Sora 2 System Card

OpenAI

September 30, 2025

Contents

1 Overview of Sora 2

Sora 2 is our new state of the art video and audio generation model. Building on the foundation of Sora, this new model introduces capabilities that have been difficult for prior video models to achieve– such as more accurate physics, sharper realism, synchronized audio, enhanced steerability, and an expanded stylistic range. The model follows user direction with high fidelity, enabling the creation of videos that are both imaginative and grounded in real-world dynamics. Sora 2 expands the toolkit for storytelling and creative expression, while also serving as a step toward models that can more accurately simulate the complexity of the physical world. Sora 2 will be available via sora.com, in a new standalone iOS Sora app, and in the future it will be available via our API.

Sora 2's advanced capabilities require consideration of new potential risks, including nonconsensual use of likeness or misleading generations. To address these, we worked with internal red teamers to identify new challenges and inform corresponding mitigations. We're taking an iterative approach to safety, focusing on areas where context is especially important or where risks are still emerging and are not fully understood.

Our iterative deployment includes rolling out initial access to Sora 2 via limited invitations, restricting the use of image uploads that feature a photorealistic person and all video uploads, and placing stringent safeguards and moderation thresholds on content involving minors. We'll continue to learn from how people use Sora 2 and refine the system to balance safety while maximizing creative potential. This system card describes the model's capabilities, potential risks, and the safety measures OpenAI has developed for a safe deployment of Sora 2.

2 Model Data & Data Filtering

Like OpenAI's other models, Sora 2 was trained on diverse datasets, including information that is publicly available on the internet, information that we partner with third parties to access, and information that our users or human trainers and researchers provide or generate. Our data processing pipeline includes rigorous filtering to maintain data quality and mitigate potential risks. We also employ a combination of safety classifiers to help prevent the use or generation of harmful or sensitive content, including explicit materials such as sexual content involving a minor.

3 Safety: Safety Stack, Evaluations and Mitigations for Sora 2

3.1 Safety Stack

To safely deploy Sora 2, we have built a robust safety stack that builds on our learnings from Sora 1, incorporates mitigations developed for other OpenAI models and products such as GPT-4o Image Generation and DALL·E, and adds safeguards specific to Sora 2. These include:

- **Text and image moderation via multi-modal moderation classifiers:** Input prompts, output video frames, audio transcripts, comments, and output scene description texts are run through various safety models:

  - **Input (prompt) blocking**: This strategy involves blocking the tool from generating a video if text or image classifiers flag the prompt as violating our policies. By preemptively identifying and blocking inputs, this measure helps prevent the generation of disallowed content before it even occurs.
  - **Output blocking**: This approach, applied after the video has been generated, uses a combination of controls including Child Sexual Abuse Material (CSAM) classifiers and a safety-focused reasoning monitor to block the output of videos that violate our policies in the event that our input blocks have been circumvented. The monitor is a multimodal reasoning model which is custom-trained to reason about content policies. By evaluating the output post-generation, this strategy aims to block any content that is disallowed under our policies, providing an additional safeguard against the creation of disallowed content.

- **Increased safeguards for minors**: We apply the mitigations above more strictly for users who may be under 18, while limiting their ability to create certain categories age-inappropriate content. Users under the age of 13 are currently prohibited from using any of OpenAI's products or services.

3.2 Product and Usage Policies

In addition to the protections we have built into the model and system to prevent the generation of violative content, we take further steps to reduce misuse across the new Sora app's product surfaces (videos, comments, profiles, and messaging).

We clearly communicate policy guidelines through in-product and publicly available Usage Policies[1], which prohibit:

- Violations of others' privacy, including use of another person's likeness without their permission
- Use of Sora to infringe on the safety and security of others, including content that threatens, harrasses, or defames others, non-consensual intimate imagery, or content intended to incite violence or the suffering of others
- Misleading others through impersonation, scams, or fraud; and
- Use of Sora in ways that exploit, endanger, or sexualize minors

As with Sora 1, some of these forms of misuse are addressed through our model and system mitigations, but others are more contextual, and require additional information to be appropriately evaluated.To support enforcement, we provide in-app reporting, combine automation with human review to detect patterns of misuse, and apply penalties or remove content when violations occur. Users are notified of enforcement actions and given an opportunity to respond. We track the effectiveness of these measures and refine them over time.

We also take steps to filter out violating content, as well as content that is inappropriate for younger audiences, from Sora's social feed.

[1]OpenAI Usage Policies: https://openai.com/policies/usage-policies/

3.3 Provenance and Transparency Initiatives

Because many of the potential problem spaces for Sora 2, such as unauthorized use of likeness, are highly context dependent, we've continued to invest in and refine our provenance tools.

For general availability, our provenance safety tooling for our First-party (1P) products will include:

- **C2PA metadata** on all assets, providing verifiable origin through an industry standard
- **Visible moving watermark** on videos downloaded from sora.com or the Sora app
- **Internal detection tools** to help assess whether a certain video or audio was created by our products.

We recognize that there is not a single solution to provenance, but we will continue to improve the provenance ecosystem to help bring more transparency to content created from our tools.

3.4 Specific Risk Areas & Mitigations

In addition to the early testing, red-teaming and safety evaluations highlighted several areas of focus.

- **Harmful Or Inappropriate Outputs:** Similar to Sora 1, Sora 2 absent mitigations may carry the risk of producing harmful or inappropriate content, including violence, self harm, terrorist material, or sexual content. To address these risks, Sora 2 uses automated detection systems that scan video frames, scene descriptions, and audio transcripts aimed to block content that violates our guidelines. We also have a proactive detection system, user reporting pathways to flag inappropriate content, and apply stricter thresholds to material surfaced in Sora 2's social feed. We are continuously monitoring trends to adapt mitigations as new risks arise. Please see the Safety Evaluations section for our evaluations across specific types of content.
- **Misuse of Likeness & Deceptive Content:** Sora 2's ability to generate hyperrealistic video and audio raises important concerns around likeness, misuse, and deception. As noted above, we are taking a thoughtful and iterative approach in deployment to minimize these potential risks. Safeguards for our initial launch include: not supporting video-to-video generation at launch, not supporting text-to-video generation of public figures, and blocking generations that include real people (other than users who consent through Sora's likeness-control cameo feature); and mechanisms requiring explicit opt-in consent and controls for the use of likeness through cameos. Where real people are featured in videos, additional model safeguards will apply. These include classifiers intended to prevent non-consensual nudity or racy output, graphic violence, or output that could be used for certain fraudulent purposes. And although some deceptive content is highly contextual and not easily detectable by classifiers, Sora's provenance measures are aimed at further mitigating those risks.
- **Child Safety:** OpenAI is committed to addressing child safety risks across all of our products, including Sora. We prioritize the prevention, detection, and reporting of Child Sexual Abuse Material (CSAM) by responsibly sourcing datasets to exclude CSAM, partnering with the National Center for Missing & Exploited Children (NCMEC), and applying robust

scanning across all inputs and outputs—including first-party and third-party use (API and Enterprise) unless customers meet strict criteria for removal. To prevent CSAM generation, we have built a dedicated safety stack that leverages system mitigations used across our other products as well as additional safeguards developed specifically for Sora. For full details on our specific CSAM safety stack can be found on our Sora System Card[2].

- **Teen Safety:** Sora 2 has a number of additional safeguards designed to protect users under 18. These include:
  - **Model output restrictions for minor users.** We are applying additional moderation thresholds for users we believe may be under the age of 18. Separately, when our classifiers detect the presence of a potential minor in an uploaded image or video (including through Sora's cameo feature) subsequent generations based on that image or video are subjected to even tighter safety thresholds to prevent additional categories of potentially harmful generations. Regardless of the user's age, classifiers seek to ensure that Sora's public feed will only include content that is in line with our under 18 policies and intended to be appropriate for teen users.
  - **Privacy & Parental Controls.** When deploying Sora 2 on sora.com and the new Sora app, we will also have stricter privacy safeguards and defaults for teens, including limits on how their likeness can be used and protections against unwanted contact or discovery by adults. In addition, we are introducing a suite of parental controls, which you can read more about in our blog[3].

As mentioned above, we are taking an iterative approach to this deployment, and specific mitigations may change in the future.

4 Red Teaming

OpenAI worked with external testers from OpenAI's Red Team Network to test Sora 2, evaluate existing safety mitigations, and provide feedback on emerging risks. Red teamers assessed new safety measures, and suggested improvements for future iterations.

Content generation red teaming focused on violative and disallowed categories under OpenAI's usage policies—including sexual content, nudity, extremism, self-harm, wrongdoing, violence and gore, and political persuasion—as well as additional policies on youth safety and likeness use. Red teamers also probed violative uploads, tested media generation, attempted to jailbreak safety systems, and stress-tested product-level safeguards. Insights from red teaming informed the design of new safety measures and refinements to existing ones, including further tuning of prompt filters, blocklists, and classifier thresholds to better align the model with our safety objectives.

5 Safety Evaluations

OpenAI evaluated Sora 2's safety stack using thousands of adversarial prompts gathered through targeted red-teaming. Each prompt was categorized by use case and policy area, run through a

[2]OpenAI Sora System Card: https://openai.com/index/sora-system-card/
[3]Shape how ChatGPT works for your family: http://chatgpt.com/parent-resources

helpful-only version of the video model to generate outputs, and then graded and converted to an automated evaluation. The production safety stack—scanning video frames, captions, and audio transcripts—was tested for two key metrics: **not_unsafe**, measuring how effectively unsafe content is blocked (recall), and **not_overrefuse**, measuring how well benign content avoids false blocks. See below a summary of our results aggregated across use cases and policies.

Table 1: Safety Evaluations

Category	**not_unsafe at output**	**not_overrefuse at output**
Adult Nudity / Sexual Content Without Use of Likeness	96.04%	96.20%
Adult Nudity / Sexual Content With Use of Likeness	98.40%	97.60%
Self-Harm	99.70%	94.60%
Violence and Gore	95.10%	97.00%
Violative Political Persuasion	95.52%	98.67%
Extremism/Hate	96.82%	99.11%

6 Continued work on safety, policy, & iterative deployment

We are committed to building Sora 2 as a system where users can create safely and confidently. Safety and creativity go hand-in-hand: people are most expressive when they can trust the product they use. While layered safeguards are in place, some harmful behaviors or policy violations may still circumvent mitigations.

To strengthen protections, we are investing in features such as age prediction and further provenance measures. Our safety stack will continue to evolve through ongoing fine tuning and feature refinement as Sora 2 usage develops. Internal teams will monitor trends, assess the effectiveness of current mitigations, and adapt policies or enforcement to address emerging risks.

7 Acknowledgements

Thank you to all of OpenAI's internal teams, including Communications, Brand Design, Global Affairs, Integrity, Intel & Investigations, Legal, Product Policy, Safety Systems and User Operations, whose support was instrumental in helping develop and implement Sora's safety mitigations as well as their contributions to this System Card.