Progress on our Child Safety Commitments

May 2025

Introduction

Anthropic is committed to building AI systems that are safe, beneficial, and respectful of human values. Central to this commitment is safeguarding against the unique risks related to generative AI and child sexual abuse and exploitation material (CSAEM). Our Usage Policy explicitly prohibits the use of our models to further sexual harms against children.

To put safety first, we have deliberately prioritized design and product choices for Claude that prevent potential misuse and strengthen our safety infrastructure. Currently, Claude models generate text-based interactions and do not generate images or videos.

In April 2024, we became a signatory to Thorn's Safety by Design for Generative AI: Preventing Child Sexual Abuse principles. This complements our other collaborations in child safety, including membership with the Tech Coalition. At the one-year anniversary of this commitment, we want to share an update on our progress in implementing these principles across its three key phases: Develop, Deploy, and Maintain.

Detection and Prevention Systems

Claude currently does not produce image or video outputs and is therefore incapable of generating image-based child sexual abuse material (CSAM) or non-consensual intimate images (NCII).

We take a multi-pronged approach to detecting and preventing abusive content. For example, we employ the following techniques:

- On our first-party services, we employ hash-matching technology to detect and report known CSAM to the National Center for Missing and Exploited Children (NCMEC) that users may upload. In the period between April 15, 2024 and March 31, 2025, we detected and reported 859 images to NCMEC. We are implementing similar tooling for detecting NCII and novel CSAM, pending our exploration of adequate technological solutions. Our third-party partners maintain their own screening and detection systems.

- We run safety classifiers on prompts and completions to identify harm. If user inputs violate our Usage Policy, we may take action such as automatically modifying the request, issuing warnings, or in serious cases, conducting user suspensions or bans. We review our usage policies, classification and detection systems, and enforcement processes on a regular cadence.
- We undertake various data preparation and cleaning processes to ensure that training data is of sufficient quality and appropriateness. We are in the process of adopting interventions to avoid ingestion of CSAM, CSEM, and NCII from our training datasets. For CSAM, we have been actively working to set up hash-matching and reporting mechanisms with third parties to prevent CSAM ingestion in our training data. For images that do come from crawl sources, we have begun to leverage a general Not Safe For Work (NSFW) filter over the training data.
- We also work with a third-party vendor to perform Open-Source Intelligence (OSINT) and send our internal team alerts related to general platform abuse.

Model Testing

We evaluate all deployed models. Red teaming for CSAM and CSEM is led by internal child safety experts, and employees involved in this process are trained on responsible red teaming procedures as well as general reporting and preservation processes for CSAM. We also integrate policy testing that we commission from outside subject matter experts to ensure that our evaluations are robust and take into account new trends in abuse. Results from red teaming are provided to our model finetuning and safeguards teams to assess for integration back into model training, model development, and deployment of safety and enforcement strategies. For example, we used feedback from child safety experts at Thorn around signals often seen in child grooming to update our classifiers, enhance our Usage Policy, fine-tune our models, and incorporate these signals into testing of future models.

Additionally, we use a phased deployment approach that ensures thorough testing and limits access before wider release . Our deployment phases typically include: (1) Internal testing with employees only; (2) Limited early access with select customers; and (3) Graduated general availability.

Monitoring and Reporting

Anthropic's Usage Policy explicitly prohibits the use of our models to facilitate sexual harms against children. For all Claude model families, we have implemented layered enforcement mechanisms, which include warnings, prompt modification, and account restrictions. In severe cases and/or repeated abuse we may ban or terminate accounts. For our models deployed on third-party platforms, we coordinate with platform partners to monitor violations and take appropriate enforcement actions. Claude is also trained to provide prevention messaging when potential CSAM solicitation is detected.

We maintain multiple channels for identifying and reporting violative content. Our in-house Safeguards experts monitor public forums and analyze emerging abuse patterns. We have also established reporting flows that allow users to flag concerning content or model behavior.

In addition, we will continue to update the metrics on our Platform Detection page on a regular basis and include information on child safety testing in our model documentation.

Conclusion

As we continue to advance our AI capabilities, we are committed to maintaining the highest standards of child protection and will continue to invest in cutting-edge detection and prevention technologies; collaborate with industry partners, civil society, and law enforcement; share our learnings to foster a safer online environment for all; and adapt our approaches as new challenges emerge.

Together with our partners in industry, government, and civil society, we will work to ensure that generative AI technologies are developed and deployed in ways that protect and benefit all members of society, including children.