OpenAI

ChatGPT Images 2.6

System Card

2026-11-04

1 Introduction

ChatGPT Images 2.6 introduces new ways to create and refine visual content. The updated Sunburst and Flare models improve editing consistency, detail preservation, and responses to sketches and image annotations. Flare is optimized for faster generation, while configurable reasoning effort gives users in Chat via the model picker or developers in the API more control over generation and editing speed. This release also adds GIF generation and photo markup in both ChatGPT and the API, plus SVG output support in the API.

The core safety stack we are using with ChatGPT Images 2.6 is based on the same safety foundations of ChatGPT Images 2.5, with additional safeguards for GIF creation. This system card describes our safety evaluations, mitigations, and limitations, including risks introduced by the new GIF feature.

2 Model Data and Training

Like OpenAI's other models, GPT-Image-2.6 Sunburst and GPT-Image-2.6 Flare were 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 use advanced data filtering processes to reduce personal information from training data. We also employ safety classifiers to help prevent or reduce the use of harmful or sensitive content, including explicit materials such as sexual content involving a minor.

3 Model Safety

3.1 Safety Stack

GPT-Image-2.6 Sunburst includes multiple layers of image-specific safety protection. Upstream, LLM-based policy checks are designed to block potentially violative image requests before image generation begins. Downstream, we use safety-focused image classifiers designed to block potentially violative input images from being fed into the final generation, combined analysis of input images and prompts to evaluate potential malicious edits,

and a final step where we analyze whether the generated image violates our policies before we show the image to the user.

Below we explain each of these safety layers in a bit more detail:

- **Upstream Refusals:** Before a request is sent to the image generation model, we run LLM-based policy checks to evaluate whether the request violates policy. Requests deemed to violate policy are refused at this stage.
- **Downstream Blocking:** After a request reaches the image generation system, we use a safety reasoning model as a monitor to moderate both image inputs to the image generation model and the generated output. This model is a safety-focused multimodal model trained to reason about prompt intent and content policies.
  - **Input blocking:** The monitor checks all text and image input provided to the image generation tool. If the monitor determines that any input violates policy, generation is blocked.
  - **Output blocking**: The monitor also checks the final output image before it is shown to the user. If the monitor determines that the output image violates policy, it is blocked.

Safety-focused image and text classifiers have been part of our image safety stack since the original GPT-4o image generation launch. Improvements made across previous releases include:

- **Improving the Safety Reasoning Model**: We've continuously improved the safety classifier, safety policies, and overall safety stack with new training to address the feedback we've collected and enhance our offline and online monitoring enforcement.
- **Safety evaluation:** We evaluate both the images generated by the model and whether moderation blocks outputs graded unsafe. Section 3.2.1 reports these outcomes separately for each model and its corresponding 2.5 baseline.
- **Additional child-safety checks:** For ChatGPT image generation, we added synchronous CSAM classification alongside existing image moderation.
- **Expanding online and offline monitoring:** We maintain broader offline and online monitoring and enforcement. We also run multiple evaluation stacks for higher-risk categories around minors.

3.2 Safety and Policy Evaluations

3.2.1 Safety Evaluations

We used an automated evaluation to measure the efficacy of selected production moderation checks for the GPT-Image-2.6 Sunburst and GPT-Image-2.6 Flare models. We evaluated saved generations from a shared set of 3,077 challenging requests designed to generate images that would violate our policy across safety categories such as violence and sexual content. Each model was compared with its corresponding GPT-Image-2.5 baseline using a common reference-grading procedure. These results are not representative of how often violative requests or outputs arise in production traffic.

Definitions:

- **Total Prompts:** Total prompts meant to generate violative images as part of our automated evaluation process.
- **Model Outputs**
  - Safe Generation: Images where the adversarial prompt resulted in a non-violative image generation, whether or not the evaluated moderation checks would allow the image.
  - Unsafe Generation Blocked: Images where the adversarial prompt resulted in a violative image generation that was blocked by the evaluated moderation checks.
  - Unsafe Generation Not Blocked: Outputs graded violative that were not blocked by the evaluated moderation checks. This is an offline evaluation outcome, not observed delivery to users.

Each model and generation setting was evaluated on the same 3,077 requests. Results are shown as percentages.

Table 1: Safety Evaluations Overall

		Model Output: Final Unsafe Outcome		
Total Prompts	**Model**	**Safe Generation (higher is better)**	**Unsafe Generation Blocked (higher is better)**	**Unsafe Generation Not Blocked (lower is better)**
	GPT-Image-2.5 Sunburst (baseline)	76.15%	21.94%	1.92%
	GPT-Image-2.6 Sunburst	75.50%	22.78%	1.72%
3,077	GPT-Image-2.5 Flare (baseline)	78.36%	20.15%	1.49%
	GPT-Image-2.6 Flare	**75.04%***	**22.81%***	2.14%

* Asterisks indicate statistically significant differences from the corresponding GPT-Image-2.5 baseline after adjustment for multiple comparisons.

Table 2: Safety Evaluations by Policy

		Model Output: Final Unsafe Outcome		
Prompts by Policy	**Model**	**Safe Generation (higher is better)**	**Unsafe Generation Blocked (higher is better)**	**Unsafe Generation Not Blocked (lower is better)**
	GPT-Image-2.5 Sunburst (baseline)	76.15%	21.94%	1.92%
	GPT-Image-2.6 Sunburst	75.50%	22.78%	1.72%
Overall[1]	GPT-Image-2.5 Flare (baseline)	78.36%	20.15%	1.49%
	GPT-Image-2.6 Flare	**75.04%***	**22.81%***	2.14%
	GPT-Image-2.5 Sunburst (baseline)	18.52%	80.42%	1.06%
	GPT-Image-2.6 Sunburst	15.87%	83.07%	1.06%
Sexual	GPT-Image-2.5 Flare (baseline)	24.34%	75.13%	0.53%
	GPT-Image-2.6 Flare	**12.70%***	**87.30%***	0.00%
	GPT-Image-2.5 Sunburst (baseline)	88.61%	10.85%	0.54%
	GPT-Image-2.6 Sunburst	88.07%	11.21%	0.72%
Hate	GPT-Image-2.5 Flare (baseline)	91.14%	8.32%	0.54%
	GPT-Image-2.6 Flare	88.97%	10.31%	0.72%

		Model Output: Final Unsafe Outcome		
Prompts by Policy	**Model**	**Safe Generation (higher is better)**	**Unsafe Generation Blocked (higher is better)**	**Unsafe Generation Not Blocked (lower is better)**
Other[2](e.g., religious and public figures)	GPT-Image-2.5 Sunburst (baseline)	79.98%	17.77%	2.25%
	GPT-Image-2.6 Sunburst	79.30%	19.12%	1.57%
	GPT-Image-2.5 Flare (baseline)	82.34%	16.42%	1.24%
	GPT-Image-2.6 Flare	**76.49%***	20.70%	2.81%
Violence / Gore	GPT-Image-2.5 Sunburst (baseline)	70.66%	28.27%	1.07%
	GPT-Image-2.6 Sunburst	70.02%	29.12%	0.86%
	GPT-Image-2.5 Flare (baseline)	73.45%	25.70%	0.86%
	GPT-Image-2.6 Flare	71.52%	27.41%	1.07%
Extremism	GPT-Image-2.5 Sunburst (baseline)	82.05%	15.38%	2.56%
	GPT-Image-2.6 Sunburst	79.49%	20.51%	0.00%
	GPT-Image-2.5 Flare (baseline)	74.36%	20.51%	5.13%
	GPT-Image-2.6 Flare	79.49%	20.51%	0.00%
Self-harm	GPT-Image-2.5 Sunburst (baseline)	60.38%	34.34%	5.28%
	GPT-Image-2.6 Sunburst	61.51%	34.34%	4.15%
	GPT-Image-2.5 Flare (baseline)	61.51%	33.58%	4.91%
	GPT-Image-2.6 Flare	61.13%	32.83%	6.04%
Political	GPT-Image-2.5 Sunburst (baseline)	90.75%	8.19%	1.07%
	GPT-Image-2.6 Sunburst	89.32%	8.54%	2.14%
	GPT-Image-2.5 Flare (baseline)	92.17%	6.76%	1.07%
	GPT-Image-2.6 Flare	90.75%	7.83%	1.42%
Abuse	GPT-Image-2.5 Sunburst (baseline)	68.10%	31.03%	0.86%
	GPT-Image-2.6 Sunburst	66.38%	31.90%	1.72%
	GPT-Image-2.5 Flare (baseline)	70.69%	28.45%	0.86%
	GPT-Image-2.6 Flare	72.41%	26.72%	0.86%

		Model Output: Final Unsafe Outcome		
Prompts by Policy	Model	Safe Generation (higher is better)	Unsafe Generation Blocked (higher is better)	Unsafe Generation Not Blocked (lower is better)
	GPT-Image-2.5 Sunburst (baseline)	89.71%	6.17%	4.12%
	GPT-Image-2.6 Sunburst	89.30%	6.58%	4.12%
Wrongdoing	GPT-Image-2.5 Flare (baseline)	90.53%	6.17%	3.29%
	GPT-Image-2.6 Flare	88.07%	7.41%	4.53%

For the evaluated configurations shown, Sunburst produced fewer violative images post-moderation than its corresponding GPT-Images-2.5 baseline. Flare's rate increased from 1.49% (46 of 3,077 requests) to 2.14% (66 of 3,077), an increase of 0.65 percentage points. Neither overall difference was statistically significant.

Before moderation, the proportion of Flare outputs graded non-violative decreased from 78.36% to 75.04%. This decline was statistically significant, with the largest contributions coming from the "Other" and sexual prompt categories. Of the 20 net additional unsafe outputs that were not blocked, 14 were in the "Other" category. No individual prompt category showed a statistically significant increase in violative outputs when safety moderation was applied.

These evaluations use a fixed set of adversarial requests that are designed around challenging edge cases with automated grading. Results apply to the model and safeguard configurations evaluated and do not measure the complete deployed safety system or the frequency of violative outputs in production.

3.2.2 GIF Risks, Safeguards, and Evaluations

GIF generation introduces risks that may not be apparent in individual still images, including sexualized movement and harmful meaning conveyed across frames. To address these risks, we apply safeguards designed to detect and block policy-violating GIF content. GIF-specific rules also block some content categories more broadly than in still images to account for inferred motion.

[1] We find that the GPT-Images-2.6 series with moderation checks generally performs on par with GPT Images-2.5. Minor numerical regressions are not statistically significant.

[2] The other category includes long-tail categories that should be protected against including religious figures, religious satire, fabricated or compromising depictions of real people, jailbreak attempts, and image-based requests.

We evaluated these safeguards using a targeted set of adversarial GIFs labeled as policy-violating. These challenging cases are not representative of typical user requests or outputs. Recall measures the proportion of policy-violating GIFs in this evaluation set that the evaluated safeguards detected. It does not estimate how often harmful requests are fulfilled or harmful GIFs reach users in production.

Motion-related violations remain challenging to detect, and the safeguards may not catch all violations.

Policy	GIF Moderation Recall
Overall	77.7%
Sexual	77.8%
Hate / abuse	90.9%
Violence / gore	89.7%
Extremism	92.9%
Self-harm	77.5%
Violent Activities	96.6%

3.3 Preparedness Framework: Image-Specific Capability Assessment and Safeguards

OpenAI's Preparedness Framework is designed to track and prepare for models with frontier capabilities that could create new risks of severe harm in three tracked categories: Biological and Chemical, Cybersecurity, and AI Self-Improvement. Because image models are unable to create and execute code in a way that would enable AI Self-Improvement, we do not have evidence that GPT-Image-2.6 Sunburst nor GPT-Image-2.6 Flare pose meaningful risk in this category.

For the Cyber and Biological categories, we adapted existing language-model capability evaluations to assess image-generation capabilities. For each task, we instructed the model to generate an image containing both a visible reasoning scratchpad and a final answer. We then graded only the final answer rendered in the image, using the evaluation's existing grading criteria. This allowed us to test the model on the same underlying cyber-capability and biological-capability questions while eliciting and evaluating its response through its native image modality.

The results indicate that neither GPT-Image-2.6 Sunburst nor GPT-Image-2.6 Flare cross the Bio High threshold or the Cyber High threshold.

Although neither model crosses these thresholds, we follow the precautionary approach used for images 2.0 and 2.5 and apply mitigations appropriate

for a model treated as High biological capability. These mitigations include an image-specific adaptation of our existing biological-risk safety policy, which we apply to all ChatGPT Images 2.6 inputs and outputs using our safety reasoning model. The general input and output blocking mechanisms are described in Section 3.1.

3.3.1 Biological-risk monitoring and enforcement

As with GPT-Images 2.0 and 2.5, we are enabling the same policy enforcement checks for biorisk that we do with our text-based models, largely outlined in this article. We use advanced reasoning models with high biological capabilities to detect biological misuse, combining our automated systems with human reviewers to monitor and enforce our policies. We have integrated the analysis from our image-based safety reasoning model into the signals we use to detect ongoing misuse. In some cases where we detect ongoing patterns of misuse we suspend the offending accounts.

4 Image Provenance

We have continued to prioritize enhancing our provenance tools. For ChatGPT Images 2.6, our expanded provenance safety tooling includes:

- A continued commitment to C2PA metadata, an industry-standard framework that enables automated disclosure of provenance information, through the C2PA Conformance Program.
- To make provenance more resilient, we are taking a multi-layered approach and incorporating watermarking through Google DeepMind's SynthID through ChatGPT, Codex, and the OpenAI API. SynthID embeds an invisible watermarking layer that complements C2PA metadata-based approaches.

We recognize that there is no single solution to provenance, but are committed to improving the provenance ecosystem, continuing to collaborate on this issue across industry and with civil society, and helping build context and transparency to content created from ChatGPT Images 2.6 and across our products.

Readers can use OpenAI's public verification tool to check supported files for provenance signals.