THORN

Safety by Design: Annual Progress Report

REPORT #4: APRIL 2024 TO APRIL 2025

Authors: Dr. Rebecca Portnoff, Michael Simpson

Design & Publication: Yena Lee, Cassie Coccaro, Justus Hyatt

Suggested Citation: Thorn (2025). *Safety by Design for Generative AI Annual Progress Report*. Available at: https://info.thorn.org/hubfs/Thorn_SafetyByDesign_AnnualProgressReport_April2024-April2025.pdf

Table of Contents

Companies' Commitment

All companies that agreed to commit to the Safety by Design principles,[1] or commit to both the principles and the more granular recommended mitigations,[2] also agreed to share the progress they have made in implementing those principles at a regular cadence.

This is the first annual report for Amazon, Anthropic, Google, Meta, Microsoft, Mistral AI, OpenAI, and Stability AI. Civitai, Invoke and Metaphysic have been reporting quarterly; this report also serves as their quarterly report.

As a result of the above, for this fourth public report, we focused our attention on all committed companies (as described in their own words):

- **Amazon** (a company that provides access to, develops, deploys, and/or hosts a variety of first- and third-party models across our services)
- **Anthropic** (an AI safety and research company)
- **Civitai** (a platform for hosting third-party generative AI models)
- **Google** (a company whose mission is to organize the world's information and make it universally accessible and useful)
- **Invoke** (a SaaS solution and OSS platform for AI image generation)
- **Meta** (a company that builds technologies that help people connect, find communities, and grow businesses)
- **Metaphysic** (a business that develops first-party generative AI models to create photorealistic generative AI video content for film studios)
- **Microsoft** (a company that designs trusted, inclusive, and intelligent products that empower people and organizations to achieve more—across cloud, AI, devices, and everyday productivity experiences)
- **Mistral AI** (Mistral AI is a pioneer company in generative artificial intelligence, empowering the world with the tools to build and benefit from the most transformative technology of our time. The company democratizes AI through high-performance, optimized, and cutting-edge open-source models, products and solutions as well as end-to-end infrastructure with Mistral Compute. Headquartered in France and independent, Mistral AI defends a decentralized and transparent approach to technology, with a strong global presence in the United States, United Kingdom, and Singapore.)
- **OpenAI** (an artificial intelligence research and deployment company)
- **Stability AI** (a company that develops generative AI models, as well as custom workflows and advanced editing tools designed primarily for enterprise use)

1 https://www.thorn.org/blog/generative-ai-principles/
2 The recommended mitigations are documented in: Thorn and ATIH. (2024) *Safety by Design for Generative AI: Preventing Child Sexual Abuse*. Thorn Repository. Available at https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

Data Collection Process

Recognizing differing levels of bandwidth, confidentiality constraints, and existing transparency efforts, companies were provided with two options for how to fulfill their transparency commitments.

For those companies with existing relevant public transparency reports in place, some chose to satisfy their transparency commitments with their existing documentation. For those companies, we share the link to those reports in this document.

For other companies, we collected information about the progress it has made taking action on the Safety by Design principles via a survey. This survey requested information on both the steps it has taken in fulfillment of its commitments and metrics to measure the impact of its commitments. In certain circumstances, we also conducted a follow-up interview to gather more detail on survey responses.

Some companies chose to both provide links to existing reports, and engage in the survey process.

For companies that responded to the survey, we indicate how these companies have taken action on the principles based on their survey responses. Where we have the data, we include metrics to measure the impact of these actions to date.

Additionally, for companies that responded to the survey who further chose to engage in a complete (i.e. across all sub-principles) or partial (i.e. across a subset of the sub-principles) analysis process, we also provide complete or partial analysis on what delta currently remains between the actions each company has taken and fulfilling the commitments it has made. Where the analysis was conducted, it appears under the sub-sections labeled “Not Yet Implemented”.

This report documents the data self-reported by companies through the survey and any follow-up interviews, and provides links to relevant existing public transparency reports. Thorn has not independently confirmed, investigated or audited the information provided in these self-reports or the public transparency reports. The data and this report are provided for general informational purposes. Thorn makes no representation or warranty of any kind, express or implied, regarding the accuracy, completeness or reliability of the data or the report, including the warranties of merchantability, fitness for a particular purpose, and non-infringement, and disclaims all liability related to creating, producing and issuing this report. All data provided to Thorn for this report is the property of the company providing such data and may be protected by applicable law. Links to third party websites are for informational purposes only, and the third party is responsible for the content on their website.

To read more about Thorn’s strategy and perspective on accountability in regards to this Safety by Design initiative, see Thorn’s progress report blog.[3]

3 Thorn. “Thorn’s Safety by Design for Generative AI: Progress Reports.” *Thorn*, 21 October 2025, https://www.thorn.org/blog/thorns-safety-by-design-for-generative-ai-progress-reports.

Existing Public Transparency Reports

The companies that chose to satisfy their transparency commitments via existing public transparency reports are listed below, and links to those existing documents are provided.

- **Amazon:** https://www.aboutamazon.com/news/policy-news-views/amazon-csam-transparency-report-2024
- **Anthropic:** https://www-cdn.anthropic.com/0fad284f89c8f9b95ee0f59bdde78928b9a7c425.pdf (hosted on Anthropic's Transparency Hub[4])
- **Microsoft:** https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Addressing-AI-and-Child-Sexual-Exploitation-and-Abuse-Risks-Microsoft%E2%80%99s-Approach.pdf (released as part of Microsoft's Digital Safety Content Report[5])
- **Google:** https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf (Google notes that more on its approach to Responsible AI can be found in its Responsible AI Progress Report, and other related resources[6])
- **Stability AI:** https://stability.ai/news/stability-ais-annual-integrity-transparency-report

Specific Findings

For a summary of progress across all committed companies, as well as additional detail regarding other aspects of Thorn's Safety by Design initiative outside of these progress reports, please see Thorn's "Safety by Design: One year of progress protecting children in the age of AI" blog.[7]

4 https://www.anthropic.com/transparency/voluntary-commitments
5 https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report
6 https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf ; https://transparencyreport.google.com/child-sexual-abuse-material/reporting; https://protectingchildren.google/#introduction; https://support.google.com/transparencyreport/answer/10330933?hl=en-au#zippy=%2Chow-does-google-combat-the-risks-of-csam-in-the-generative-ai-genai-space
7 Thorn. "Safety by Design: One year of progress protecting children in the age of AI." *Thorn*, 21 October 2025, https://www.thorn.org/blog/safety-by-design-one-year-of-progress.

PRINCIPLE 1

DEVELOP: Develop, build and train generative AI models that proactively address child safety risks.

Sub-principle 1: Responsibly source and safeguard our training datasets from child sexual abuse material (CSAM) and child sexual exploitation material (CSEM).

This is essential to helping prevent generative models from producing AIG (AI generated)-CSAM and CSEM. The presence of CSAM and CSEM in training datasets for generative models is one avenue in which these models are able to reproduce this type of abusive content. For some models, their compositional generalization capabilities further allow them to combine concepts (e.g. adult sexual content and non-sexual depictions of children) to then produce AIG-CSAM. We are committed to avoiding or mitigating training data with a known risk of containing CSAM and CSEM. We are committed to detecting and removing CSAM and CSEM from our training data, and reporting any confirmed CSAM to the relevant authorities. We are committed to addressing the risk of creating AIG-CSAM that is posed by having depictions of children alongside adult sexual content in our video, images and audio generation training datasets.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, it undertakes various data preparation and cleaning processes to ensure that training data is of sufficient quality and appropriateness. Anthropic reports it is in the process of adopting interventions to avoid ingestion of CSAM, CSEM, and NCII from its training datasets.

With respect to CSAM, Anthropic reports that it has been actively working to set up hash-matching against industry standard CSAM hashlists, and reporting mechanisms with third parties to prevent CSAM ingestion in its training data. Anthropic further reports that for images from crawl sources, it has begun to leverage a general NSFW filter over its training data, to exclude such content.

Anthropic further reports that between April 15, 2024 and March 31, 2025, it did not train or release models with image, video, or audio output capabilities, and does not train or release open source models.

Civitai

CIVITAI REPORTS

According to Civitai, because it does not develop first-party generative AI models (it provides a platform for hosting of third-party generative AI models), it does not have any training datasets to curate or clean.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented and what it committed to implementing.

Google

GOOGLE REPORTS

According to Google, it developed a process that filters for CSAE at multiple stages during data preparation, to safeguard their training datasets from abusive content. Google notes Gemma as one example where it implemented this process, reporting that it applied data cleaning and filtering methods to Gemma's training data, such as CSAM filtering, at multiple stages in the data preparation process. According to Google, another example of this process is its integration of hash-matching and child safety classifiers to remove CSAM, as well as other exploitative and illegal content, from training datasets.

Google further reports that many of its training datasets are taken from sources already built or integrated with child safety protections, such as Google Search, which reported and removed over 1 million URLs from the Search index for CSAM in 2024 — over 400,000[8] URLs between January and June 2024 and over 880,000[9] between July and December 2024. In 2025, Google reports that it detected a number of instances of CSAM in training datasets during its preparation of these datasets pre-model training, which were subsequently reported to the National Center for Missing and Exploited Children (NCMEC).

Invoke

INVOKE REPORTS

According to Invoke, because it does not develop first-party generative AI models (it provides a SaaS solution and OSS platform for AI image generation), it does not have any training datasets to curate or clean.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented and what it committed to implementing.

Meta

META REPORTS

According to Meta, it conducts dataset filtering to enact this sub-principle for its generative AI models. Meta reports that its policy is to filter any data sources used in research to help ensure training data does not contain images or videos of CSAM. According to Meta, this filtering occurs for all datasets except those that have been fully reviewed by humans or come from sources where Meta has evidence the platform implements industry standard mitigations for monitoring, screening and removing CSAM. Meta reports that it assesses source platforms based on criteria such as whether the platform includes terms prohibiting CSEM and CSAM, and whether the platform publicly represents that it proactively monitors, screens, moderates, and removes CSAM using industry standard tools.

Meta reports that this dataset filtering practice is comprehensive across all its data sources used in research, including the data used for its foundational Llama series (Llama, Llama 2, Llama 3, Llama 4), as well as additional salient models that its Fundamental AI Research division (FAIR) has released, including Chameleon,[10] Segment

8 https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en&lu=urls_deindexed&urls_deindexed=period:2024H1
9 https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en
10 https://ai.meta.com/blog/generative-ai-text-images-cm3leon/

Anything Model (SAM),[11] SAM 2,[12] Meta Motivo,[13] Video Seal,[14] Movie Gen,[15] Audiobox,[16] Seamless Communication,[17] Dinov2,[18] and OpenDAC.[19]

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has four primary strategies to enact this sub-principle. We address each of these strategies below:

1. Studio consent: Metaphysic reports that all data used for its generative AI models is sourced directly from the film studios with which it collaborates. As part of its contracts with these studios, Metaphysics reports that it requires the studio to warrant that no illegal material is present in these datasets.

2. User consent: Metaphysic reports that as part of its contracts with film studios it requires that studios also receive the consent of the individuals depicted in the data. It requires this consent for Metaphysic's use of both the data and its derivatives.

3. Human review: Metaphysic reports that upon receipt of the data, human moderators review every piece of data to confirm that no illegal or unethical content is present in the data.

4. Machine learning (ML)/AI dataset segmentation: Metaphysic reports that it uses proprietary ML/AI detection, to detect and separate out sexual content from depictions of children (such that its generative AI models are not trained on a combination of this content). Metaphysic reports its proprietary models for sexual content detection have an accuracy of around 95%. It reports more difficulty with the tools it uses for age estimation, with performance of these tools generally lower than the tools it uses for detecting sexual content.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining the commitments):

- The percentage of its datasets that have been audited and updated for CSAM and CSEM: 100%; 100%
- The number of instances of CSAM detected in its datasets: 0
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 0 and 0

Metaphysic further reports that in that process, it did not discover any CSAM in its datasets due to the nature of its business model as Metaphysic works exclusively with consensual data provided by clients and studios, and therefore has not submitted any reports to NCMEC or other reporting hotlines.

11 https://segment-anything.com/
12 https://ai.meta.com/sam2/
13 https://metamotivo.metademolab.com/
14 https://ai.meta.com/research/publications/video-seal-open-and-efficient-video-watermarking/
15 https://ai.meta.com/research/movie-gen/
16 https://audiobox.metademolab.com/
17 https://ai.meta.com/research/seamless-communication/
18 https://dinov2.metademolab.com/
19 https://open-dac.github.io/

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it has two primary strategies to enact this sub-principle. We address both of these strategies below:

1. Pre-ingestion dataset filtering: Mistral AI reports that it utilizes proprietary multimodal classifiers to conduct filtering, to ensure problematic content is excluded from training data prior to ingesting the data. With respect to image material, Mistral AI further reports that it employs an approach that uses its classifiers to first remove all nudity content, and does not analyze the data at the content level, in order to minimize the internal team's exposure to sensitive content. Mistral AI chose not to disclose performance metrics for its classifiers.

2. Pre-ingestion domain exclusion: Mistral AI reports it excludes certain domains when collecting data, where those domains are known to be associated with illegal and problematic content.

Mistral AI further reports that it maintains detailed processes for handling CSAM exposure, and has defined specific training data and model development policies with respect to CSAM. According to Mistral AI, it has established training and policies to minimize staff exposure to sensitive content. Holistically, Mistral AI reports that it considers this process of curation and filtering to be continuous for all of its models, one that regularly evolves and is never finished.

Mistral AI reports that it does not currently offer any models with image, video or audio generation capabilities, and relies on its provider Black Forest Labs[20] for image generation.

IMPACT METRICS

Mistral AI reports the following metrics (as measured since joining the commitments):

- The percentage of its datasets that have been audited and updated for CSAM and CSEM: 100%[21]

OpenAI

OPENAI REPORTS

According to OpenAI, it takes steps to remove CSAM and other harmful content from training data, and also reduces processing of personal data in its datasets.

20 https://bfl.ai/models/flux-kontext

21 This metric refers to Mistral AI's NSFW filtering efforts. CSAM contains unique characteristics that may not manifest in other types of content with nudity, including low production value and attempts to obscure the location where it was produced. CSEM may include non-nude images of children in sexualized poses and settings.

Sub-principle 2: Incorporate feedback loops and iterative stress-testing strategies in our development process.

Continuous learning and testing to understand a model's capabilities to produce abusive content is key in effectively combating the adversarial misuse of these models downstream. If we don't stress test our models for these capabilities, bad actors will do so regardless. We are committed to conducting structured, scalable and consistent stress testing of our models throughout the development process for their capability to produce AIG-CSAM and CSEM within the bounds of law, and integrating these findings back into model training and development to improve safety assurance for our generative AI products and systems.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, it evaluates all deployed models, with red teaming for CSAM and CSEM capabilities led by its internal child safety experts. Anthropic reports employees involved in this process are trained on responsible red teaming procedures as well as general reporting and preservation processes for CSAM.

Anthropic further reports that it integrates policy testing commissioned from outside subject matter experts[22] to ensure that its evaluations are robust and take into account new trends in abuse. According to Anthropic, results from internal and external red teaming are provided to its model finetuning and safeguards ("Trust and Safety") teams to assess for integration back into model training, model development, and deployment of safety and enforcement strategies. In some cases, Anthropic reports that it has used this feedback to further update its safety classifiers, enhance its usage policy, and update its internal testing strategy for future models.

IMPACT METRICS

Anthropic reports the following metrics (as measured since joining the commitments):

- The percentage of generative AI models that have been stress-tested for CSAM and CSEM capabilities: 100%; 100%

Civitai

CIVITAI REPORTS

According to Civitai, because it does not develop first-party generative AI models (it provides a platform for hosting of third-party generative AI models), it does not have any first-party models to red team or otherwise stress test.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

22 https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

Google

GOOGLE REPORTS

According to Google, before releasing any models publicly, it conducts testing to the extent legally permissible to identify and resolve potential vulnerabilities, and work to mitigate the possibility of CSAE material being produced or its products being misused to enable CSAE. Google notes that it conducts adversarial child safety testing across text, image, video, and audio for potential risks and violations.

Google further reports that it developed an evaluation approach that may include red teaming, to the extent legally permissible, and child safety adversarial testing. By inserting adversarial prompts into a model or product and evaluating outputs, Google states that it can provide data and feedback to developers. To conduct these child safety evaluations, Google reports that it developed over 19,000 adversarial prompts using resources, such as intel reports, synthetic prompt development, and subject-matter expertise, to target known risk vectors. According to Google, these prompts fit a range of modalities, such as text prompts focusing on grooming or image generation focused on sexualized images, and are developed based on the capabilities of the model or product. Google notes that in 2024, these exercises resulted in the evaluation of over 700,000 model responses. Google reports that these techniques play a critical role in its approach to proactively testing AI systems for weaknesses and identifying emerging risks.

According to Google, upon public release of its models, it continues to monitor adversarial trends and test its models for additional risks and adversarial pivots that may emerge; for example, by using intelligence vendors or social monitoring. Google reports that its approach is continually evolving, incorporating new measurement techniques as they become available as well as insights from resources such as intel reports.

Invoke

INVOKE REPORTS

According to Invoke, because it does not develop first-party generative AI models (it provides a SaaS solution and OSS platform for AI image generation), it does not have any first-party models to red team or otherwise stress test. However, Invoke does report that it has performed red teaming exercises to test the robustness of its internal prompt monitoring solution, validating it in parallel with its previous prompt monitoring solution[23] before migration.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

Meta

META REPORTS

According to Meta, its safety process for building models includes red teaming and systematically evaluating its models for CSE and CSAM-adjacent violations. Meta reports that it mitigates concerns discovered through this process, including through fine-tuning. Meta further reports that because it is illegal under federal law to

23 askvera.io

create or attempt to create CSAM — even in the context of testing or training child safety systems — there are legal limitations to what it can red team for.

According to Meta, it is working within the bounds of existing federal laws to help ensure that its testing is as extensive as legally permissible, and believes that its testing in this space is highly effective in finding and addressing vulnerabilities in its models, as well as protecting against CSAM. Meta reports that while it believes its safety processes have been effective at preventing the creation of violating imagery, it wants to do more to ensure this misuse does not occur and is looking at pathways to make further advancements/improvements to tackle misuse.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has not yet incorporated consistent red teaming into its model development process due to its data governance model, which does not require emergency red teaming.

NOT YET IMPLEMENTED

Metaphysic has not yet:

- Incorporated consistent red teaming for CSAM and CSEM capabilities

Metaphysic notes that the team chose to prioritize the work on data curation instead. The company has stated its intention to begin implementing consistent red teaming into its workflow in early 2025.

IMPACT METRICS

Metaphysic reports that it has conducted two red teaming exercises as “dry-runs” in advance of its planned implementation efforts in 2025.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it does not currently offer any models with image or video generation capabilities, instead relying on its image generation provider Black Forest Labs.

Mistral AI further reports that it plans to enhance its efforts on model capability evaluation, in accordance with the requirements of the EU AI Act. Mistral AI reports that its current CSEM evaluations for its text-generation and multimodal models are automated, conducted using a combination of in house data and external data providers. According to Mistral AI, if a known vulnerability of its models surfaces pre-deployment, it has the governance structure in place to ensure mitigations are put in place, in advance of releasing the model. Mistral AI reports that these mitigations may be at the model level (e.g. post-training or model fine-tuning), or at a different layer of its safety stack, depending on the particular issue that has been discovered.

According to Mistral AI, the company has identified further red teaming as a safety priority for 2025. Mistral AI reports focusing on ensuring compliance and establishing a cohesive red teaming and evaluation process aligned with European regulations.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Mistral AI self-reports having implemented, and what it committed to implementing.

OpenAI

OPENAI REPORTS

According to OpenAI, it has a multi-layered approach to stress testing its models for CSAM and CSEM capabilities, including post-training, red-teaming, and use of safety models. OpenAI reports that this process holistically applies across all of its image, video, audio, and text-generating models.

Sub-principle 3: Employ content provenance with adversarial misuse in mind.

Bad actors use generative AI to create AIG-CSAM. This content is photorealistic, and can be produced at scale. Victim identification is already a needle in the haystack problem for law enforcement: sifting through huge amounts of content to find the child in active harm's way. The expanding prevalence of AIG-CSAM is growing that haystack even further. Content provenance solutions that can be used to reliably discern whether content is AI-generated will be crucial to effectively respond to AIG-CSAM. We are committed to developing state of the art media provenance or detection solutions for our tools that generate images and videos. We are committed to deploying solutions to address adversarial misuse, such as considering incorporating watermarking or other techniques that embed signals imperceptibly in the content as part of the image and video generation process, as technically feasible.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, between April 15, 2024 and March 31, 2025, Anthropic did not have models with image, video, or audio output capabilities.

Civitai

CIVITAI REPORTS

In some cases, Civitai offers access to cloud-hosted third-party generative AI models on its platform. In these cases, Civitai has the necessary access to incorporate content provenance into the generated content (after generation). In cases where third-party generative AI models are cloud-hosted on Civitai's platform, Civitai reports that it currently relies on metadata to assess model origin, where the metadata is incorporated into images and videos generated on the platform, allowing tracking back to the creator using job IDs.

NOT YET IMPLEMENTED

Civitai has not yet:

- Assessed and iterated on its provenance solutions to ensure they are effective with adversarial misuse[24] in mind

24 Solutions that rely exclusively on metadata are vulnerable to adversarial misuse (e.g. metadata stripping).

Civitai reports it is actively exploring options to incorporate content provenance solutions post-generation, pending further industry standardization and technical developments.

IMPACT METRICS

Civitai reports that 100% of images and videos generated on its platform include metadata describing the provenance of that content.

Google

GOOGLE REPORTS

Google reports that in 2024, it joined the Coalition for Content Provenance and Authenticity (C2PA)[25] as a steering committee member, where it partners[26] with others in the industry to develop interoperable provenance standards and technology to explain how content was produced. Google further notes that its internal launch requirements[27] for its applications address risks and include testing and design guidance.

According to Google, once models are deployed into its products and services, applications that generate audiovisual content are required to incorporate a robust provenance solution like SynthID.[28] Google notes that these requirements are based on the nature of the product, its intended user base, planned capabilities, and the types of output involved. For example, Google highlights that an application made available to minors may have additional requirements in areas such as parental supervision and age-appropriate content.

Invoke

INVOKE REPORTS

According to Invoke, all images created within its SaaS solution and OSS platform include metadata that contains a graph describing exactly how the image was created, along with other general metadata about the image. Invoke further reports that this metadata is embedded within the image file itself, and can not be viewed by the majority of photo viewing applications, making it relatively difficult for the average user to remove or change the metadata.

Invoke reports that deciding what metadata to store within the images was a long process, and it continues to regularly assess and update that decision.

NOT YET IMPLEMENTED

Invoke has not yet:

- Assessed and iterated on its provenance solutions to ensure they are effective with adversarial misuse[29] in mind

IMPACT METRICS

Invoke reports that 100% of images created within its SaaS solution and OSS platform include metadata describing the provenance of that image content.

25 https://c2pa.org/google-to-join-c2pa-to-help-increase-transparency-around-digital-content/
26 https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/
27 https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf
28 https://deepmind.google/technologies/synthid/
29 Solutions that rely exclusively on metadata are vulnerable to adversarial misuse (e.g. metadata stripping).

Meta

META REPORTS

According to Meta, it launched its first set of generative AI provenance measures, IPTC metadata[30] as well as visible and invisible watermarking, on Imagine, its first product with photorealistic AI-generated image content. According to Meta, the invisible markers used for Meta AI images align with Partnership on AI (PAI)[31] best practices. Meta further reports that it collaborates with other companies in industry to develop common standards for identifying AI-generated content through forums like PAI.

According to Meta, these approaches reflect what is currently technically feasible, but may not fully address adversarial attempts to e.g. strip out invisible markers. Meta reports that it is pursuing a range of options to help mitigate this, including:

- Develop classifiers that can help it to automatically detect AI-generated content, even if the content lacks invisible markers.
- Research technical strategies to make it more challenging to remove or alter invisible watermarks, e.g. FAIR's Stable Signature.[32]

Meta reports that it continues to invest in research and cross-industry conversations to develop even more robust solutions.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, C2PA is now implemented by default across its data pipelines.

NOT YET IMPLEMENTED

Metaphysic has not yet:

- Assessed and iterated on its C2PA implementation to ensure it is effective with adversarial misuse[33] in mind

IMPACT METRICS

Metaphysic reports that 100% of its generative AI models have been developed with built-in content provenance.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it does not currently offer any models with image or video generation capabilities, instead relying on its image generation provider Black Forest Labs to do so. As such, Mistral AI currently relies on

30 https://iptc.org/standards/photo-metadata/iptc-standard/
31 https://partnershiponai.org/
32 https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/
33 C2PA has built a strong technology foundation for companies to adopt. However, C2PA was not built with adversarial misuse in mind (e.g. it is vulnerable to metadata stripping).

Black Forest Labs' own provenance tools. Black Forest Labs relies on the C2PA[34] protocol for track provenance and modification of digital content. Separately, Mistral AI is currently awaiting final details of the EU AI Act to establish a cohesive machine-readable provenance strategy, aligned with regulation.

NOT YET IMPLEMENTED

Mistral AI has not yet:

- Assessed and iterated on its C2PA implementation to ensure it is effective with adversarial misuse[35] in mind

OpenAI

OPENAI REPORTS

According to OpenAI, it uses C2PA metadata on all assets along with internal tooling to help assess whether a certain image is created by our products.

OpenAI reports it is continuing to work on improving its content provenance solutions in adversarial settings. To address industry-wide challenges in this space, OpenAI further reports it has supported legislative efforts such as the NO FAKES act pending in Congress.[36]

34 https://c2pa.org/
35 C2PA has built a strong technology foundation for companies to adopt. However, C2PA was not built with adversarial misuse in mind (e.g. it is vulnerable to metadata stripping).
36 https://www.congress.gov/bill/118th-congress/senate-bill/4875

PRINCIPLE 2

DEPLOY: Release and distribute generative AI models after they have been trained and evaluated for child safety, providing protections throughout the process.

Sub-principle 1: Safeguard our generative AI products and services from abusive content and conduct.

Our generative AI products and services empower our users to create and explore new horizons. These same users deserve to have that space of creation be free from fraud and abuse. We are committed to combating and responding to abusive content (CSAM, AIG-CSAM and CSEM) throughout our generative AI systems, and incorporating prevention efforts. Our users' voices are key, and we are committed to incorporating user reporting or feedback options to empower these users to build freely on our platforms.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, it has multiple strategies to enact this sub-principle across its Claude model families for its first-party users. We address each of these strategies below:

1. Detection at the inputs: Anthropic reports that it employs hash-matching technology to detect known CSAM that users may upload to its first-party services. Anthropic further reports it runs safety classifiers on prompts and completions to identify harm and violations of its usage policy, which explicitly prohibits the use of its models to facilitate sexual harms against children. According to Anthropic, the safety classifier includes detection of text CSEM. Anthropic reports it is implementing similar tooling for detecting NCII and novel CSAM, pending exploration of adequate technological solutions.

2. Enforcement at the outputs: Anthropic reports that its usage policy explicitly prohibits the use of its models to facilitate sexual harms against children. Anthropic reports it has implemented layered enforcement mechanisms across all Claude model families, which include warnings, prompt modification, and account restrictions. According to Anthropic, in severe cases and/or repeated abuse, it may ban or suspend accounts.

3. Prevention messaging: Anthropic reports that Claude is trained to provide prevention messaging when potential CSAM solicitation is detected.

4. User reporting: Anthropic reports that it has established reporting flows that allow users to flag concerning content or model behavior, including its reporting site[37] which directs users to an email address they can

37 https://support.anthropic.com/en/articles/7996906-reporting-harmful-or-illegal-content

use to report harmful or illegal content. According to Anthropic, users can also report content real-time in-product on Claude.ai through the thumbs up/thumbs down buttons.

Anthropic notes that Claude currently does not produce image or video outputs and is therefore incapable of generating image-based CSAM or NCII. Anthropic further reports that it reviews its usage policies, classification and detection systems, and enforcement processes on a regular cadence.

According to Anthropic, its third-party partners maintain their own screening and detection systems for CSAM, monitor violations, and take appropriate enforcement actions.

IMPACT METRICS

Anthropic reports the following metrics (as measured since joining into the commitments):

- The number of instances of CSAM detected at the inputs: 859
- The number of reports sent to NCMEC for CSAM as a result of the above: 859

Civitai

CIVITAI REPORTS

According to Civitai, its terms of service prohibit CSAM, CSEM, and AIG-CSAM. Civitai reports it has four primary strategies to enact this sub-principle, for those cloud-hosted third-party generative AI models on its platform. Civitai further reports that it does not cloud-host any text-generating models, and its image/video-generating cloud-hosted models do not allow for image/video uploads by users as input to said models.

We address each of these strategies below:

1. Detection at the inputs (i.e. where users submit prompts to the model): Civitai reports that these input-level detection defenses are a layered combination of automated filters and human review of content generation requests and subsequently generated media.

   Civitai reports that it combines keyword detection with ML/AI detection to identify prompts indicating an attempt to produce AIG-CSAM. Civitai's ML/AI prompt detection incorporates information from previous prompts submitted by users, to attempt to capture intent and broader context of the potentially violating prompt. Civitai further reports it is iterating on a new version of this system, and will have accuracy metrics to provide regarding the new system in time for the next report.

   According to Civitai, all prompts that are flagged by the automated filtering system are then sent to human review. For generated media that is confirmed by the human reviewer to be AIG-CSAM, a corresponding report is sent to NCMEC.

2. Enforcement at the outputs: Civitai reports using ML/AI detection to scan all cloud-model outputs for indications of minors, and sexually explicit or mature content. Civitai further reports that these efforts rely on in-house detection models, with reported accuracy rates of 75% to 80%. According to Civitai, all images that are flagged by the automated filtering system are then sent to human review. For generated media that is confirmed by the human reviewer to be AIG-CSAM, a corresponding report is sent to NCMEC. Civitai further reports that violations of its terms of service are enforced via account bans, content takedowns, and hash-based blocking of re-uploads.

3. User reporting: Civitai reports that its users have the ability to report all uploaded content, including user accounts, models, model sample images, reviews, review images, comments, and outputs from cloud-hosted third-party models. Reported media items go into an internal queue for human review, where any verified CSAM and AIG-CSAM is then reported to NCMEC. Civitai further reports that it collects user input through comments, discussions, social media, and Twitch streams.

   According to Civitai, the reporting process for models and other users involves a longer form than the media report, requiring evidence of the violating behavior or capabilities (e.g. timestamps and metadata). For problematic models, a user report further requires evidence that the violative generated content was actually generated by the reported model itself. Civitai reports that once a model has been flagged as problematic, it is removed from access, and added to an internal Civitai hashlist such that future uploads of this same model are automatically blocked.

4. Prevention messaging: Civitai reports that when the automated filters detect that a user is attempting to prompt for AIG-CSAM, the user receives a real-time warning notification. Repeated attempts result in account suspension.

Civitai further notes that defining consistent guidelines around stylized or anime-styled children, in particular distinguishing between innocent and potentially exploitative content, has been challenging. Civitai reports it addresses this challenge by escalating edge cases to its senior moderation team to align internal decisions, and continuously refine its judgement criteria through team review sessions.

NOT YET IMPLEMENTED

We currently do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of violative CSEM prompts detected at the inputs: 338,292
- The number of user reports submitted for various CSAM and CSEM related model violations: 13,156
- The number of prevention messages surfaced due to violative CSEM prompts: 1,691,46038
- The number of instances of AIG-CSAM detected at the outputs: 191
- The number of reports sent to NCMEC for AIG-CSAM as a result of the above: 191

Google

GOOGLE REPORTS

According to Google, it incorporates several preventative measures to detect and respond to CSAE content, including using machine learning to identify CSAE-seeking prompts and uploads, implementing model guardrails to prevent models from producing exploitative outputs, and reporting any confirmed CSAM matches to NCMEC. For example, Google notes that in 2024 it reported to NCMEC more than 600 instances of apparent CSAM uploaded as part of a user prompt to its generative AI products by using hash-matching.

38 Civitai reports that this number is higher than the number of violative prompts detected, because Civitai surfaces prevention messaging earlier in its overall process of establishing intent and broader context of potentially violating prompts.

Google further reports that it employs several feedback reporting mechanisms. According to Google, users have the option of using these mechanisms, such as Gemini's in-product report a problem feature[39] or Google's report content feature,[40] to report any issues.

Invoke

INVOKE REPORTS

According to Invoke, it has four primary strategies to enact this sub-principle, for its SaaS solution. We address each of these strategies below:

1. Detection at the inputs (i.e. where users submit prompts to the model): Invoke reports that its input-level detection defenses are implemented via prompt monitoring, such that Invoke can detect, ban, and report any users attempting to create abusive content on its hosted products. Invoke further reports that it has migrated its input-level prompt monitoring detection to a self-managed solution for detecting abusive inputs. According to Invoke, whenever violations of acceptable use are detected on its platform, it regularly warns, bans, and reports users based on the severity of the attempted generation. Invoke reports that its detection solution errs on the side of false positives vs. false negatives, as the company has not yet identified a case where the solution has missed abusive inputs such that the user inputting the problematic inputs was not reported. Invoke further reports that it has implemented more rigorous fingerprinting and blocking to prevent abusive users who have already been banned from accessing the platform through secondary or alternative accounts.

   According to Invoke, it commits time every day to monitoring the actions detected by the above measures to review and respond to them accordingly.

2. Customer feedback: Invoke reports that it has existing workflows and channels (including support email, ticketing system, and its Discord community) to allow for customer feedback on any and all issues related to the generated media its SaaS solution customers produce using its platform, including any feedback related to content that may contain illegal or unethical material. Invoke further reports that it has published resources for reporting abusive content found, and invited users with concerns to reach out to Invoke's support team with additional details where necessary.

3. Prevention messaging: Invoke reports that when a user is detected attempting to use a model that has been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM) for CSAM, Invoke will subsequently direct the user to redirectionprogram.com.

4. Model suppression: Invoke reports that it makes use of Thorn's hashlist of models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM). According to Invoke, it uses this hashlist to ensure that all uploads of these models are automatically blocked.

In regards to its OSS offering, Invoke reports that it has two primary strategies to enact this sub-principle: prevention messaging and model suppression. According to Invoke, both of these strategies are implemented in its OSS offering, in the same way as they are implemented for its SaaS solution. Invoke further notes that for its

39 https://support.google.com/gemini/answer/13275746?sjid=15516951918566296069-NC
40 https://support.google.com/legal/troubleshooter/1114905?sjid=15516951918566296069-NC

OSS offering, it has found that this form of open source deployment both allows its business to receive far more QA and testing than may be the norm within its field, but also results in its services being leveraged in ways it cannot fully control.

NOT YET IMPLEMENTED

With respect to its SaaS solution, Invoke has not yet:

- Implemented detection for CSAM[41] and image/video CSEM that users may upload to its generative AI systems
- Implemented detection for AIG-CSAM and image/video CSEM that may output from its generative AI models

With respect to its OSS offering, Invoke has not yet:

- Implemented a pathway for users of its OSS offering to report models that generate AIG-CSAM and CSEM, to the appropriate organizations
- Implemented prevention or deterrence efforts to combat and respond to AIG-CSAM that users may prompt for via its OSS offering

Invoke reports it will continue to evaluate ways it can improve its prevention strategy long-term.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- The number of violative CSEM prompts detected at the inputs: 3302[42]
- The number of user reports submitted for various CSAM and CSEM related model violations: 0
- The number of reports sent to NCMEC for AIG-CSAM as a result of the above: 3302

According to Invoke, all of the above metrics are sourced from its SaaS solution, as Invoke does not have telemetry or access to collect metrics for its OSS platform. Invoke further noted that it did not expect the frequency at which people would attempt to perform such actions on a commercially hosted product.

Meta

META REPORTS

According to Meta, it has three main strategies to enact this sub-principle for its cloud-hosted models, e.g. Meta AI and AI Studio. We address each of these strategies below:

1. Detection at the inputs: Meta reports that it has trained its models to identify prompts related to child exploitation or sexualization and subsequently block responses. Meta further reports that for its LlamaAPI offering, it incorporates hash-matching for known CSAM.

2. Enforcement at the outputs: Meta reports that it has policies against child nudity, abuse and exploitation, both real and AI-generated. According to Meta, it enforces its policies by running detection technology on

41 E.g. via using hashing and matching against verified CSAM lists to detect known CSAM as part of input-level detection defenses of its SaaS solution

42 Invoke further reports that, while the number of violative prompts detected continues to increase, in the most recent quarter of reporting it observed a significant drop in number of prompt violations, compared to previous quarters.

AI-generated responses from its cloud-hosted models before they are shown to users. Meta reports that violatory outputs are subsequently blocked, to reduce the likelihood of potentially unsafe experiences. Meta further reports that any violations that meet the threshold for NCMEC reporting are sent to NCMEC in accordance with its legal obligations.

3. Prevention messaging: Meta reports that when prompts related to child exploitation or sexualization are detected as inputs to its models, in addition to blocking responses it further provides the user resources to global hotlines.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has three primary strategies to enact this sub-principle. We address each of these strategies below:

1. Controlled access: Metaphysic reports that no one outside of the employees at Metaphysic has access to its generative AI models. According to Metaphysic, film studios only receive the requested outputs that they have contracted with Metaphysic to produce. Metaphysic reports that this is part of its larger strategy to ensure that, from a business and ethics perspective, the generative AI models it builds are only used to generate content for the specific use case in which it has been contracted.

2. Human moderation: As noted in the analysis on the sub-principle "Responsibly source and safeguard our training datasets from CSAM and CSEM," Metaphysic reports that it employs human moderators to review every piece of received film studio data for illegal and unethical content. Metaphysic similarly reports employing human moderators to review every piece of generated media for the same purpose.

3. Customer feedback: Metaphysic reports that it has existing workflows to allow for customer feedback on any and all issues related to the generated media it produces for its customers, including any feedback related to content that may contain illegal or unethical material.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining into the commitments):

- The number of instances of CSAM detected at the inputs: 0
- The number of instances of AIG-CSAM detected at the outputs: 0
- The number of user reports submitted for various CSAM and CSEM related model violations: 0
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 0 and 0

Metaphysic reports that with the above strategies in place, it has not discovered any CSAM or AIG-CSAM produced by its generative AI models, and therefore has not submitted any reports to NCMEC or other reporting hotlines.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it has three primary strategies to enact this sub-principle, for its cloud-hosted offering (Mistral Cloud). We address each of these strategies below:

1. Detection at the inputs: Mistral AI reports that it employs proprietary ML/AI detection built in-house to flag policy violating text prompts and images sent to its cloud hosted generative models. Mistral AI further reports it previously used Thorn's Safer hash-matching service to detect for known CSAM images uploaded as input to its cloud hosted generative models; it no longer uses Thorn's service. Mistral AI chose not to disclose performance metrics for its classifiers, but it does note that it may be more conservative in how it treats content at different stages across the develop, deploy, maintain life cycle of its generative models.

2. Enforcement mechanism: Mistral AI reports that its terms of service clearly prohibit illegal activities and have a zero tolerance policy regarding CSAM. Mistral AI reports that any generation or attempt to generate CSAM on its services is strictly prohibited. According to Mistral AI, it reports all actual or suspected CSAM to the relevant law enforcement authorities and terminates the account of any user found in violation.

3. User reporting: Mistral AI reports that it has various feedback channels for users to flag content, including direct product feedback in LeChat, email reporting, La Plateforme reporting, and via its community channels on Discord and Reddit. Mistral AI reports that these mechanisms allow users to report any concerns about generated content.

4. Prevention efforts: Mistral AI reports that its terms of service, usage policy, detection at the inputs and enforcement mechanism prevent attempts from users to prompt for AIG-CSAM or provide CSAM and CSEM as input to its models. Mistral AI further reports that it flags and responds to such problematic inputs in its conversational application.

In regards to its open source models, Mistral AI reports that it relies on user reporting, in the same way as it is implemented for its cloud-hosted solution.

NOT YET IMPLEMENTED

With respect to its cloud-hosted offering, Mistral AI has not yet:

- Implemented detection for AIG-CSAM and CSEM that may output from its generative AI models

With respect to its open source models, Mistral AI has not yet:

- Implemented prevention or deterrence efforts[43] to combat and respond to CSAM, AIG-CSAM and CSEM that users may prompt for or provide as input to its models

IMPACT METRICS

Mistral AI reports the following metrics (as measured since joining into the commitments):

- The number of violative CSEM prompts detected at the inputs: Mistral AI reports it is still in the process of auditing its text content, so cannot provide this metric at this time.

43 E.g. by fine-tuning the model responses to surface prevention messaging with violatory prompt requests

- The number of instances of CSAM detected at the inputs: 0
- The number of user reports submitted for various CSAM and CSEM related model violations: 0
- The number of reports sent to NCMEC for CSAM as a result of the above: 0

OpenAI

OPENAI REPORTS

According to OpenAI, it has multiple strategies to enact this sub-principle across its web interface, direct-to-consumer apps, API Platform, and Enterprise offerings. We address each of these strategies below:

1. Detection at the inputs: According to OpenAI, it employs multiple detection solutions to detect and refuse harmful inputs that violate its policies. OpenAI reports that it uses Thorn's CSAM classifier and Safer service to detect both known and novel CSAM uploads to its systems for image and video uploads. OpenAI further reports that it has implemented proprietary ML/AI detection solutions to detect image and video CSEM, as well as text-based safety violations in user prompts, including evidence of grooming and child exploitation.

   According to OpenAI, it leverages a multi-modal moderation classifier it has built in-house to detect and moderate any sexual content that involves minors via text, image, and video input, along with Thorn's CSAM classifier. OpenAI further reports that it has developed and employs a classifier[44] to predict from text and images whether a minor (under the age of 18) is depicted in the content, and restricts edits to images of minors. According to OpenAI, it also implements NCMEC's "Take It Down" program to remove any user-uploaded non-consensual intimate imagery (NCII) of children.

   OpenAI reports that any flags from its detection systems trigger human review. OpenAI reports conducting human review on every flagged image and video, including hash-matched CSAM, to reduce the burden on law enforcement.

2. Enforcement at the outputs: According to OpenAI, its terms of service explicitly prohibit the use of its products for harming children. OpenAI reports it enforces its policies through model refusals, blocking violatory content and user bans, as well as reporting to NCMEC and law enforcement when appropriate. OpenAI further reports that it has developed and employs technology to identify banned users attempting to create new accounts.

   According to OpenAI, it runs Thorn's Safer ML/AI technology as well as the Safer moderation tool to detect and block outputs of AIG-CSAM. OpenAI further reports it deploys its proprietary ML/AI detection solutions to detect and block image and video CSEM, as well as text-based safety violations in model outputs. According to OpenAI, it additionally has a separate tool that detects heightened cases of human review for reports to NCMEC on text-based crimes.

   OpenAI further reports that it bans child nudity and non-nude child exploitation content, with certain thresholds in place for context-specific cases (e.g., medical images). According to OpenAI, in addition to using Thorn's CSAM classifier, it employs a multi-modal moderation classifier it has built in-house to detect and block any generated sexual content that involves minors.

44 Metrics regarding the performance of this classifier can be found here: https://openai.com/index/sora-system-card/

3. User reporting: OpenAI reports that it has an online portal (https://openai.com/form/report-content/) where users can report potential policy violations and illegal content.

With respect to its API Platform offerings, OpenAI reports that repeated violations of OpenAI's CSAM, child sexualization, and other safety policies result in OpenAI banning the developer.

Sub-principle 2: Responsibly host our models.

As our models continue to achieve new capabilities and creative heights, a wide variety of deployment mechanisms manifests both opportunity and risk. Safety by design must encompass not just how our model is trained, but how our model is hosted. We are committed to responsible hosting of our first party generative models, assessing them e.g. via red teaming or phased deployment for their potential to generate AIG-CSAM and CSEM, and implementing mitigations before hosting. We are also committed to responsibly hosting third party models in a way that minimizes the hosting of models that generate AIG-CSAM. We will ensure we have clear rules and policies around the prohibition of models that generate child safety violative content.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, it does not host any third-party generative AI models on its platform.

When considering the first-party models it builds, Anthropic reports it evaluates all deployed models for child safety risks. As noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Anthropic reports that it conducts red teaming for CSAM and CSEM risks, both by its internal child safety experts and external experts. Anthropic further reports it uses a phased deployment approach to ensure thorough testing and limits access before wider release. According to Anthropic, its deployment phases typically include: (1) internal testing with employees only; (2) limited early access with select customers; and (3) graduated general availability.

IMPACT METRICS

Anthropic reports the following metrics (as measured since joining the commitments):

- The percentage of newly hosted generative AI models that have been assessed for their ability to produce AIG-CSAM and CSEM before being made accessible: 100%; 100%

Civitai

CIVITAI REPORTS

Civitai reports that it has established terms of service that prohibit the use and upload of third-party generative AI models on its platform for generating AIG-CSAM, sexually exploitative depictions of children, or photorealistic depictions of minors. According to Civitai, it enforces these policies by employing a combination of human moderation and automated review. Civitai reports it uses a combination of in-house and external solutions (specifically, Hive's Visual Moderation API and Hive's Demographic API) for the automated review, such that the metadata, tags, filenames, and images associated with the generative model are assessed for presence

of minors. In a final pass, these predicted labels are combined with (where relevant) the prompt via Civitai's automated review system, as in some cases harmful model behaviors only appear with certain prompts (that may be additionally surfaced via community reporting).

Civitai further reports that when violative models are identified through user reporting, Civitai takes action by either removing them from the platform (see "User reporting" in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle above) or implementing mitigations to prevent misuse (in addition to account bans). According to Civitai, it utilizes semi-permeable membranes[45] (SPMs) to ensure that cloud-hosted generative AI models do not produce AIG-CSAM or other harmful content as part of its proactive safety measures. Civitai further reports that certain models are restricted to cloud-hosted generation only, where filtering and SPM mitigations can be applied effectively.

Civitai further reports that the volume and variety of uploaded models is an ongoing challenge. In the absence of formal stress testing, Civitai reports that its strategy may delay detection of violatory models. Civitai reports it is working on more proactive methods and plans to expand its automated checks for earlier detection.

NOT YET IMPLEMENTED

Civitai has not yet:

- Implemented CSAM and CSEM capability assessments for newly uploaded third-party generative models, before the models are hosted on its platforms
- Implemented processes to retroactively assess its currently hosted third-party generative models for CSAM and CSEM capabilities, or make use of information[46] collected in child safety sections of a model card to assess where direct assessments are not possible
- Incorporated mitigations for those third-party models on its platform with CSAM[47] and CSEM capabilities that are hosted on its site

Civitai reports that identifying and mitigating third-party models capable of generating child safety violative content remains a complex and evolving challenge. Civitai reports that while comprehensive retroactive assessments of hosted models remain a challenge due to the lack of automated model assessment technology, it has conducted early research evaluating possible scalable approaches — such as leveraging a dedicated GPU cluster to test models with predefined prompts and automated ML/AI detection at the outputs. Civitai reports that while current hardware limitations prevent full-scale implementation, it anticipates launching a beta system later this year.

Civitai further reports that it is actively working toward incorporating mitigations for Stable Diffusion 1.5 models and its derivatives hosted on its platform. Civitai reports that it has successfully implemented SPM mitigation in its cloud-hosted generative AI models and it now aims to extend these safeguards to all hosted models on

45 Lyu, Mengyao, et al. One-Dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications. *CVPR, 2024*. https://doi.org/10.48550/arXiv.2312.16145.

46 E.g. include questions to third-party developers on what technologies were used to implement data cleaning and curation, where answers with insufficient detail, or other indications that the provided response is false, could result in Civitai disallowing the model to be hosted on its platform

47 E.g. Stable Diffusion 1.5 and its derivatives. These models have been confirmed (see Thiel, D., Stroebel, M., and Portnoff, R. (2023) Generative ML and CSAM: Implications and Mitigations. Stanford Digital Repository. Available at https://doi.org/10.25740/jv206yg3793) as capable of generating AIG-CSAM. As of November 2024, Stable Diffusion 1.5 was the most popular base model used by offenders on the dark web dedicated to the sexual abuse of children. (see Partnership on AI (2024). *Mitigating the risk of generative AI models creating Child Sexual Abuse Materials: An analysis by child safety nonprofit Thorn*. Available at https://partnershiponai.org/wp-content/uploads/2024/11/case-study-thorn.pdf)

its platform, to ensure a consistent and effective approach to mitigating potential risks. Civitai reports it has evaluated the integration of its SPM technology in these additional models to enhance safeguards and eliminate the capability to produce AIG-CSAM. According to Civitai, its current approach has been implemented on a single series of models and testing is being done to expand the safety enhancement to the entirety of the Stable Diffusion 1.5 ecosystem of models.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of hosted generative AI models taken down and removed from platform access, due to discovering they are capable of producing AIG-CSAM and CSEM: 282
- Of those models, the number of models for which mitigations were incorporated and the model was re-uploaded: 15

Google

GOOGLE REPORTS

According to Google, its approach to AI is grounded in its AI Principles,[48] which guide the safety and reliability of Google AI products, focusing on oversight, due diligence, and feedback mechanisms. Google notes that these principles ensure it aligns with user goals, social responsibility, and widely accepted principles of international law and human rights.

Google further reports that its policies and procedures[49] for mitigating harm in areas such as child safety have been informed by years of research, user feedback, and expert consultation. According to Google, these policies guide Google's models and products to minimize certain types of harmful outputs and dictate behavior that is prohibited on its products.

Google reports that its Generative AI prohibited use policy[50] states: "Do not engage in dangerous or illegal activities, or otherwise violate applicable law or regulations. This includes generating or distributing content that relates to child sexual abuse or exploitation." Google notes that as part of its responsible AI approach, it expects to iterate[51] on these policies as both the technology and the risk landscape evolve. Google highlights that these policies are also incorporated into its Cloud Terms of Service,[52] along with its Cloud Acceptable Use Policy[53] which prohibits using Google Cloud Services "to engage in, promote or encourage illegal activity, including child sexual exploitation, child abuse, or terrorism or violence that can cause death, serious harm, or injury to individuals or groups of individuals."

According to Google, it employs a gradual approach to deployment as a critical risk mitigation. Google further notes that it employs a multi-layered approach — starting with testing internally, then releasing to trusted testers externally, then opening up to a small portion of its user base[54] (noting as an example, that it may release models to Gemini Advanced users first). Google reports that it phases its country and language releases,

48 https://ai.google/responsibility/principles/
49 https://transparency.google/?_gl=1*1ly846y*_up*MQ..*_ga*OTk4MTU3MjkwLjE3Mzk4MDg3Mjc.*_ga_7VR0QEE3V8*MTczOTgwODcyNy4xLjAuMTczOTgwODcyOS4wLjAuMA.
50 https://policies.google.com/terms/generative-ai/use-policy
51 https://blog.google/feed/were-updating-our-generative-ai-prohibited-use-policy/
52 https://cloud.google.com/terms/
53 https://cloud.google.com/terms/aup
54 https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf

constantly testing to ensure mitigations are working as intended before expanding. Google further notes that it has protocols and additional testing and mitigations required before a product is released to users under the age of 18.

Invoke

INVOKE REPORTS

According to Invoke, it does not serve as a platform for third-party developers to distribute or merchandise their models, nor does it build any first-party models. However, Invoke reports it has proactively established terms of service that prohibit customer use of Invoke's services in a way that violates any law, regulation or court order, including the use of third-party generative AI models (within its SaaS solution and OSS systems) to generate AIG-CSAM and other sexually exploitative depictions of children. Invoke further reports it has established user policies and enforcement mechanisms around the upload and subsequent use of models that are capable of generating AIG-CSAM (such as Stable Diffusion 1.5 models and its derivatives), as noted in the discussion around the principle "Safeguard our generative AI products and services from abusive content and conduct."

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

Meta

META REPORTS

When considering the first-party models it builds, as noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Meta reports it conducts adversarial red teaming as part of its standard process, working to mitigate concerns discovered through this process, including through fine-tuning. Meta further reports that the Acceptable Use Policy explicitly prohibits the use of its open source models for "exploitation or harm to children, including the solicitation, creation, acquisition, or dissemination of child exploitative content or failure to report Child Sexual Abuse Material."

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not host any third-party models. However, when making use of third-party models internally, Metaphysic reports that it assesses every model, prior to using said model, for a variety of ethical issues (including child safety violations). If any issues are found, Metaphysic does not use the model.

When considering the first-party models it builds, as noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Metaphysic reports it has not yet incorporated red teaming into its processes. However, Metaphysic does report that it practices non-CSAM/CSEM specific model assessment and phased deployment of its models. According to Metaphysic, this model assessment is currently manual. Metaphysic reports it is working towards solutions to conduct these assessments systematically and in an automated fashion, but doing so requires significant resources to build. Finally, as noted in the discussion on "Safeguard our generative AI products and services from abusive content

and conduct," Metaphysic reports that no individuals or organizations outside of Metaphysic have direct access to its generative AI models.

NOT YET IMPLEMENTED

Metaphysic has not yet:

- Implemented CSAM and CSEM capability assessments for its first-party models before hosting

IMPACT METRICS

Metaphysic reports that 100% of its first-party models undergo phased deployment.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it does not host any third-party models on its platform.

When considering the first-party models it builds, as noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Mistral AI reports it has not yet incorporated comprehensive external red teaming into its processes. However, Mistral AI does report that it performs systematic internal red teaming for CSAM and CSE and practices model assessment, further noting that as a European company, it is subject to the enforcement of the AI Act (which makes model assessment a legal requirement). Mistral AI further reports that as a European company that is committed to fostering open source, it intends to continue to align its practices around assessment against the requirements of the regulation to come.

Mistral AI further reports that it practices phased deployment of its models. According to Mistral AI, it typically begins with deployment to 5% of users, observes metrics, verifies that everything is performing as expected, and then gradually increases to 100% of traffic. Mistral AI reports that this monitoring occurs on Mistral Cloud, where it tracks observability metrics for live performance alongside upstream evaluations.

NOT YET IMPLEMENTED

Mistral AI has not yet:

- Implemented processes to retroactively assess currently hosted first-party models (that have not already been assessed) for CSAM and CSEM capabilities, and update them with mitigations where necessary
- Confirmed what interventions its third-party image/video provider Black Forest Labs has in place to minimize these image/video models' CSAM and CSEM capabilities, and assessed and iterated on whether these strategies are effective

Mistral AI reports it is currently awaiting final details of the EU AI Act to ensure compliance and to establish a cohesive red teaming and evaluation process aligned with regulation.

IMPACT METRICS

Mistral AI reports the following metrics (as measured since joining the commitments):

- The percentage of newly hosted generative AI models that have been assessed for their ability to produce CSEM before being made accessible: 100%

OpenAI

OPENAI REPORTS

According to OpenAI, it does not host any third-party models on its platform.

Sub-principle 3: Encourage developer ownership in safety by design.

Developer creativity is the lifeblood of progress. This progress must come paired with a culture of ownership and responsibility. We encourage developer ownership in safety by design. We will endeavor to provide information about our models, including a child safety section detailing steps taken to avoid the downstream misuse of the model to further sexual harms against children. We are committed to supporting the developer ecosystem in their efforts to address child safety risks.

Anthropic

ANTHROPIC REPORTS

Anthropic reports that it includes information on child safety testing in its model documentation and has incorporated a child safety section into its model cards during the reporting period.

Civitai

CIVITAI REPORTS

Civitai reports it requires all model uploads to comply with its terms of service, which explicitly prohibit CSAM, AIG-CSAM, and related harms. Civitai further reports that its built-in moderation systems and community feedback options serve to hold developers accountable for unsafe content, and support the broader goal of ensuring developers consider downstream risks before uploading. According to Civitai, one ongoing challenge is early user education, as not all users understand the risks of uploading or training models. Civitai reports it is working to improve its upload guidelines and flagging systems to catch issues sooner.

NOT YET IMPLEMENTED

Civitai has not yet:

- Incorporated a child safety section into its model cards[55] detailing steps taken by third-party model developers to avoid the downstream misuse of the model to further sexual harms against children

According to Civitai, it is actively evaluating ways to enhance transparency and safety in model submissions. Civitai reports that one approach under consideration is requiring developers to confirm that their training data has been properly curated and cleaned, in line with the "Develop" sub-principle "Responsibly source and safeguard our training datasets from CSAM and CSEM." Civitai further reports a key blocker is finding the most effective format, such that it is aligned with moderation policies, avoids friction for compliant users, and the additional information included enhances safety without inadvertently guiding bad actors toward models lacking proper safeguards.

55 Civitai's model card equivalent consists of the model "details" section for third-party model developers to fill in before they upload their model.

As part of this effort, Civitai reports it is also exploring how submitted safety information could be used as a factor in approving models for hosting, aligning with its broader commitments to responsible deployment as stated in the "Deploy" sub-principle "Responsibly host our models."

Google

GOOGLE REPORTS

Google reports that it regularly publishes external model cards and technical reports as transparency artifacts. According to Google, its technical reports[56] provide details about how its most advanced AI models are created and how they function. Google further notes that this includes offering clarity on the intended use cases, any potential limitations of the models, and how its models are developed in collaboration with safety, privacy, security, and responsibility teams.

Google additionally reports that it publishes model cards[57] for its most capable models and open models. According to Google, these cards offer summaries of technical reports in a "nutrition label" format to surface vital information needed for downstream developers or to help policy leaders assess the safety of a model.

Invoke

INVOKE REPORTS

Invoke reports that it does not develop first-party models, nor does it serve as a platform for third-party developers to distribute or merchandise their models, and therefore it does not make use of model cards in either capacity. In regards to the third-party models that are uploaded to its SaaS solution or OSS offerings, according to Invoke it offers a "Name" and "Description" field to customers, where customers can choose to input details regarding the third-party model they are using.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing. However, there is an opportunity here for Invoke to point customers towards other existing documentation (e.g. model cards included on model hosting platforms) that provide more context and relevant information to its customers regarding what child safety interventions were put into place as part of model development.

Meta

META REPORTS

Meta reports that it provides a Developer Use Guide, formerly known as its Responsible Use Guide, as a resource to support developers in building with Llama safely and in line with best practices. According to Meta, this guide includes a detailed overview of Llama 4 models, information on system-level safety alignment and best practices, and responsibility considerations for building responsible agents using model reasoning.

56 https://arxiv.org/pdf/2408.07009
57 https://modelcards.withgoogle.com/

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has incorporated into its datasets and models an associated card. Metaphysic reports that this card contains information listed in the "Model Card: Child Safety" additional resource included in Thorn & ATIH's paper associated with this Safety by Design initiative.[58]

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports that 100% of its datasets and models have the above described card implemented.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it has established clear safety policies, guidelines and training for its internal team. Mistral AI is currently awaiting final details of the EU AI Act to incorporate a child safety section into its model cards, which is aligned with regulation and which details steps taken to avoid the presence of CSAM and CSEM in its training data sets.

NOT YET IMPLEMENTED

Mistral AI has not yet:

- Incorporated a child safety section into its model cards detailing steps taken to avoid the downstream misuse of the model to further sexual harms against children

OpenAI

OPENAI REPORTS

OpenAI reports that it encourages developer ownership in safety by design through close collaboration and communication with its product team.

58 Thorn and ATIH. (2024) *Safety by Design for Generative AI: Preventing Child Sexual Abuse*. Thorn Repository. Available at https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

PRINCIPLE 3

MAINTAIN: Maintain model and platform safety by continuing to actively understand and respond to child safety risks.

Sub-principle 1: Prevent our services from scaling access to harmful tools.

Bad actors have built models specifically to produce AIG-CSAM, in some cases targeting specific children to produce AIG-CSAM depicting their likeness. They also have built services that are used to "nudify" content of children, creating new AIG-CSAM. This is a severe violation of children's rights. We are committed to removing from our platforms and search results these models and services.

Anthropic

ANTHROPIC REPORTS

Anthropic reports that it did not offer search functionality during this reporting period.

Civitai

CIVITAI REPORTS

According to Civitai, known and verified problematic models (discovered via user reporting) are removed from access, and added to an internal Civitai hashlist such that future uploads of this same model are automatically blocked. Additionally, Civitai reports it retroactively checks its existing corpus of hosted models, running a daily batch job to detect and remove said models such that the newly discovered problematic models do not appear anywhere else in its collection.

Civitai further reports it has updated its policies such that any AI workflows, models, or tools designed with the intention of removing clothing or otherwise "nudifying" individuals in any context ("real people" or otherwise) is explicitly prohibited. According to Civitai, it has incorporated periodic manual moderation efforts to enforce these policies, where moderators will search for indicators in the title and description of resources and workflow that indicate they violate Civitai's policies regarding nudifying individuals.

Additionally, Civitai reports that its existing policies against "suggestive" or "sexual" content depicting real people, combined with its use of prompt filters and SPMs for cloud-generated images captures a significant portion of nudifying activity.

NOT YET IMPLEMENTED

Civitai has not yet:

- Implemented processes to consistently detect and remove from its model hosting platform those services[59] that are used to "nudify" content of children

59 E.g. nudifying AI workflows, models and tools

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of models optimized to produce AIG-CSAM, retroactively removed from its platforms: 258
- The number of prevented attempts to upload to its platforms a model optimized to produce AIG-CSAM: 380
- The number of nudifying services retroactively removed from its platforms: 19

Google

GOOGLE REPORTS

According to Google, one way it takes action to fight CSAM online is by reporting and removing URLs from its Search index (Google Search aggregates and organizes information published on the web). Google notes that it does not have control over the content on third-party web pages. Google further reports that when it identifies CSAM on third-party web pages, it reports, de-indexes and removes that URL from Search results, but has no ability to remove the content from the third-party page itself. According to Google, Google Search reported and removed over 1 million URLs from the Search index for CSAM in 2024 — over 400,000[60] URLs between January and June 2024 and over 880,000[61] between July and December 2024. Google notes that this metric is a combination of both automated and manual removals.

According to Google, its policy is to block search results that lead to CSAM that appears to sexually victimize, endanger, or otherwise exploit children.[62] Google notes that this includes AIG-CSAM. According to Google, it constantly updates its algorithms to combat these evolving threats. Google further reports that it always removes CSAM when it is identified and demotes[63] all content from sites with a high proportion of CSAM content.

Google reports that it has made significant updates[64] in Search to help people affected by non-consensual sexually explicit fake content. Google further reports that these changes were developed based on feedback from experts and victim-survivors. According to Google, these changes include updates to its removal processes to make it easier for people to remove this content from Search and updates to its ranking systems to keep this type of content from appearing high up in Search results.

Invoke

INVOKE REPORTS

According to Invoke, it makes use of Thorn's hashlist of models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM). According to Invoke, it uses this hashlist to ensure that all uploads of these models (in both its SaaS solution and its OSS offering) are automatically blocked (see "Model suppression" in the "Safeguard our generative AI products and services from

60 https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en&lu=urls_deindexed&urls_deindexed=period:2024H1
61 https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en
62 https://support.google.com/transparencyreport/answer/10330933?hl=en#zippy=%2Cwhat-does-google-do-to-deter-users-from-seeking-out-csam-on-search
63 https://developers.google.com/search/docs/appearance/ranking-systems-guide#removals
64 https://blog.google/products/search/google-search-explicit-deep-fake-content-update/

abusive content and conduct" sub-principle above). Invoke further reports that it retroactively checks uploaded models in its SaaS solution, when new models are added to Thorn's hashlist of models.

In regards to the upload and use of models that are intended for "nudifying" imagery that are used to create AIG-CSAM, Invoke reports that it does not have the necessary contextual information (e.g. the advertising language used by the provider of the model indicating it is a "nudifying" model) to reliably distinguish between a model that has been built for the express purpose of "nudifying" imagery vs. models that are capable of nudifying imagery but were not built for that express purpose. As a result, Invoke reports that the user policies and enforcement mechanisms noted in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle outline its strategy for addressing this type of misuse.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- The number of prevented attempts to upload a model optimized to produce AIG-CSAM: 0
- The number of models optimized to produce AIG-CSAM, retroactively removed: 0

According to Invoke, the above metrics are sourced from its SaaS solution offering, as Invoke does not have telemetry or access to collect metrics for its OSS platform. Invoke further notes that it has never had a user attempt to upload to its SaaS solution system a model from Thorn's hashlist of models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM).

Meta

META REPORTS

According to Meta, it offers search functionality via Facebook, Instagram, Whatsapp and Meta AI.[65] Meta further reports its social platforms include Facebook, Instagram, Messenger, WhatsApp, Threads, and Horizon.

Meta reports that its Community Standards, applicable across multiple Meta technologies, prohibit adult nudity, sexual activity, and content or activity that sexually exploits or endangers children. Meta further reports these standards include a specific policy that prohibits the promotion of NCII apps and services. According to Meta, its content guidelines for apps made available on the Horizon store also prohibit this content. Meta reports it removes services, applications, or instructions that promote, threaten to share, or offer to make NCII when Meta becomes aware of them, even if there is no nude or near-nude commercial or non-commercial imagery shared, including those referred to as "nudify apps." According to Meta, it updated its policies last year to account for this evolving trend and prohibit these services.

Meta reports it enforces these policies primarily using automated tools to check apps, ads and business assets against its policies. Meta further reports it applies techniques used in combating other coordinated adversarial spaces to accounts promoting NCII services, such as identifying and taking action against coordinated networks.

65 https://about.fb.com/news/2025/04/introducing-meta-ai-app-new-way-access-ai-assistant/

According to Meta, its ad review process may include review of specific components of an ad, such as images, video, text and targeting information, as well as associated landing pages or other destinations. Meta reports this review process starts automatically before ads begin running and is typically completed within 24 hours. According to Meta, ads may be reviewed again after they are live and may be rejected or restricted for violating policies at any time. Meta reports that ad providers who fail to comply may have their accounts terminated.

Meta notes that reviewing ads from millions of advertisers globally against its Advertising Standards can present challenges, particularly given the highly adversarial nature of this space, e.g. offenders creating new domain names to host applications after previous websites have been blocked. To help address these challenges, Meta reports it monitors and assesses new risks, regularly evaluating its policies and works to improve enforcement mechanisms to address these evolving tactics.

Meta further reports that it is a founding member of Lantern,[66] a program from the Tech Coalition that enables tech companies to share signals about predatory accounts and behaviors, such that participating companies can use this information to conduct investigations on their own platforms and take action. According to Meta, it has recently begun signal sharing on Lantern for nudify apps.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not host third-party models or services, or offer search functionality as part of its business model.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it does not host third-party models. Mistral further reports that it relies on a third-party provider (Brave) for its user-facing search engines. Brave does active scans for child sexual abuse material (CSAM), both internally and using a third party (ActiveFence) and blocks such content.

OpenAI

OPENAI REPORTS

According to OpenAI, it does not host third-party models, but it does offer search functionality as part of its services.

With respect to search, OpenAI reports it has incorporated the Internet Watch Foundation (IWF) URL watchlist[67] into its ChatGPT search functionality, to prevent access to web content that IWF has confirmed contains images and videos of child sexual abuse.

66 https://www.technologycoalition.org/newsroom/announcing-lantern
67 https://www.iwf.org.uk/our-technology/our-services/url-list/

In regards to removing search results for services that are used to "nudify" content of children, OpenAI reports it has conducted investigations to identify these problematic sites and has implemented both

whitelists and blacklists to ensure those disallowed sites are not surfaced to users via ChatGPT's search functionality. According to OpenAI, its models are trained not to surface content from the dark web.

OpenAI further reports its policies explicitly prohibit the creation of "sexually explicit or suggestive content" and ban tools targeting minors.

Sub-principle 2: Invest in research and future technology solutions.

Combating child sexual abuse online is an ever-evolving threat, as bad actors adopt new technologies in their efforts. Effectively combating the misuse of generative AI to further child sexual abuse will require continued research to stay up to date with new harm vectors and threats. For example, new technology to protect user content from AI manipulation will be important to protecting children from online sexual abuse and exploitation. We are committed to investing in relevant research and technology development to address the use of generative AI for online child sexual abuse and exploitation. We will continuously seek to understand how our platforms, products and models are potentially being abused by bad actors. We are committed to maintaining the quality of our mitigations to meet and overcome the new avenues of misuse that may materialize.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, it has invested in research via its work on AI alignment, interpretability, and preventing harmful deployment, as detailed in its Transparency Hub.[68] Anthropic further reports that its Societal Impacts team undertakes research on the values built into AI and how these can express themselves to help prevent harmful deployments or use cases.

With respect to technology solutions, Anthropic reports that when new capabilities are released, its safeguards team deploys classifiers and detection systems to prevent malicious deployment or use of new capabilities. Anthropic reports that prior to launching computer use capabilities, it built custom tooling to detect and prevent prompt injection.

Anthropic further reports that it leverages Open Source Intelligence (OSINT) capabilities by working with a third-party vendor that sends alerts related to general platform abuse to its internal team. According to Anthropic, its in-house Safeguards experts additionally monitor public forums and analyze emerging abuse patterns.

Civitai

CIVITAI REPORTS

According to Civitai, it has invested in and deployed future technology solutions via its line of work around SPM-based interventions. Civitai further reports that it monitors its user community for emerging risks and relies

68 https://www.anthropic.com/transparency

on outside partners to also monitor trends and emerging risks. Civitai reports it regularly updates its tagging, metadata checks, and moderation systems to respond to new forms of abuse. Additionally, Civitai reports continuous effort improving the ML/AI detection technology it builds in-house, evaluating changes in model behavior over time and adjusting its internal review tools accordingly. Civitai further reports that staff are trained to apply updated standards as policies and technology evolve.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- How much investment (e.g. R&D time, grants/funding, etc.) has been made into research and/or technology development to address the use of generative AI for online child sexual abuse and exploitation: ~25% of all development time has been spent optimizing and improving moderation tools, accounting for nearly 20% of payroll costs over the last year, with 2 Contractors and 1 FTE hired in the last year to support these efforts
- Cadence at which mitigations are assessed against the business's tech stack, to ensure effective performance: Quarterly

Google

GOOGLE REPORTS

According to Google, one of its key priorities is protecting against, and responding to, new and unique risks for potential child sexual abuse and exploitation (CSAE) that generative AI might pose. Google reports that it invests heavily in fighting CSAE online and employs a combination of automated detection tools and specially trained reviewers working around the clock to deter, detect, remove, and report content that is illegal or violates its policies on its platform, including technology-facilitated CSAE. Google reports several examples of ways that it has invested in combating CSAE:

- Robust Open Online Safety Tools (ROOST): ROOST[69] is a cross-industry initiative which aims to build scalable, interoperable safety infrastructure ready for the AI era. According to Google, it collaborates with companies such as OpenAI, Discord, and Roblox through the ROOST initiative to share technology and help organizations of all sizes create safer online platforms. Google further reports that this includes providing access to AI safety tools for detecting, reviewing, and reporting CSAE that will work to close the digital safety gap.
- Researching best practices and new solutions, including using AI, for addressing CSAE: According to Google, it works with the Digital Trust & Safety Partnership[70] to research and develop industry best practices for using AI towards detecting and removing policy-violative generative AI content, such as AIG-CSAM. Google notes that, while human oversight remains crucial, AI offers significant potential to enhance the moderation of harmful CSAE content and improve child safety by improving efficiency

69 https://roost.tools/
70 https://dtspartnership.org/

in tackling CSAE while reducing human reviewers' exposure to content that could cause psychological harm.

- Using research-backed insights to enhance its approach: According to Google, between 2023 and 2024, Google commissioned a number of internal reports focused on CSAE with an external provider of expert intelligence insights, with several focused specifically on generative AI and how bad actors are using or planning on using AI to abuse children. Google further notes that these reports were then used to inform Google's approach in a variety of ways, including to update adversarial prompts for child safety evaluations and aid in further understanding and addressing the risks identified.
- Ad Grants: According to Google, it offers Ad Grants Program to make it easier for organizations working to fight against CSAE to reach victims and help report child safety concerns. Google notes that it offers up to $10K of free advertising every month to reporting hotline operators or victim support providers.
- Funding and technical guidance: According to Google, it sends Google engineers to expert child safety organizations to help increase their technical capacity via its Googler in Residence Program. Google further reports that it additionally funds technical fellowships at organizations dedicated to fighting child sexual abuse, and invested[71] in funding to promote teen safety and wellbeing in Europe.
- Google's Child Safety Toolkit: According to Google, since 2014 its partners (including Snap and Adobe) have used its free Child Safety Toolkit (consisting of the Content Safety API and CSAI Match) to analyze billions of images and videos each month for potential CSAE. Google further reports that in October 2024, the Content Safety API was updated to offer a feature that allows partners that either cannot send raw image bytes to Google, or have high-volume requirements, to leverage its CSAM prioritization tooling.
- Tech Coalition: The Tech Coalition[72] is a global alliance of technology companies working to end online CSAE. According to Google, during 2023 and 2024, as a member of the Tech Coalition, Google led several working groups focused on understanding child safety risks in generative AI. Google further reports that as a result of these working groups, the Tech Coalition developed member resources, including a reporting template for industry reports of AI-generated CSAE to NCMEC. In addition, Google notes that through its membership, it also supports the Tech Coalition's research initiatives, including research[73] supported via the Tech Coalition Safe Online Research Fund focused on generative AI and CSAE.

Invoke

INVOKE REPORTS

According to Invoke, it has invested in and deployed future technology solutions via its work building its own detection mechanisms and systems. Invoke further reports it has developed new checks for blocked accounts based on card fingerprints from payment processors to prevent repeat abusers from accessing its platform.

In addition, Invoke reports that it leverages its access to OSINT using forums such as Github, Reddit, Discord etc. to monitor for emerging risks. Invoke further reports that all new features created on its platform are architected with the explicit goal of avoiding the creation of abusive content in mind.

71 https://blog.google/technology/families/new-10m-funding-to-promote-teen-safety-and-wellbeing-in-europe/
72 https://technologycoalition.org/
73 https://technologycoalition.org/news/tech-coalition-announces-new-generative-ai-research/

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- How much investment (e.g. R&D time, grants/funding, etc.) has been made into research and/or technology development to address the use of generative AI for online child sexual abuse and exploitation: $244,000 in R&D time and tools
- Cadence at which mitigations are assessed against the business's tech stack, to ensure effective performance: Daily

Meta

META REPORTS

According to Meta, it has invested in research and technology solutions through the development and open sourcing of multiple tools to combat child sexual exploitation, including image and video matching technology[74] and content moderation tools.[75] Meta further reports it is currently collaborating with NCMEC on a technology that aims to increase the detection of AIG-CSAM, and plans to share its learnings with industry when it is further in the process. Meta also reports that it worked with NCMEC to develop the Take It Down[76] tool, which gives teens control of their nude or near-nude images, including AI-generated ones, and helps prevent online sharing of those images.

According to Meta, it has further deployed technology solutions through its Llama Protections umbrella project, which brings together tools and evaluations to help the community build responsibly with open generative AI models. Meta reports that it makes available safety tooling such as Llama Guard 4, which supports protections for text and image understanding across modalities and was aligned to safeguard against the standardized MLCommons hazards taxonomy,[77] including CSE and sexual content. Meta further reports that it has released Llama Firewall, a security guardrail tool to detect system risks such as prompt injections, and Llama Prompt Guard 2, which helps prevent jailbreaks and prompt injection attempts.

Meta additionally reports that it is working to make the AI ecosystem even safer through content provenance resources, including launching its Llama Defenders Program[78] in April 2025, which includes new audio watermarking and detection technology that provides industry-leading detection performance on accuracy, imperceptibility, and speed.

With respect to research, Meta reports that it conducts internal research to understand how its products may be used by bad actors in order to ensure its protections remain impactful in this adversarial space.

Meta reports that it is an active member of the Tech Coalition, where industry collaborates to combat child exploitation, including generative AI-facilitated exploitation. Meta reports that it is a founding member of

74 https://about.fb.com/news/2019/08/open-source-photo-video-matching/
75 https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/
76 https://takeitdown.ncmec.org/
77 https://mlcommons.org/ailuminate/
78 https://ai.meta.com/blog/ai-defenders-program-llama-protection-tools/

Lantern,[79] a program from the Tech Coalition that enables tech companies to share signals about predatory accounts and behaviors, such that participating companies can use this information to conduct investigations on their own platforms and take action. Meta reports it provided the Tech Coalition with the technical infrastructure behind the Lantern program and continues to maintain it.

Metaphysic

METAPHYSIC REPORTS

As noted in the discussion on "Safeguard our generative AI products and services from abusive content and conduct," according to Metaphysic no individuals or organizations outside of Metaphysic have direct access to its generative AI models. As a result of this controlled access, Metaphysic reports it has not made use of OSINT or other strategies to understand how bad actors are potentially misusing its products and services.
In regards to investing in research and technology, Metaphysic reports that it intends to publish its findings around its efforts to build ML/AI dataset segmentation technologies. Metaphysic further reports (as outlined in the discussion on "Responsibly host our models") its investment in building scalable, automated model assessment mechanisms.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining into the commitments):

- Cadence at which mitigations are assessed against the business's tech stack, to ensure effective performance: Once per month

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, it has invested in and deployed future technology solutions via its work building its own detection mechanisms and systems. Mistral AI further reports that its commitment to open source constitutes an important process in its continuous effort to improve its generative models and products through feedback from the community, open source tooling and OSINT. Additionally, Mistral AI reports that maintaining mitigations is a continuous process, and it constantly invests in improving its mitigations, including iterating on its in-house safety classifiers.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Mistral AI self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Mistral AI reports the following metrics (as measured since joining into the commitments):

- Cadence at which mitigations are assessed against the business's tech stack, to ensure effective performance: Quarterly

79 https://www.technologycoalition.org/newsroom/announcing-lantern

OpenAI

OPENAI REPORTS

According to OpenAI, it has invested in and deployed technology solutions via its efforts building in-house detection solutions for image, video and text-based child sexual exploitation and abuse. OpenAI further reports that it continuously works on improving its holistic safety stack, with a particular focus on child safety. According to OpenAI, it maintains the quality of its mitigations through human auditing, review, and regular assessment. OpenAI reports testing its detection technologies against relevant datasets to ensure necessary recall, monitoring performance and updating its classifiers when necessary to maintain high precision of its technologies or expand to detect new abuse vectors, such as sexual extortion.

OpenAI further reports that it leverages OSINT monitoring across platforms and the dark web to understand how its platforms, products and models are potentially being abused by bad actors. OpenAI reports maintaining a dedicated team to identify malicious activity. OpenAI additionally reports having joined The Tech Coalition's Lantern program[80] for signal sharing.

Sub-principle 3: Fight CSAM, AIG-CSAM and CSEM on our platforms.

We are committed to fighting CSAM online and preventing our platforms from being used to create, store, solicit or distribute this material. As new threat vectors emerge, we are committed to meeting this moment. We are committed to detecting and removing child safety violative content on our platforms. We are committed to disallowing and combating CSAM, AIG-CSAM and CSEM on our platforms, and combating fraudulent uses of generative AI to sexually harm children.

Anthropic

ANTHROPIC REPORTS

According to Anthropic, the strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are comprehensive across its first-party deployments. According to Anthropic, for manual reports where Anthropic is able to identify that the content is AIG-CSAM, it ensures that its CyberTipline reports supply the correct generative AI file annotation.

Civitai

CIVITAI REPORTS

According to Civitai, it employs a multi-layered approach to safeguarding its platform, utilizing the same core strategies outlined in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle: detection, user reporting, and prevention messaging.

In addition to using the in-house detection models discussed in the previously mentioned sub-principle, Civitai reports that when conducting ML/AI detection to scan uploaded images for indications of minors, sexually

80 https://www.technologycoalition.org/newsroom/announcing-lantern

explicit or mature content, it also leverages external tools such as Hive moderation. Civitai further reports it maintains an internal hash database of removed images to prevent the re-upload of previously flagged content, ensuring that identified violations do not resurface. Additionally, Civitai reports it detects uploads of images depicting known, real humans (in order to prevent sexual deepfakes of known individuals) checking input images against an unspecified database of "known individuals". According to Civitai, confirmed violations lead to model removal, content takedown, user bans, NCMEC reports, and hash-based reupload blocking.

Civitai further reports that it ensures reports of AIG-CSAM submitted to NCMEC's CyberTipline include all necessary parameters for accurate reporting and intervention, inclusive of information regarding the model used to generate the offending image, when that information is known. According to Civitai, it periodically reviews its reporting workflows, updating them if necessary to remain aligned with NCMEC standards and any new reporting formats it issues. Civitai reports that it also conducts internal audits to ensure proper annotation is applied in each relevant report.

Civitai further reports that, with respect to its low-rank adaptation (LoRA) training functionality,[81] it has implemented dataset cleaning for those third-party datasets that are uploaded by its users. According to Civitai, all uploaded datasets are moderated using a combination of metadata analysis, Hive tools, and human review to flag CSAM, CSEM, or AIG-CSAM risks, including any photorealistic depictions of minors. Civitai reports that as part of its review process, any datasets that contain any depictions of minors are further reviewed to assess whether they contain any adult sexual content. According to Civitai, confirmed violations result in account bans, content takedowns, model removal, hash-based blocking of re-uploads and reports to NCMEC.

NOT YET IMPLEMENTED

Civitai has not yet:

- Implemented industry-standard tools for hashing and matching against third-party owned, maintained and verified CSAM lists to detect known CSAM hosted on its platform
- Implemented prevention messaging as part of safeguarding the search functionality[82] on its site

Civitai reports that it is working to expand its moderation capabilities by integrating additional industry-standard tools. Civitai reports it is actively pursuing access to Microsoft's pDNA license, which would allow for integration with NCMEC's verified CSAM hashlist.

Civitai further reports it is exploring improvements to its search functionality to incorporate prevention messaging to ensure that certain flagged search terms trigger warnings or deterrent messaging. According to Civitai, blockers include defining accurate triggers, designing non-accusatory messaging, and the technical work needed to implement these interventions.

For more detail on progress, please see the discussion in previous principles.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of instances of CSAM detected on its site: 91

81 https://education.civitai.com/using-civitai-the-on-site-lora-trainer/
82 E.g. entering the terms "child abuse model" into its in-site search functionality does not surface prevention messaging

- The number of user reports submitted for various violations on its site: 710,256
- The number of instances of AIG-CSAM detected on its site: 242
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 91 and 242

Civitai further reports that as a result of these various violations, 21,842 accounts have been banned.

With respect to the efforts to clean third-party LoRA training datasets, Civitai reports the following metrics (as measured since joining the commitments):

- The percentage of third-party uploaded LoRA datasets that have been audited and updated: 100%
- The number of instances of CSAM detected in third-party uploaded LoRA datasets: 74
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 74 and 0

For more detail on impact metrics, please see the discussion in previous principles.

Google

GOOGLE REPORTS

According to Google, it invests heavily in fighting child sexual abuse and exploitation online and uses its proprietary technology to deter, detect, remove and report offences on its platforms. Google notes that its approach includes identifying and appropriately reporting obscene visual representations (OVR) of children, such as CSAM cartoons, and computer-generated imagery (CGI) CSAM.[83] Google notes that these types of abuse material are created using technology such as image editing and the AI generation of CSAM. Google states that in 2024, it reported hundreds of thousands of pieces of CGI CSAM and OVR CSAM that appeared on its platforms to NCMEC. According to Google, in 2024 it reported more than 5,000,000 pieces of content to NCMEC, with more than 1,000,000 Cybertipline reports. Google further notes that it brought enforcement action on over 600,000 accounts in 2024 for CSAM violations. Google states that its annual Transparency Report shares additional data regarding its global efforts and resources to combat CSAE. Google reports several examples of ways that it has invested in combating CSAE:

- **Acceptable use policies:** Google reports that its products that use generative AI or may process generative AI content also prohibit the use of its technology to abuse or exploit children via product-specific policies. Google highlights several examples:
  - Google's policy guidelines[84] for the Gemini app include threats to child safety, stating that Gemini should not "generate outputs, including Child Sexual Abuse Material, that exploit or sexualize children."
  - Google Play's Play Console Help Center[85] contains a section focused on understanding Google Play's AI-Generated Content policy[86] that outlines violative AI-generated content, such as AI-generated non-consensual deepfake sexual material, and prohibits content that may exploit or abuse children[87]

83 https://support.google.com/transparencyreport/answer/10330933?hl=en#zippy=%2Chow-does-google-combat-risks-of-csam-in-the-generative-ai-genai-space
84 https://gemini.google/policy-guidelines/?hl=en
85 https://support.google.com/googleplay/android-developer/?hl=en-GB#topic=3450769
86 https://support.google.com/googleplay/android-developer/answer/14094294?hl=en&ref_topic=12798386&sjid=15648833734686571924-NC
87 https://support.google.com/googleplay/android-developer/answer/9878809?sjid=15648833734686571924-NC

  - Google's Cloud Platform's Acceptable Use Policy[88] requires users to agree not "to engage in, promote or encourage illegal activity, including child sexual exploitation, child abuse, or terrorism or violence that can cause death, serious harm, or injury to individuals or groups of individuals."

- **Google Priority Flagger Program:** Google reports that as part of its Priority Flagger Program,[89] it partners with expert third parties that flag potentially violative content, including content that raises child safety issues, for its teams' review.

- **Regulatory reporting requirements in order to identify and assess risks:** Google reports that it is complying with regulatory requirements, such as the European Union's Digital Services Act (DSA) under which it conducts Systemic Risk Assessments (SRAs)[90] to help make the internet more safe, transparent, and accountable. Google notes that these reports identify and address risks, including CSAE on online platforms. Google further reports that the recently released 2024 SRA notes that generative AI may be used by bad actors to create new outputs that exacerbate some existing child safety risks and introduce new risks. Google outlines that mitigations for these risks include Google's product policies and testing its generative AI products before launch.[91]

- **Partnerships and technology building:** Google reports that it helped build the Hash Matching API tool, supporting NCMEC with filtering duplicate files.[92] Google further highlights its Child Safety Toolkit,[93] and ongoing proactive engagement with child safety experts from industry, academia, government, and civil society, including its annual Growing Up in the Digital Age summit.[94]

Invoke

INVOKE REPORTS

According to Invoke, the strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are comprehensive across its SaaS solution and OSS offerings. According to Invoke, when reporting AIG-CSAM to NCMEC its content moderation team ensures that its CyberTipline reports supply all of the correct parameters. Invoke further reports that it has observed a reduction in the number of fake accounts after introducing email verification.

NOT YET IMPLEMENTED

Invoke has not yet:

- Implemented detection for CSAM[95] for its SaaS solution where users may upload training data for model building and fine-tuning purposes

For more detail on progress, please see the discussion in previous principles.

88 https://cloud.google.com/terms/aup?hl=en
89 https://transparency.google/tools-programs/partner-programs/
90 https://transparencyreport.google.com/report-downloads?lu=report-155&hl=en%E2%80%9D%20with%20%E2%80%9Chttps://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf
91 https://support.google.com/transparencyreport/answer/10330933#zippy=%2Chow-does-google-contribute-to-the-child-safety-ecosystem-to-combat-csam%2Chow-does-google-combat-risks-of-csam-in-the-generative-ai-genai-space
92 https://safety.google/intl/en_us/stories/hash-matching-to-help-ncmec/
93 https://protectingchildren.google/tools-for-partners/
94 https://blog.google/technology/families/new-10m-funding-to-promote-teen-safety-and-wellbeing-in-europe/
95 E.g. via using hashing and matching against verified CSAM lists to detect known CSAM as part of input-level detection defenses of its SaaS solution

IMPACT METRICS

For more detail on impact metrics, please see the discussion in previous principles.

Meta

META REPORTS

According to Meta, it has policies that prohibit content or activity that exploits or endangers children, including rules against CSAM, child sexualization, child nudity, abuse and exploitation. Meta reports that these policies cover both synthetic (e.g. AI-generated) and non-synthetic media. According to Meta, it removes explicit sexualization of children when it becomes aware of the content, and has policies against implicit sexualization of children, even when the content itself appears to be benign. Meta reports that it removes accounts, profiles or pages dedicated to sharing images of children, (including AI-generated images) where captions and comments focus on children's appearance when Meta becomes aware of them.

According to Meta, it disables accounts and profiles for severe violations of its child safety policies, such as malicious distribution of CSAM or sexual solicitation of children, when it becomes aware of them. Meta further reports that it simultaneously disables other accounts held by the account holder, restricts the device from setting up future accounts, and disables linked Facebook or Instagram accounts.

Meta reports that it uses detection technology to proactively identify both known and unknown CSAM content, including AI-generated CSAM, and reports all known apparent instances to NCMEC in line with legal obligations. According to Meta, when it becomes aware of AIG-CSAM, it reports to NCMEC using the NCMEC-created annotation for generative AI when possible, and utilizes the NCMEC reporting template created by industry through the Tech Coalition in collaboration with NCMEC.

Meta reports that as of spring 2024, it has implemented C2PA and IPTC signal detection for generative AI-created or edited image content uploaded to its platforms. Meta further reports that it displays "AI info" labels for content detected as generated by an AI tool, and shares information on whether the content is labeled as such due to industry-shared signals or user self-disclosure. According to Meta, for content it detects as only modified or edited by AI tools, the "AI info" label has been moved to the three-dot button in the post's menu.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not build or offer access to platforms that allow for the solicitation or distribution of any material (regardless of the type of material that is solicited or distributed). In regards to preventing the creation and storing of this material, Metaphysic notes its efforts outlined in the discussion around the principle "Develop, build and train generative AI models that proactively address child safety risks."

NOT YET IMPLEMENTED

For more detail on progress, please see the discussion in previous principles.

IMPACT METRICS

For more detail on impact metrics, please see the discussion in previous principles.

Mistral AI

MISTRAL AI REPORTS

According to Mistral AI, the strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are comprehensive across its cloud-based and OSS offerings. According to Mistral AI, its content moderation team is prepared to ensure that reports of AIG-CSAM to NCMEC supply all of the correct parameters, though it has not detected any AIG-CSAM instances to date.

NOT YET IMPLEMENTED

Mistral AI has not yet:

- Implemented policies and processes to combat fraudulent uses[96] of agentic AI[97] to sexually harm children[98]
- Implemented policies and processes to combat fraudulent use of code generation capabilities[99] to create models specifically built to produce AIG-CSAM, or services that are used to "nudify" content of children
- Implemented interventions [100]for its services where users may upload training data for model fine-tuning purposes,[101] to prevent users from fine-tuning its models to learn AIG-CSAM and CSEM capabilities

For more detail on progress, please see the discussion in previous principles.

IMPACT METRICS

For more detail on impact metrics, please see the discussion in previous principles.

OpenAI

OPENAI REPORTS

According to OpenAI, the strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are comprehensive across its web interface, direct-to-consumer apps, API Platform and Enterprise offerings.

OpenAI additionally reports it is a member of The Tech Coalition,[102] an industry-member organization working to defend against the sexual exploitation and abuse of children online via knowledge, information sharing, collective action, and the innovation and adoption of new technologies. OpenAI further reports it has joined the Beneficial AI for Children Coalition,[103] a multi-stakeholder initiative led by the Paris Peace Forum and Everyone. AI to support practical, evidence-based guidelines to safeguard and promote children's cognitive and socio-emotional well-being in the age of AI.

96 Early indications of agentic misuse in other non-child safety domains include relationship-centric influence operations. E.g. Anthropic. *Operating Multi-Client Influence Networks Across Platforms*. April 2025, https://cdn.sanity.io/files/4zrzovbb/website/45bc6adf039848841ed-9e47051fb1209d6bb2b26.pdf.
97 https://mistral.ai/news/agents-api
98 E.g. attempts to sexually exploit minors that make use of rapid relationship building or content creation
99 https://mistral.ai/news/codestral
100 E.g. hashing and matching against verified CSAM lists to prevent users from uploading CSAM datasets
101 E.g. Mistral AI's fine-tuning offering https://docs.mistral.ai/capabilities/finetuning/classifier_factory/
102 https://www.technologycoalition.org/
103 https://parispeaceforum.org/initiatives/beneficial-ai-for-children-coalition/

Definitions

AI-generated child sexual abuse material (AIG-CSAM)	Visual depiction (image/video) of sexually explicit conduct involving a minor, the creation of which has been facilitated by generative AI technologies. This may range from a fully generated image/video to generated elements applied to a pre-existing image/video.
Child sexual abuse material (CSAM)	Visual depiction (image/video) of sexually explicit conduct involving a minor. Does not require that the material depict a child engaging in sexual activity. Covers lewd and lascivious content, as well as content with a focus on genitalia. N.B. The definition of minor will vary depending on your legal jurisdiction.
Child sexual exploitation material (CSEM)	Used as a shorthand for the full list of: image/video/audio content sexualizing children, grooming text, sexual extortion text, CSAM advertising, CSAM solicitation, and text promoting sexual interest in children.
CSAM advertising	Noting where child sexual abuse material can be found. This may be a URL or advertisement of CSAM for sale.
CSAM solicitation	The act of requesting, seeking out, or asking for access to, or the location of, child sexual abuse material.
Detect	The method or act of scanning through a larger set of data to attempt to identify the target material (e.g. CSAM or CSEM). Can include both manual and automated methodologies.