System Card: ChatGPT Images 2.0 and Thinking mode

April 21, 2026

1 Introduction

The ChatGPT Images 2.0 model is a major step forward in image generation capabilities, including significantly enhanced world knowledge, instruction following, and generating detail and complexity such as dense text. The new thinking mode capability introduced along with the model adds reasoning and tool use to the image generation process, allowing the system to integrate live web search data, generate multiple images from a single prompt, and use our reasoning stack to turn a basic prompt into a well-researched and thought-through final image.

The core safety stack we are using with ChatGPT Images 2.0 and thinking mode is based on the same foundations as our ChatGPT Images 1.5 safety stack, with additional safeguards to address new risks that emerge as models become more capable.

2 Observed Safety Challenges, Evaluations, and Mitigations

2.1 New Safety Challenges

Compared to our past GPT-4o Image Generation (1.0) and 1.5 deployments ChatGPT Images 2.0 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 strong protections in place to help prevent them from reaching users, including safeguards at the prompt (text) and image layers.

2.2 Safety stack

ChatGPT Images 2.0 includes multiple layers of image-specific safety protection. We have special-purpose safety text classifiers 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, 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 use safety classifiers 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 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 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.

2.3 Safety and Policy Evaluations

2.3.1 Automated Evaluations and Adversarial Testing

We used an automated evaluation to measure the efficacy of our safety stack with the ChatGPT Images 2.0 model both in standard as well as in thinking mode (which uses reasoning and tools to improve the final image quality). 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:

- **G:** Images generated from adversarial prompts as part of our automated evaluation process
- **V:** Images which were judged by consensus by the evaluation's grader to be violating our policy. The grader is run with high reasoning effort and we use the consensus of 10 total judgements.

- **IT:** Of the images graded as violative (V below), the total fraction caught by the image-layer safety model. This measures how well the image review layer detects bad generated images.
- **PT:** Of the images in V, the total fraction caught by the prompt-layer safety model. This measures how well the prompt review layer detects violative generated images.
- **PI:** Prompt-Only incremental: Of the images in V, the fraction caught by the prompt layer that were not also caught by the image layer. This shows the incremental protection added by prompt filtering beyond the image layer.
- **CT:** Combined total detection rate (also known as recall). Of the images in V, the percent caught by either the prompt layer or the image layer. This is the union of both layers and best reflects end-to-end blocking of bad images.
- **RU:** Remaining undetected: Of the images in V, the fraction missed by both prompt layer and image layer. These are the violative outputs that would not have been blocked by the safety stack.
- **SO:** Safe output - of all images generated in the eval, the fraction that either were graded to not violate policy or were caught by the measured safety stack.

Table 1: **Results**

Model	G: Images Generated	V: Violative outputs (as judged by grader consensus)	IT: Image-Layer: Total Detected	PT: Prompt-Layer: Total Detected	PO: Prompt-Only Incremental Detected	CT: Combined Total Detected	RU: Remaining Undetected	SO: Safe Output
Images 2.0	3112	685/3112 (22.0%)	598/685 (87.3%)	529/685 (77.2%)	60/685 (8.8%)	658/685 (96.1%)	27/685 (3.9%)	3085/3112 (99.1%)
Images 2.0 Thinking	6944*	464/6944 (6.7%)	357/464 (76.9%)	222/464 (47.8%)	49/464 (10.6%)	406/464 (87.5%)	58/464 (12.5%)	6886/6944 (99.2%)

** Note: due to the very small number of violative images generated on the first run of 3112 prompts against thinking mode, we re-ran the same prompts additional times to collect enough data to more accurately measure recall. That is why we have a larger number of total generation attempts (6944 vs 3112) for thinking vs. instant mode.*

It is worth emphasizing that this evaluation was conducted on prompts that were designed to elicit an image which violates our policies, and don't reflect expected outputs from a general sample of user queries.

2.3.2 Analysis

The image-layer total detected **(IT)** and prompt-layer total detected (**PT)** columns are standalone measurements over the same set of images graded to be violative (**V**): they answer "would this layer have caught the bad output/request?" The "Prompt-Only Incremental" column shows the additional bad cases caught by the prompt layer that were not caught by the image layer. "Combined Caught" is the union of prompt-layer and image-layer catches, which is the relevant end-to-end production safety metric.

The relative number of blocks occurring in each layer differs to some degree between instant and thinking mode (e.g. 22% blocking at the generation layer for the base model vs. 6.7% for thinking). This is because the thinking model is trained to safely transform (via Safe Completions) adversarial requests into safe ones rather than simply producing the requested violative content. As a result, the pool of truly violative generated images in thinking mode is substantially smaller. In the pooled evaluations of our final thinking mode checkpoint, the offline grader marked 464/6944* generated images (**V** above) as policy-violating, a 6.7% bad-image production rate. This compares to 685/3112 (22.0%) in instant mode.

Importantly, this does not indicate a weakness in end-to-end safety of the instant model. Instant mode is still strongly protected by the production safety stack, whose effect is reflected in **CT** above. Our safety reasoning model catches 598/685 of the images deemed by the grader to violate policy (87.3% recall), and the combined recall of prompt + image stack catches 658/685 (96.1% combined recall), resulting in 3085/3112 (99.1%) adversarial prompts leading to a safe output. Our thinking mode checkpoint improves the safety profile in a different way: it produces fewer bad images upstream, and after the combined safety stack reaches 6886/6944 (99.2%).

2.4 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 allow scaling of Cybersecurity attacks or enable AI Self-Improvement, we do not have evidence that ChatGPT Images 2.0 poses meaningful risk in these categories.

In the Biological and Chemical domain, we tested ChatGPT Images 2.0 with a series of prompts that were designed to elicit an output image (e.g. an infographic) that could assist a novice in creating dangerous substances such as biotoxins. We then asked a bioweapons expert to validate these images for risk. In limited cases, they determined that the resulting outputs were accurate enough to potentially provide novice uplift on harmful tasks. We therefore set our mitigations as if this model were high capability in biology. This included developing a new, image-specific variant of our existing biological risk safety policy, and we are applying this policy to all ChatGPT Images 2.0 inputs and outputs using our safety reasoning model.

2.4.1 Live Blocking

We use a safety reasoning model to detect and block all ChatGPT Images 2.0 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 with 772 images that fell into this risk category. We used those images to evaluate our safety monitor's performance, and 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.

2.4.2 Offline conversation review, flagging, blocking

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.

3 Image Provenance

We have continued to prioritize enhancing our provenance tools. For ChatGPT Images 2.0, 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.
- Integrating an imperceptible, robust, and content-specific watermark alongside internal tooling to help assess whether a certain image was created by our products.

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 2.0 and across our products.