GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum

OpenAI

November 12, 2025

1 Introduction 2
2 Baseline Model Safety Evaluations 2
2.1 Disallowed Content Evaluations 2
Early signal on prevalence of undesired responses for sensitive situations 2
Mental Health, Emotional Reliance, and Self Harm and Suicide 3
2.2 Jailbreaks 3
2.3 Vision 3
3 Preparedness Framework 4
References 4

1 Introduction

As described in our blog, GPT-5.1 Instant and GPT-5.1 Thinking are the next iteration of our GPT-5 models. GPT-5.1 Instant is more conversational than our earlier chat model, with improved instruction following and an adaptive reasoning capability that lets it decide when to think before responding. GPT-5.1 Thinking adapts thinking time more precisely to each question. GPT-5.1 Auto will continue to route each query to the model best suited for it, so that in most cases, the user does not need to choose a model at all.

The comprehensive safety mitigations for these models are largely the same as we described in the GPT-5 System Card. This system card addendum provides updated baseline safety metrics for these new model versions. As we noted in our recent GPT-5 system card addendum on sensitive conversations, we have expanded the baseline safety evaluations that we conduct as part of pre-deployment safety review to include evaluations for mental health (covering situations where there are signs that a user may be experiencing isolated delusions, psychosis, or mania) and for emotional reliance (covering output related to unhealthy emotional dependence or attachment to ChatGPT).

In this card we also refer to GPT-5.1 Instant as gpt-5.1-instant, and GPT-5.1 Thinking as gpt-5.1-thinking.

2 Baseline Model Safety Evaluations

2.1 Disallowed Content Evaluations

We conducted benchmark evaluations across disallowed content categories. We report here on our Production Benchmarks, a new more challenging evaluation set with conversations representative of challenging examples from production data. As we noted in previous system cards, we introduced these Production Benchmarks to help us measure continuing progress given that our earlier Standard evaluations for these categories had become relatively saturated.

These evaluations were deliberately created to be difficult. They were built around cases in which our existing models were not yet giving ideal responses, and this is reflected in the scores below. Error rates are not representative of average production traffic. The primary metric is not_unsafe, checking that the model did not produce output that is disallowed under the relevant OpenAI policy.

Table 1: Production Benchmarks (higher is better)

Category	gpt-5-thinking	gpt-5.1-thinking	gpt-5-instant-aug15	gpt-5-instant-oct3	gpt-5.1-instant
illicit/non-violent	0.865	0.860	0.700	0.807	0.853
personal data	0.966	1.000	0.966	1.000	1.000
harassment	0.815	0.747	0.683	0.745	0.836
sexual	0.906	0.895	0.782	0.951	0.917
extremism	1.000	1.000	0.922	0.978	0.989
hate	0.883	0.839	0.74	0.806	0.897
violence	0.946	0.930	0.829	0.953	0.938
sexual/minors	0.953	0.901	0.862	0.961	0.957
Illicit/violent	0.954	0.934	0.783	0.862	0.918
self-harm/intent	0.959	0.958	0.893	0.893	0.909
self-harm/instructions	0.979	0.950	0.858	0.943	0.950
mental health*	0.466	0.684	0.251	0.944	0.883
emotional reliance*	0.812	0.785	0.688	0.986	0.945

*New evaluations, as introduced in the GPT-5 update on sensitive conversations.

Overall, both gpt-5.1-thinking and gpt-5.1-instant show comparable safety performance to their GPT-5 predecessors on these particularly challenging evaluations, which are designed to target areas where our models still have room to improve.

The new gpt-5.1-thinking model shows light regressions relative to gpt-5-thinking for content involving harassment and hateful language, as well as disallowed sexual content. We are working on further improvements for these categories.

The new gpt-5.1-instant model outperforms gpt-5-instant-aug15 on all above evaluations, and performs slightly worse than gpt-5-instant-oct3 on the evaluations for disallowed sexual content, violent content, mental health, and emotional reliance. We provide further context on the latter two safety categories below.

Early signal on prevalence of undesired responses for sensitive situations

In addition to these offline evaluations, we also share here some very early signal on the prevalence of undesired responses for sensitive situations based on online measurements that we ran during A/B testing. Given the extremely low prevalence of undesired model responses for sensitive situations, combined with the relatively small size of A/B tests, these online measurements have wide error bars. However, they can help provide early signal on potential improvements or regressions. After launch, we continue to run these measurements in order to gain more precise signal on prevalence of undesired responses in real-world usage, which more fully informs whether further mitigations are needed (such as routing to specific safer models). We report more information on the results of these early online measurements for mental health, emotional reliance, and self harm and suicide below.

Online measurements and offline evaluations capture different elements of safety performance. Online measurements can provide real-time signals on the prevalence of risks in deployment, and are able to capture shifts in live user behavior with our models. In contrast, our offline evaluations focus on challenging conversations closer to a "worst case," and are typically very long conversations seeded with undesired behavior from past models in the previous turns.

Mental Health, Emotional Reliance, and Self Harm and Suicide

Mental health: On offline evaluations (i.e., the Production Benchmarks table shown above), gpt-5.1-instant shows a slight regression relative to gpt-5-instant-oct3, but still outperforms gpt-5-instant-aug15. gpt-5.1-thinking improves relative to gpt-5-thinking. On early online measurements, both gpt-5.1-instant and gpt-5.1-thinking show a slight, but low statistical confidence, improvement relative to gpt-5-instant-oct3 and gpt-5-thinking, respectively.

As mentioned above, our evaluations capture challenging conversations that may not be representative of average production traffic. We will continue to investigate the mental health performance of this model post-launch.

Emotional reliance: On offline evaluations, both gpt-5.1-instant and gpt-5.1-thinking show a slight regression relative to gpt-5-instant-oct3 and gpt-5-thinking, respectively. gpt-5.1-instant still improved relative to gpt-5-instant-aug15. Preliminary online measurements also show a regression for gpt-5.1-instant compared to gpt-5-instant-oct3, although the regression is low statistical confidence. Even with this possible regression, gpt-5.1-instant still performs better than gpt-5-instant-aug15 on online measurements. gpt-5.1-thinking shows an improvement in preliminary online measurements relative to gpt-5-thinking with high statistical confidence. We are further investigating the performance of these models on emotional reliance and are committed to improving the models' behavior and updating our safeguards where needed.

Self harm and suicide: Our preliminary online measurements were neutral for gpt-5.1-instant relative to gpt-5-instant-oct3, and showed improvements for gpt-5.1-thinking relative to gpt-5-thinking. However, these estimates have low statistical confidence.

2.2 Jailbreaks

We evaluate the robustness of models to jailbreaks: adversarial prompts that purposely try to circumvent model refusals for content it's not supposed to produce.

Below is an adaptation of the academic jailbreak eval, StrongReject [1]. This eval inserts a known jailbreak into an example from disallowed content evals. We then run it through the same policy graders we use for disallowed content checks. We test jailbreak techniques on base prompts across harm categories, and evaluate for not_unsafe according to relevant policy.

Table 2: StrongReject

metric	gpt-5-thinking	gpt-5.1-thinking	gpt-5-instant- aug15	gpt-5-instant-oct3	gpt-5.1-instant
not_unsafe	0.974	0.967	0.683	0.850	0.976

We find that gpt-5.1-instant performs better than its predecessor, and that gpt-5.1-thinking is on par with its predecessor.

2.3 Vision

We ran the image input evaluations introduced with ChatGPT agent, that evaluate for not_unsafe model output, given disallowed combined text and image input.

Table 3: Image input evaluations, with metric not_unsafe (higher is better)

Category	gpt-5-thinking	gpt-5.1-thinking	gpt-5-instant- aug15	gpt-5-instant-oct3	gpt-5.1-instant
hate	0.984	0.980	0.982	0.990	0.993
extremism	0.991	0.993	0.986	0.986	0.996
illicit	0.994	0.980	0.986	1.000	0.992
attack planning	1.000	1.000	1.000	1.000	1.00
self-harm	0.976	0.936	0.983	0.975	0.960
harms-erotic	0.990	0.990	0.994	0.999	0.999

We find that both the instant and thinking variations of GPT-5.1 perform generally on par with their predecessors. We are observing a regression of gpt-5.1-thinking on self-harm prompts with image inputs and are working on further improvements.

3 Preparedness Framework

GPT-5's frontier capabilities are assessed under the Preparedness Framework as described in the original GPT-5 system card. As we did for GPT-5 at launch, we are continuing to treat GPT-5.1 as High risk in the Biological and Chemical domain, and continuing to apply the corresponding safeguards. For cybersecurity and AI self-improvement, evaluations of near-final checkpoints indicate that, like their GPT-5 predecessor models, GPT-5.1 models do not have a plausible chance of reaching a High threshold.

References

[1] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, *et al.*, "A strongreject for empty jailbreaks," *arXiv preprint arXiv:2402.10260*, 2024.