existing programs, namely Microsoft Bounty Terms and Conditions and their bounty Safe Harbor policy.	Microsoft [2025]
Robust provenance or watermarking for audio	✓	Microsoft watermarks to the speech outputs created with the personal voice feature.	Microsoft [2024a]

Continued on next page

Indicator Scores for Microsoft – *Continued from previous page*

Indicator	Score	Justification	Source
Robust provenance or watermarking for visual content	✓	Microsoft attaches provenance metadata generated by their AI services	Microsoft [2023]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Content Credentials Verify is a tool for users to inspect the Content Credentials of AI-generated content via DALL-E through the Azure OpenAI endpoint.	Microsoft [2023]
Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI	✓	Microsoft co-founded C2PA and co-developed the C2PA technical specification.	Microsoft [2023]
Report capabilities	✓	Technical report for Phi-4 reports capabilities through benchmark results.	Marah Abdin et al. [2024]
Report limitations	✓	Technical report for Phi-4 discloses limitations such as limited skill scope, bias in generation-based benchmarks, and limitations of multiple-choice tasks.	Marah Abdin et al. [2024]
Report domains of appropriate use	✓	The Phi 4 model card enumerates intended and out-of-scope use cases.	Microsoft [2024d]
Report domains of inappropriate use	✓	The Phi 4 model card enumerates intended and out-of-scope use cases.	Microsoft [2024d]
Report safety evaluations	✗	Technical report for Phi-4 discloses safety evaluation results, but not for CBRN.	Marah Abdin et al. [2024]
Report on societal risks	✗	No relevant evidence found.	-
Report on adversarial testing used to determine appropriateness of deployment	✓	Technical report for Phi-4 describes the results from their red-teaming exercise.	Marah Abdin et al. [2024]
Empower trust and safety teams	✗	While Microsoft has created a responsible AI council, there is no relevant evidence around empowering its trust and safety team.	Crampton [2023]
Advance AI safety research	✓	Microsoft Azure AI releases tooling to advance AI safety.	Zarfati [2023]
Advance privacy	✓	Microsoft Azure releases tooling to advance AI privacy.	Microsoft [2024e]
Protect children	✓	We award this point for their partnerships with Thorn on combatting CSAM.	Microsoft [2024b]
Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats	✓	Microsoft launched AI for Health, a philantropic program to support researchers tackling global health challenges.	Microsoft Research [n.d.a]

Continued on next page

Indicator Scores for Microsoft – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI	✓	Microsoft launched an AI Skills Initiative to help students and workers develop AI skills. They also developed an Education AI Toolkit is intended for educators.	Behnchen [2023], Microsoft [2025]
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	While Microsoft hosted a series featuring Trevor Noah to discuss the potential of AI in addressing global issues, this initiative is a one-time occurence and not sustained.	Microsoft Research [n.d.b]

Table 12: Indicator Scores for Nvidia

Indicator	Score	Justification	Source
Internal red-teaming	✓	Nvidia has an internal AI red team whose testing covers types of misues, societal risks, and security concerns.	Pearce and Lucas [2023]
External red-teaming	✓	Nvidia participated in the Generative AI Red Teaming Challenge	Intelligence [2023]
Red teaming coverage of risks	✗	While Nvidia describes the high-level risks their red teaming targets, there is nsufficient coverage of risk areas, such as CBRN and cyber capabilities.	Pearce and Lucas [2023]
Information sharing with companies	✗	No relevant evidence found.	-
Information sharing with government	✗	While Nvidia has made generous contributions to NAIRR and they are collaborating with DARPA to identify AI-generated images, we do not count these towards fulfilling the commitment. There is no evidence that Nvidia has shared non-public information with governments about trust and safety risks, AI system capabilities, or attempts to circumvent safeguards.	Berry [2024a], Finkle [2023]
Forum or mechanism for information sharing	✓	Nvidia is a member of the AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members to prevent misuse or security vulnerabilities.	-
Model weight cybersecurity practices	✗	While Nvidia provides related services to clients, they do not clearly disclose that they implement these practices in training their models.	Jr. [2023]
Insider threat detection program	✗	While Nvidia provides related services to clients, they do not clearly disclose that they implement this threat detection internally.	Nvidia [n.d.]
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	No relevant evidence found.	-
Include AI systems in their existing bug bounty programs	✗	No relevant evidence found.	-
Robust provenance or watermarking for audio	✓	Nvidia's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Robust provenance or watermarking for visual content	✓	Nvidia's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-

Continued on next page

Indicator Scores for Nvidia – *Continued from previous page*

Indicator	Score	Justification	Source
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Nvidia's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI	✓	Nvidia's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Report capabilities	✓	Technical report for he Nemotron-4-340B Instruct reports capabilities through benchmark results.	Nvidia [2024d]
Report limitations	✓	The Nemotron Hugging Face model card report limitations in toxic language, unsafe content, and societal biases.	Nvidia [2024a]
Report domains of appropriate use	✓	The Nemotron Hugging Face model card reports intended uses.	Nvidia [2024b]
Report domains of inappropriate use	✓	The NemoTron license outlines inappropriate uses.	Nvidia [2024c]
Report safety evaluations	✗	Technical report for he Nemotron-4-340B Instruct details safety evaluation results, but none for CBRN.	Nvidia [2024d]
Report on societal risks	✓	Technical report covers evaluations for safety and societal risks.	Nvidia [2024d]
Report on adversarial testing used to determine appropriateness of deployment	✓	Technical report for he Nemotron-4-340B Instruct details safety evaluation results from red-teaming.	Nvidia [2024d]
Empower trust and safety teams	✗	No relevant evidence found.	-
Advance AI safety research	✓	Nvidia developed an open source toolkit for developing safe and trustworhy LLM conversational systems.	Chockalingam and Varshney [2023]
Advance privacy	✗	They do not publicly indicate that they concretely take actions to advance privacy.	-
Protect children	✗	No relevant evidence found.	-
Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats	✓	Nvidia detailed partnerships to improve diagnostics and healthcare delivery and advance climate modeling efforts.	Caulfield [2024]
Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI	✗	While NVIDIA maintains a certification for educators through its Deep Learning Institute University Ambassador Program, it does not target these efforts to workers or the broader public.	Stewart [2024], Berry [2024b]

Continued on next page

Indicator Scores for Nvidia – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	No relevant evidence found.	-

Table 13: Indicator Scores for OpenAI

Indicator	Score	Justification	Source
Internal red-teaming	✓	OpenAI documents internal red-teaming in o1 system card in areas including misuse, societal risks, and national security concerns.	OpenAI [2024b]
External red-teaming	✓	OpenAI launched the OpenAI Red Teaming Network, a community of experts to inform their risk assessment.	OpenAI [2023d]
Red teaming coverage of risks	✓	The o1 system card details full coverage of the risk areas in the voluntary commitments under their external red-teaming efforts	OpenAI [2024b]
Information sharing with companies	✓	OpenAI is a member of the Frontier Model Forum, which facilitates information-sharing among companies.	Frontier Model Forum [n.d.]
Information sharing with government	✓	OpenAI provides the U.S. AI Safety Institute with new models before and and following their releases.	NIST [2024]
Forum or mechanism for information sharing	✓	OpenAI is a member of the U.S. AI Safety Institute Consortium and the Frontier Model Forum.	NIST [2024], Frontier Model Forum [n.d.]
Forum or mechanism shares information on risks	✓	The Frontier Model Forum shares information as desired.	Frontier Model Forum [n.d.]
Model weight cybersecurity practices	✓	OpenAI created secure research environments dedicated to model security, including protecting model weights.	OpenAI [2024e, 2023c]
Insider threat detection program	✗	No evidence found.	Rafieyan [2024]
Limiting weight-level access to relevant personnel	✓	OpenAI built a service called AccessManager to manage internal authorization and access to sensitive resources, including model weights.	OpenAI [2024e]
Establish bounties, contests, or prizes	✓	OpenAI launched a bug bounty program for their systems.	OpenAI [2023b]
Include AI systems in their existing bug bounty programs	✓	OpenAI expanded their bug bounty to include model issues.	OpenAI [2023b]
Robust provenance or watermarking for audio	✓	OpenAI has implemented tamper-resistant watermarking on their generated audiovisual content.	OpenAI [2024g]
Robust provenance or watermarking for visual content	✓	OpenAI has implemented tamper-resistant watermarking on their generated audiovisual content.	OpenAI [2024g]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	OpenAI built a tool that predicts the likelihood an image was generated by DALL-E 3.	OpenAI [2024g]

Continued on next page

Indicator Scores for OpenAI – *Continued from previous page*

Indicator	Score	Justification	Source
Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI	✓	OpenAI joined C2PA and is working towards developing the C2PA standard.	OpenAI [2024g]
Report capabilities	✓	OpenAI describes o1's capabilities in the model release announcement.	OpenAI [2024b]
Report limitations	✗	While the o1 system card details a number of safety challenges, it does not provide a comprehensive description of the model limitations.	OpenAI [2024b]
Report domains of appropriate use	✗	While OpenAI's usage policies describes prohibited uses, it does not report domains of appropriate use.	OpenAI [2024f]
Report domains of inappropriate use	✓	OpenAI's usage policies describes prohibited uses.	OpenAI [2024f]
Report safety evaluations	✓	o1 system card includes safety evaluation.	OpenAI [2024b]
Report on societal risks	✓	o1 system card includes assessment on measuring CBRN risks and impact on different industries / occupations according to prepardness framework.	OpenAI [2024b]
Report on adversarial testing used to determine appropriateness of deployment	✓	o1 system card details adversarial testing results for external red-teaming.	OpenAI [2024b]
Empower trust and safety teams	✗	While OpenAI formed a Safety and Security Committee, there is no relevant mention of a trust and safety team.	OpenAI [2024a]
Advance AI safety research	✓	OpenAI developed a Prepardness Framework as a proactive, risk-based approach to AI development.	OpenAI [2024d]
Advance privacy	✓	The OpenAI consumer privacy documentation details how they aim to minimize personal information in model training and in model generations.	OpenAI [2024c]
Protect children	✓	We award this point for their partnerships with Thorn and the National Center for Missing & Exploited Children on combatting CSAM.	OpenAI [2023a, 2024d]
Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats	✓	OpenAI is working with Color Health to accelerate cancer patients' access to treatement.	OpenAI [n.d.]

Continued on next page

Indicator Scores for OpenAI – *Continued from previous page*

Indicator	Score	Justification	Source
Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI	✓	OpenAI partners with Common Sense media to provide free AI training courses for K-12 educators, helping them understand and responsibly implement the basics of AI into their work.	Common Sense Media [2024]
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	Microsoft and OpenAI launched a $2 million Societal Resilience Fund to further AI education and literacy among voters and vulnerable communities. However, we do not award this point because the fund is described to support initiatives that create better understanding of AI capabilities, and not the nature, limitations, and impact of the technology.	Hutson [2024]

Table 14: Indicator Scores for Palantir

Indicator	Score	Justification	Source
Internal red-teaming	✗	No relevant evidence found.	-
External red-teaming	✗	No relevant evidence found.	-
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✗	No relevant evidence found.	-
Information sharing with government	✗	While Palantir extensively engages the US government in procurement relationships, we find no evidence of information sharing with the government outside of these procurement relationships and the procurement relations/contracts are also not described publicly.	Palantir [2024b]
Forum or mechanism for information sharing	✓	Palantir is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share information on risks.	-
Model weight cybersecurity practices	✓	Palantir is not a frontier model developer to our knowledge, so we default their score to ✓ for this commitment.	-
Insider threat detection program	✓	Palantir is not a frontier model developer to our knowledge, so we default their score to ✓ for this commitment.	-
Limiting weight-level access to relevant personnel	✓	Palantir is not a frontier model developer to our knowledge, so we default their score to ✓ for this commitment.	-
Establish bounties, contests, or prizes	✗	No relevant evidence found.	-
Include AI systems in their existing bug bounty programs	✗	No mention of AI in existing bug bounties.	-
Robust provenance or watermarking for audio	✓	Palantir's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Robust provenance or watermarking for visual content	✓	Palantir's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Palantir's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Palantir's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Report capabilities	✓	Documentation details platform capabilities.	Palantir [n.d.a]
Report limitations	✗	No relevant evidence found.	-

Continued on next page

Indicator Scores for Palantir – *Continued from previous page*

Indicator	Score	Justification	Source
Report domains of appropriate use	✓	Palantir platform describes many domains of use.	Palantir [n.d.b]
Report domains of inappropriate use	✓	Palantir's terms and conditions enumerate illegal and improper uses for their products	Palantir [n.d.d]
Report safety evaluations	✗	No relevant evidence found.	-
Report on societal risks	✗	While high-level ethical principles are discussed, no specifics on risks associated with the platform in particular.	Palantir [2023]
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✗	No relevant evidence found.	-
Advance AI safety research	✗	No relevant evidence found.	-
Advance privacy	✓	Palantir has a Privacy and Civil Liberties Engineering team and released a series of writing around their privacy-by-design engineering approach.	Palantir [2024c]
Protect children	✓	We award this point for the Palantir-NCMEC partnership.	Palantir [2024a]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	Palantir and Tree Energy Solutions partnered to leverage Palantir AI software to drive the green energy transition.	Palantir [2024d]
Support initiatives that foster the education and training of students and workers	✗	No evidence found; we do not award this for a generic scholarship program on STEM subjects.	Palantir [n.d.c]
Support initiatives that help citizens understand the technology	✗	No relevant evidence found.	-

Table 15: Indicator Scores for Salesforce

Indicator	Score	Justification	Source
Internal red-teaming	✗	Salesforce has conducted 19 internal red teaming exercises, but there is no evidence that these exercise covers types of misuse, societal risks, and national security concerns.	Salesforce [2024a]
External red-teaming	✗	While Salesforce has conducted two external red teaming exercises, we cannot gauge whether these are structured, organized operations that constitute a detailed red teaming regime based on available information.	Salesforce [2024a]
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✓	Salesforce is a member of the AI Alliance.	**?**
Information sharing with government	✗	While Salesforce publishes many (20+) articles publicly that provide value to the government and to other companies, given this information is public, we do not consider it as eligible to award this point and find no other evidence.	Salesforce [2024a]
Forum or mechanism for information sharing	✓	Salesforce is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share information on risks.	-
Model weight cybersecurity practices	✗	No information found on model weight security specifically.	-
Insider threat detection program	✗	While Salesforce provides insider threat detection services, there is no clear indication that these programs are implemented in relation to Salesforce's AI systems.	**?**
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	While Salesforce has a bug bounty program to prevent AI-powered cyber threats, Salesforce does not specify that their AI systems are covered under the scope of this program.	Orlando Lugo [2024]
Include AI systems in their existing bug bounty programs	✗	While Salesforce has a bug bounty program to prevent AI-powered cyber threats, Salesforce does not specify that their AI systems are covered under the scope of this program.	Orlando Lugo [2024]
Robust provenance or watermarking for audio	✓	Salesforce's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Robust provenance or watermarking for visual content	✓	Salesforce's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Salesforce's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-

Continued on next page

Indicator Scores for Salesforce – *Continued from previous page*

Indicator	Score	Justification	Source
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Salesforce's flagship models do not generate audio or images.	-
Report capabilities	✓	The XGen-7B technical report describes model capabilities.	Salesforce [2023]
Report limitations	✓	The XGen-7B technical report describes limitations.	Salesforce [2023]
Report domains of appropriate use	✓	The Salesforce Einstein model cards include intended uses.	Salesforce [2024b]
Report domains of inappropriate use	✓	The Salesforce Einstein model cards include out-of-scope uses.	Salesforce [2024b]
Report safety evaluations	✗	No relevant evidence found.	-
Report on societal risks	✗	No relevant evidence found.	-
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✗	While Salesforce has funded user research headcount to focus on trust and responsible AI, there is no relevant evidence of empowering its trust and safety team.	Salesforce [2024a]
Advance AI safety research	✓	Salesforce AI Research has published research on trust and safety evaluations for LLMs and have open-sourced a library for auditing generative AI for trustworthiness.	Salesforce [2024a]
Advance privacy	✓	We award this point for the Einstein Trust Layer, which reduces the presence of PII in prompts to LMs, which may in turn reduce privacy risks associated with prompt-conditioned LM generation and/or the storage of prompts.	Salesforce [2024a]
Protect children	✓	We award this point for their role in the Technology Task Force for CSAM.	Thorn [2019]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	Salesforce launched the Salesforce Accelerator to support purpose-driven nonprofits in developing AI-powered climate solutions.	Salesforce [2024a]
Support initiatives that foster the education and training of students and workers	✓	Salesforce provides grants to U.S. school districts and global education nonprofits to introduce AI literacy.	Salesforce [2024c]
Support initiatives that help citizens understand the technology	✗	While Salesforce provides information that could be intelligible to laypersons, we find no evidence of initiatives specifically aimed at increasing citizen's AI literacy.	-

Table 16: Indicator Scores for Scale AI

Indicator	Score	Justification	Source
Internal red-teaming	✗	No internal red-teaming described for Donovan or other models built by Scale AI, though we recognize that Scale does conduct such pre-deployment red-teaming for other developers' models and creates related services.	Scale AI [n.d.d, 2023b]
External red-teaming	✓	Scale AI participated in the Generative AI Red Teaming Challenge	Intelligence [2023]
Red teaming coverage of risks	✗	No relevant evidence found of red-teaming for specific risks; we do not award this point for creating safety benchmarks like WMDP nor for red-teaming OpenAI's GPT-4 model.	Yue and Berrios [2024], Dylan Slack [2023]
Information sharing with companies	✗	While Scale AI engages in commercial partnerships with other companies, we do not award this point given that the primary objective of these partnerships is not to advance the trust and safety of AI.	CSIS [2023], Murthy [2023]
Information sharing with government	✗	While ScaleAI provides services procured by the federal government, no information is disclosed about what is provided as part of the procurement process, nor are separate mechanisms discussed for information sharing about the models	Scale AI [2024a]
Forum or mechanism for information sharing	✓	Scale AI is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✗	Based on public information, the NIST AISIC does not share sensitive information on risks that must be carefully controlled and restricted to forum members to prevent misuse or security vulnerabilities.	-
Model weight cybersecurity practices	✗	No relevant evidence found.	-
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	While Scale AI launched the Humanity's Last Exam with a prize pool, this project is not aimed at incentivizing the responsible disclosure of weaknesses, such as unsafe behaviors.	Scale AI [2024b]
Include AI systems in their existing bug bounty programs	✗	No relevant evidence found.	-
Robust provenance or watermarking for audio	✓	Scale AI's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Robust provenance or watermarking for visual content	✓	Scale AI's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-

Continued on next page

Indicator Scores for Scale AI – *Continued from previous page*

Indicator	Score	Justification	Source
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	Scale AI's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	Scale AI's flagship models do not generate audio or images, and so we default their score to 1 for this commitment.	-
Report capabilities	✓	Scale reports that Donovan has capabilities for military and intelligence operations planning and for target analysis.	Scale AI [n.d.a]
Report limitations	✗	While Scale describes limitations of LMs in the context of their work evaluation other developers' LMs, we do not find any description of the limitations of their own tools.	Scale AI [n.d.c]
Report domains of appropriate use	✓	We award this point for the description of uses of Donovan.	Scale AI [n.d.a]
Report domains of inappropriate use	✓	The Scale Acceptable Use Policy outlines inappropriate uses of its services.	Scale AI [2022a]
Report safety evaluations	✗	While Scale conducts safety evaluations of other developers' models, we do not find evidence of safety evaluations for Donovan.	Scale AI [n.d.c]
Report on societal risks	✗	While Scale conducts safety evaluations, and provides a holistic framework for testing and evaluation that addresses societal risks, they do not report on such matters for Donovan.	Dylan Slack [2023], Scale AI [n.d.c]
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✗	Scale AI launched the Safety, Evaluations, and Alignment Lab aimed at enhancing transparency and standardization in AI safety.	Scale AI [2023a]
Advance AI safety research	✓	We award this point for their safety evaluation benchmark: the Weapons of Mass Destruction Proxy.	Scale AI [2024c]
Advance privacy	✗	No relevant evidence found.	-
Protect children	✗	No relevant evidence found.	-
Support research and development of frontier AI systems that can help meet society's greatest challenges	✓	We award this point for the CSIS partnership on international security.	CSIS [2023], Scale AI [2022b]
Support initiatives that foster the education and training of students and workers	✗	While Scale AI partnered with Chegg to develop proprietary LLMs for personalized learning, this initiative is not accessible to all students and workers and does not benefit them in developing AI skills.	Chegg [2023]
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	While Scale provides public leaderboards, we find no evidence of direct methods for increasing lay citizen AI literacy.	Scale AI [n.d.b]

Table 17: Indicator Scores for Stability AI

Indicator	Score	Justification	Source
Internal red-teaming	✗	While the model card for Stable Diffusion details red-teaming as part of integrity evaluation, the company does not distinguish between internal and external red-teaming in its description.	Stability AI [2024a]
External red-teaming	✓	Stability AI participated in the Generative AI Red Teaming Challenge	Intelligence [2023]
Red teaming coverage of risks	✗	No relevant evidence found.	-
Information sharing with companies	✗	While Stability AI engages in commercial partnerships with other companies, we do not award this point given that the primary objective of these partnerships is not to advance the trust and safety of AI.	Dominguez [2024]
Information sharing with government	✗	Although Stability AI participated in the U.S. Senate AI Insight Forum, we do not award them this point because they did not initiate engagement and this engagement is not ongoing.	Stability AI [2023b]
Forum or mechanism for information sharing	✓	Stability AI is a member of the U.S. AI Safety Institute Consortium.	NIST [2024]
Forum or mechanism shares information on risks	✓	We award this point for the commitments made in relation to sharing of information on CSAM.	Stability AI [2024]
Model weight cybersecurity practices	✗	No relevant evidence found.	-
Insider threat detection program	✗	No relevant evidence found.	-
Limiting weight-level access to relevant personnel	✗	No relevant evidence found.	-
Establish bounties, contests, or prizes	✗	No relevant evidence found.	-
Include AI systems in their existing bug bounty programs	✗	No relevant evidence found.	-
Robust provenance or watermarking for audio	✗	No relevant evidence found.	-
Robust provenance or watermarking for visual content	✓	We award this point for the integration of Content Credentials into the Stability AI API.	Stability AI [2023a]
Develop tools or APIs to determine if a particular piece of content was created within their tools	✓	We award this point for the integration of Content Credentials into the Stability AI API.	Stability AI [2023a]
Work with industry peers and standards-setting bodies towards developing a technical framework	✓	We award this point for their collaboration with the Content Authenticity Initiative and their work on adopting the C2PA standard for metadata.	Stability AI [2024b]
Report capabilities	✓	Stability AI reports model capabilities through benchmark results in the Stable Diffusion 3 technical report	Esser et al. [2024]

Continued on next page

Indicator Scores for Stability AI – *Continued from previous page*

Indicator	Score	Justification	Source
Report limitations	✓	Model card for Stable Diffusion 3.5 on Hugging Face describes some limitations.	Stability AI [2024a]
Report domains of appropriate use	✓	Model card on Hugging Face for Stable Diffusion 3.5 details intended uses.	Stability AI [2024a]
Report domains of inappropriate use	✓	Stability AI reports domains of inappropriate use in their Acceptable Use Policy	Stability AI [2024c]
Report safety evaluations	✗	No relevant evidence found.	-
Report on societal risks	✗	No relevant evidence found.	-
Report on adversarial testing	✗	No relevant evidence found.	-
Empower trust and safety teams	✗	No evidence found; external partnerships do not clearly empower their trust and safety team.	Stability AI [n.d.a]
Advance AI safety research	✗	No relevant evidence found that Stability explicitly advances AI safety, though the open release of model weights with documentation of datasets does enable others to do better AI safety research.	Stability AI [n.d.b]
Advance privacy	✗	No relevant evidence found.	-
Protect children	✓	We award this point for their partnerships with Thorn and All Tech is Human on CSAM.	Stability AI [2024]
Support research and development of frontier AI systems that can help meet society's greatest challenges	✗	No relevant evidence found.	-
Support initiatives that foster the education and training of students and workers	✗	No relevant evidence of significant support of initiatives beyond a one-off six-week program led by HUG Studios that we deem insufficient to award this point. While Stability AI has a learning hub around setting up and using its models, these courses and guides do not adequately provide the training students and workers need to prosper from the benefits of AI.	Studios [2024]
Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology	✗	No relevant evidence found.	-

D Company-Model Mapping

We map from the 16 signatory companies to their flagship models (or systems if the company generally does not build models) as follows: (Adobe, Firefly Image 3), (Amazon, Nova), (Anthropic, Claude 3.5 Sonnet (new)), (Apple, Apple Intelligence Foundation Language Models), (Cohere, Command R+), (IBM, Granite 3.0), (Inflection, Infection 3.0/Pi 3.0), (Google, Gemini 1.0), (Meta, Llama 3.3), (Microsoft, Phi-4), (Nvidia, Nemotron-4 340B), (OpenAI, o1), (Palantir, AIP), (Salesforce, xgen), (Scale AI, Donovan), (Stability AI, Stable Diffusion 3.5).

E Scoring Criteria

Table 18: Scoring Criteria for Indicators

Indicator	Commitment Text	Criteria
Internal red-teaming	"Commit to internal ... red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas."	The company differentiates internal and external red-teaming initiatives. Its internal red-teaming efforts cover, at minimum, areas of misuse, societal risks, and national security concerns. These are areas listed in the commitment.
External red-teaming	"Companies commit to advancing this area of research, and to developing a multi-faceted, specialized, and detailed red-teaming regime, including drawing on independent domain experts, for all major public releases of new models within scope."	We consider a detailed red-teaming regime to be a structured, organized exercise with the sole focus of external red-teaming. We consider a disclosure about a detailed red-teaming regime with independent domain experts to be a description of the role and activities of the external red team. The company differentiates internal and external red-teaming initiatives. A bug bounty does not constitute an external red teaming regime.
Red teaming coverage of risks	"In designing the regime, they will ensure that they give significant attention to the following: • Bio, chemical, and radiological risks, such as the ways in which systems can lower barriers to entry for weapons development, design, acquisition, or use • Cyber capabilities, such as the ways in which systems can aid vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful • Cyber capabilities, such as the ways in which systems can aid vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system. • The effects of system interaction and tool use, including the capacity to control physical systems • The capacity for models to make copies of themselves or "self-replicate". • Societal risks, such as bias and discrimination"	The company provides full coverage of the risk areas outlined in the voluntary commitments in their red-teaming efforts.

Continued on next page

Scoring Criteria for Indicators – *Continued from previous page*

Indicator	Commitment Text	Criteria
Information sharing with companies	"Work toward information sharing among companies ... regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards"	The information shared must be outside of publicly available knowledge. The information shared must relate to trust and safety risks, AI system capabilities, or attempts to circumvent the safeguards of their models. The primary objective of information-sharing is to advance AI trust and safety. Information-sharing through commercial partnerships do not satisfy this commitment.
Information sharing with government	"Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards"	The information shared must be outside of public knowledge. Information-sharing is facilitated through a sustained and proactive mechanism. Senate testimonies do not fullfill this commitment because the companies did not initiate the enegagement and the engagement is not ongoing. The primary objective of information-sharing is to advance AI trust and safety. The information shared must relate to trust and safety risks, AI system capabilities, or attempts to circumvent safeguards of their models.
Forum or mechanism for information sharing	"They commit to establish or join a forum or mechanism through which they can develop, advance, and adopt shared standards and best practices for frontier AI safety, such as the NIST AI Risk Management Framework or future standards related to red-teaming, safety, and societal risks."	The company establishes or joins a forum that is dedicated to developing standards and best practices for frontier AI safety.
Forum or mechanism shares information on risks	"The forum or mechanism can facilitate the sharing of information on advances in frontier capabilities and emerging risks and threats, such as attempts to circumvent safeguards, and can facilitate the development of technical working groups on priority areas of concern."	The forum restricts who can join and what they do with the shared information. Information on risks should be sensitive information beyond public knowledge that must be controlled and restricted to forum members to prevent misuse or security vulnerabilities.
Model weight cybersecurity practices	"In addition, it requires storing and working with the weights in an appropriately secure environment to reduce the risk of unsanctioned release."	The company implements security controls for their models - specifically, the controls provide protection around the model weights. General best practices for cloud security or cybersecurity do not fulfill this commitment.
Insider threat detection program	"This includes ... establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets."	The company implements insider threat detection programs in relation to their models or AI services.
Limiting weight-level access to relevant personnel	"This includes limiting access to model weights to those whose job function requires it ..."	The company restricts access to model weights to only authorized personnel who require it for their role. Access control is specific to model weight. Restricting access to data and code do not fulfill this commitment as types of assets distinct from model weights.

Continued on next page

Scoring Criteria for Indicators – *Continued from previous page*

Indicator	Commitment Text	Criteria
Establish bounties, contests, or prizes	"They commit to establishing for systems within scope bounty systems, contests, or prizes to incent the responsible disclosure of weaknesses, such as unsafe behaviors, or to include AI systems in their existing bug bounty programs"	The company establishes a bug bounty, contest, or prize. The bounty incentivizes the responsible disclosure of model vulnerabilities and safety issues. The company allows for external participation in the bug bounty. Company-wide bounty programs do not count. The bug bounty, contest, or prize covers their flagship model or AI system. The company provides sufficient information for the public to gauge the activities of the bug bounty, contest, or prize, and mechanisms for participation.
Include AI systems in their existing bug bounty programs	"They commit ... to include AI systems in their existing bug bounty programs"	The company expands their existing bug bounty to include AI systems. This could include covering additional components of their flagship model or including additional categories of vulnerabilities.
Robust provenance or watermarking for audio	"To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed."	The company develops and implements provenance and/or watermarking systems for audio content created by any of their publicly available AI models.
Robust provenance or watermarking for visual content	"To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed."	The company develops and implements provenance and/or watermarking systems for visual content created by any of their publicly available AI models.
Develop tools or APIs to determine if a particular piece of content was created within their tools	"They will also develop tools or APIs to determine if a particular piece of content was created with their system."	The company develops tools to verify if a piece of content was created with their tools.
Work with industry peers and standards-setting bodies towards developing a technical framework	"More generally, companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI."	The company collaborates with others to work towards developing a standard technical framework for content authentication in AI.
Report capabilities	"Publicly report model or system capabilities ...".	The company provides public documentation of their flagship model and system capabilities.
Report limitations	"These reports should include ... significant limitations in performance that have implications for the domains of appropriate use"	The company documents limitations that are specific to their model and systems, and not limitations about AI systems generally.
Report domains of appropriate use	"Publicly report ... domains of appropriate and inappropriate use"	The company publicly reports the intended uses for their models and systems.

Continued on next page

Scoring Criteria for Indicators – *Continued from previous page*

Indicator	Commitment Text	Criteria
Report domains of inappropriate use	"Publicly report ... domains of appropriate and inappropriate use"	The company publicly discloses the unintended and prohibited uses for their models and systems.
Report safety evaluations	"These reports should include the safety evaluations conducted (including in areas such as dangerous capabilities, to the extent that these are responsible to publicly disclose)"	The company publicly reports the safety evaluations conducted on their models and systems. These evaluations target severe safety risks such as CBRN and child safety and not just model toxicity and bias.
Report on societal risks	"These reports should include ... discussion of the model's effects on societal risks such as fairness and bias"	The company describes the effects of model on societal risks, such as fairness and bias, in public reports.
Report on adversarial testing used to determine appropriateness of deployment	"These reports should include ... the results of adversarial testing conducted to evaluate the model's fitness for deployment."	The company publicly reports the results of adversarial testing conducted to evaluate the model's fitness for deployment.
Empower trust and safety teams	"Companies commit generally to empowering trust and safety teams"	The company empowers an explicitly mentioned trust and safety team. Empowerment entails providing the team with sufficient resources and authority to monitor and address potential risks such as bias and misinformation.
Advance AI safety research	"Companies commit generally to ... advancing AI safety research"	The company produces research or develops research tools to implement safeguards and conduct safety evaluations.
Advance privacy	"Companies commit generally to ... advancing privacy"	Efforts to reduce privacy risks must be associated with prompt-conditioned LM generation and/or the storage of user data.
Protect children	"Companies commit generally to ... protecting children, and working to proactively manage the risks of AI so that its benefits can be realized."	"The company partners with organizations such as Thorn and NCMEC to combat CSAM. It designs AI services and releases tools to reduce the risk of AI misuse for child exploitation. Funding external research on CSAM does not constitute protecting children."
Support research and development of frontier AI systems that can help meet society's greatest challenges	"Companies making this commitment agree to support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats."	The contributions the company makes must extend beyond funding to advance research through the deployment of their flagship models, close collaboration, or resource sharing. Initiatives must be driven by public benefit rather than commercial gain. Efforts should target fundamental and widely recognized societal issues.

Continued on next page

Scoring Criteria for Indicators – *Continued from previous page*

Indicator	Commitment Text	Criteria
Support initiatives that foster the education and training of students and workers	"Companies also commit to supporting initiatives that foster the education and training of students and workers to prosper from the benefits of AI"	The company supports initiatives to train students and workers in developing AI literacy and the skills to harness the benefits of AI. This support can be in the form of funding for educational programs dedicated to AI literacy. These educational initiatives should be accessible to all students and workers. These initiatives should be focused on AI literacy and not investments with potential downstream educational impact. These initiatives should not be profit-driven. These investments should be long-term rather than a one-time occurrence.
Support initiatives that help citizens understand the technology	"Companies also commit ... to helping citizens understand the nature, capabilities, limitations, and impact of the technology"	The goal of these initiatives must be to improve citizens' understanding of the nature, capabilities, limitations, and impact of the technology. Initiatives that are focused solely on the public engagement or input process without an educational component are out of scope. These initiatives should provide full coverage of the nature, capabilities, limitations, and impact of the technology. These initiatives should be sustained, rather than a one-time occurrence.

F Limitations

Our analysis has notable limitations. First, our assessment relies on company disclosures, which selectively reflect company practices and may omit internal efforts. We prefer this information as it is directly from companies whereas other sources, such as media coverage, may be biased towards certain high-profile companies. Second, our scoring approach simplifies compliance into binary indicators, which may not capture partial adherence. We focus on the sharp distinction between 0 and 1 to reduce subjectivity. Third, despite codifying our scoring criteria, our interpretation involve subjective judgment. We provide the full set of criteria in Appendix D to enable independent verification. Fourth, our search methodology leverages AI-powered search tools which may produce incorrect, biased, or incomplete information. To ensure accuracy, we manually review and verify all AI-generated outputs. Finally, we score whether companies provide sufficient evidence to meet the commitments but not evaluate the effectiveness or impact of their practices. This focus on evidence of compliance — rather than effectiveness — reflects the scope we chose given the vague commitments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are consistent with the paper's contributions and scope. We assess whether companies follow through on voluntary commitments made to the White House, and find substantial variation in implementation. These findings expose a structural shortcoming in voluntary AI governance approaches, which we discuss further in **??**. Full details of our methodology and findings are provided in **??** and **??**.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper acknowledges several limitations, including the reliance on company disclosures, the potential subjectivity of scorers, and the use of AI-powered search tools. These limitations are outlined in **??** and detailed in **??**.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms with its guidelines. Our study does not involve human subjects or raise data-related concerns, and we do not foresee potentially harmful consequences from this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the policy implications of our work in **??**. Our research into the voluntary commitments leads us to consider future-looking policy design, as well as current corporate practices. This work has potential positive societal impacts by informing stronger governance mechanisms that can influence corporate behavior and ensure public accountability.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: We describe our usage of LLMs in **??**. In our information gathering process, we utilized the Perplexity API to search for additional resources that address how companies assess, mitigate, or communicate risks associated with their generative AI system. Each response was manually reviewed for relevance. We will include our code, along with documentation for reproducing our analyses, in a forthcoming public release.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.