OpenAI

ChatGPT Images 2.5

System Card

2026-09-08

1 Introduction

ChatGPT Images 2.5 is a major step forward in image generation capabilities. The model produces more consistent results when editing images, and improves infographic accuracy and layout. Users can change an image's setting, style, or composition while retaining more of the details. New features such as sketch-based generation, templates, and shared prompts give users more control over the generation process, and faster generation lets users create and iterate in less time. The core safety stack we are using with ChatGPT Images 2.5 is based on the same foundations as our ChatGPT Images 2.0 safety stack, with added safeguards for addressing new risks that emerge as models become more capable.

2 Model Data and Training

Like OpenAI's other models, GPT-Image-2.5-Sunburst and GPT-Image-2.5-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 Observed Safety Challenges, Evaluations, and Mitigations

3.1 New Safety Challenges

Compared to our past ChatGPT Images 2.0 deployment, ChatGPT Images 2.5 allows for heightened realism that could, absent safeguards, allow more convincing deepfakes, including political, sexual, or otherwise sensitive imagery of real people, places or events.

Such images violate our usage policies, and we have protections in place to help prevent them from reaching users, including safeguards at the prompt (text) and image layers.

3.2 Safety Stack

GPT-Image-2.5-Sunburst includes multiple layers of image-specific safety protection. We have special-purpose safety prompts designed to block potentially violative image requests before the image generation process begins, 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, and we have continuously improved the system since then. Most notably, recent improvements 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.
- **Making evaluation more product-grounded**: We've evaluated how harms manifest in production, and shifted from a raw taxonomy-matching approach to a more outcome-based evaluation of real harmful-output risk.

- **Expanding online and offline monitoring:** We now maintain broader offline and online monitoring and enforcement. We also run multiple evaluation stacks for higher-risk categories around minors.

3.3 Safety and Policy Evaluations

3.3.1 Safety Evaluations

We used an automated evaluation to measure the efficacy of our safety stack for the GPT-Image-2.5-Sunburst model and the GPT-Image-2.5-Flare model. We tested our end-to-end safety system using challenging prompts we specifically designed to generate images that would violate our policy broadly across all safety categories we track (e.g. violence, sexual content), so the numbers below are not representative of how often such prompts arise in production traffic.

Definitions:

- **Total Prompts:** Total prompts meant to generate violative images as part of our automated evaluation process.
- **Model Outputs**
  - **Safe Generated:** Images where the adversarial prompt resulted in a safe image generation.
  - **Unsafe Generation Blocked:** Images where the adversarial prompt was blocked by our moderation stack and resulted in no image shown.
  - **Unsafe Generation Presented:** Remaining undetected images. These are violative generations that would not have been blocked by the safety stack.

Table 1: Safety Evaluations Overall

Total Prompts	Model	Model Output: Final Unsafe Outcome		
		Safe Generation (higher is better)	**Unsafe Generation Blocked** (higher is better)	**Unsafe Generation Presented** (lower is better)
Total Prompts	GPT-Image-2.5-Sunburst	77.0%*	21.9%*	1.09%
	GPT-Image-2.5-Flare	79.4%*	19.2%*	1.41%
	ChatGPT Images 2.0 (baseline)	75.2%	23.1%	1.64%

Table 2: Safety Evaluations by Policy

Category	Model	Model Output: Final Unsafe Outcome		
		Safe Generation (higher is better)	**Unsafe Generation Blocked** (higher is better)	**Unsafe Generation Presented** (lower is better)
Sexual	GPT-Image-2.5-Sunburst	10.4%	89.1%	0.52%
	GPT-Image-2.5-Flare	14.5%	84.5%	1.04%
	ChatGPT Images 2.0 (baseline)	10.4%	87.6%	2.07%
Hate	GPT-Image-2.5-Sunburst	87.7%*	11.9%*	0.36%
	GPT-Image-2.5-Flare	89.5%*	10.1%*	0.36%
	ChatGPT Images 2.0 (baseline)	82.7%	16.9%	0.36%
Violence / Gore	GPT-Image-2.5-Sunburst	83.1%	16.3%	0.64%
	GPT-Image-2.5-Flare	85.9%*	12.2%*	1.93%
	ChatGPT Images 2.0 (baseline)	81.6%	16.5%	1.93%
Extremism	GPT-Image-2.5-Sunburst	84.6%	15.4%	0.00%
	GPT-Image-2.5-Flare	84.6%	15.4%	0.00%
	ChatGPT Images 2.0 (baseline)	76.9%	20.5%	2.56%
Self-harm	GPT-Image-2.5-Sunburst	76.6%	22.3%	1.13%
	GPT-Image-2.5-Flare	78.9%	20.0%	1.13%
	ChatGPT Images 2.0 (baseline)	76.2%	22.3%	1.51%
Political	GPT-Image-2.5-Sunburst	90.4%	9.2%	0.35%
	GPT-Image-2.5-Flare	90.8%	8.5%	0.71%
	ChatGPT Images 2.0 (baseline)	89.4%	9.6%	1.06%
Abuse	GPT-Image-2.5-Sunburst	72.9%	25.4%	1.69%
	GPT-Image-2.5-Flare	76.3%	21.2%	2.54%
	ChatGPT Images 2.0 (baseline)	75.4%	22.0%	2.54%
Nonviolent & Violent Wrongdoing[1]	GPT-Image-2.5-Sunburst	86.9%	10.2%	2.86%
	GPT-Image-2.5-Flare	87.3%	9.4%	3.27%
	ChatGPT Images 2.0 (baseline)	87.8%	9.8%	2.45%

Other[2] (e.g., religious figures)	GPT-Image-2.5-Sunburst	74.5%	23.9%	1.64%
	GPT-Image-2.5-Flare	77.9%*	20.6%*	1.53%
	ChatGPT Images 2.0 (baseline)	72.3%	25.6%	2.08%

Within each policy area, the three outcome columns account for all evaluated prompts. An asterisk (*) indicates $p < 0.05$ versus ChatGPT Images 2.0 for the same outcome and policy area (two-sided exact McNemar test; unadjusted for multiple comparisons). No unsafe-shown difference meets this threshold.

3.3.1.1 Limitations

Automated policy labels may contain errors, and varying policy-specific sample sizes impact statistical precision. The findings apply to a fixed adversarial test set as well as the model and safeguard configurations evaluated.

3.4 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-2.5-Sunburst nor GPT-Image-2.5-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.

[1] *We find that the GPT-Images-2.5 series generally performs on par with or better than GPT Images 2.0. Minor 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.*

The results indicate that neither GPT-Image-2.5-Sunburst nor GPT-Image-2.5-Flare cross the Bio High threshold or the Cyber High threshold.

We, regardless and as done in GPT-Images 2.0, take a precautionary approach and apply mitigations appropriate for a model treated as High capability in the biological-risk domain. These mitigations include an image-specific adaptation of our existing biological-risk safety policy, which we apply to all ChatGPT Images 2.5 inputs and outputs using our safety reasoning model.

3.4.1 Live Blocking

As with GPT-Images 2.0, we use a safety reasoning model to detect and block all ChatGPT Images 2.5 outputs that are flagged as violating the above-noted image-specific variant of our biological safety policy. To test the performance of the safety monitor, we developed an evaluation set of images and used them to evaluate our safety monitor's performance. We found that it has comparable rates of recall and precision to our similar live text-based biological risk mitigation systems. As with our other policy blocks, we apply this blocking at both the image input (editing) and image output layers and generation stops if any image is detected to violate.

3.4.2 Offline conversation review, flagging, blocking

As with GPT-Images 2.0, 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 2.5, 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 2.5 and across our products.