GPT-5.3 Instant System Card

OpenAI

March 3, 2026

1 Introduction

GPT-5.3 Instant is the newest addition to the GPT-5 series. As described in our blog, GPT-5.3 Instant responds faster, delivers richer and better-contextualized answers when searching the web, and reduces unnecessary dead ends, caveats, and overly declarative phrasing that can interrupt the flow of conversation. The comprehensive safety mitigation approach for this model is largely the same as that described for GPT-5.2 Instant in the GPT-5.2 System Card.

In this card we also refer to GPT-5.3 Instant as gpt-5.3-instant.

2 Model Data and Training

Like OpenAI's other models, this model 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 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.

Note that comparison values from previously-launched models are from the latest versions of those models, so may vary slightly from values published at launch for those models.[1]

3 Safety

3.1 Disallowed Content

We conducted benchmark evaluations across disallowed content categories. We report here on our Production Benchmarks, an 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 metric is not_unsafe, checking that the model did not produce output that is disallowed under the relevant OpenAI policy.

Values from previously launched models are from the latest versions of those models, and are subject to some variation. Values may vary slightly from values published at launch for those models.

[1] GPT-5.3 Instant is intended to be used in accordance with OpenAI's Usage Policies, Service Terms, and Terms of Use. These policies apply universally to OpenAI services and are designed to ensure safe and responsible usage of AI technology. You can review OpenAI's Usage Policies at openai.com/policies/usage-policies/ .

If you need assistance with respect to GPT-5.3 Instant, you can find further information on OpenAI's website (openai.com), or you can contact OpenAI Support by opening the chat bubble icon displayed at the bottom-right of help.openai.com.

Table 1: Production Benchmarks (higher is better)

Content type	**gpt-5.1-instant**	**gpt-5.2-instant**	**gpt-5.3-instant**
violent illicit behavior	0.962	0.965	0.926
nonviolent illicit behavior	0.656	0.832	0.921
Self-harm	0.874	0.923	0.895
biology	1.000	1.00	1.00
sexual content	0.930	0.926	0.866
extremism (propaganda praise or assistance)	1.000	0.979	0.981
hate (abusive conduct protected class)	0.959	0.851	0.868
graphic violence or physical injury	0.889	0.852	0.781

On average, the model performs above gpt-5.1-instant and below gpt-5.2-instant on our disallowed content evaluations. gpt-5.3-instant shows regressions relative to gpt-5.2-instant and gpt-5.1-instant for disallowed sexual content, and relative to gpt-5.2-instant for self-harm on both standard and dynamic evaluations. The regressions for graphic violence and violent illicit behavior have low statistical significance. For other categories, gpt-5.3-instant is either on par or improves upon previous launches.

In addition to the benchmark evaluations, we reviewed safety during online experimentation and did not observe an increase in undesirable responses for self-harm. We will continue monitoring after launch to verify our test results and investigate the disparities between our offline evaluations and online testing.

For disallowed sexual content, we deploy system-level safeguards in ChatGPT intended to mitigate this behavior. We are continuing to improve our safeguards in this area and these learnings will inform any future releases.

Table 2: Dynamic Mental Health Evaluations

Evaluation	**gpt-5.1-instant**	**gpt-5.2-instant**	**gpt-5.3-instant**
mental health*	0.832	1.000	0.985
emotional reliance*	0.945	0.952	0.992
Self-harm *	0.845	0.920	0.911

*Ahead of the GPT-5.3 Instant launch, we implemented dynamic multi-turn evaluations for mental health, emotional reliance, and self-harm that simulate extended conversations across these domains. Rather than assessing a single response within a fixed dialogue, these evaluations allow conversations to evolve in response to the model's outputs, creating varied trajectories during testing that better reflect real user interactions. This approach helps identify potential issues that may only emerge over the course of long exchanges and provides an even more rigorous

test than prior static multi-turn methods. By utilizing realistic, yet adversarial user simulations, these evaluations have enabled continued improvements in safety performance, particularly in areas where earlier evaluation frameworks had reached saturation.

Our standard evaluations measure whether the final model response violates our policies. In these dynamic conversations, we instead evaluate whether any assistant response violates policy and report the percentage of policy-compliant responses. The metric used is not_unsafe, representing the share of assistant messages that do not violate safety policies.

4 Health Performance

4.1 HealthBench

Chatbots can empower consumers to better understand their health and help health professionals deliver better care [1] [2]. We evaluate GPT-5.3 on HealthBench [3], an evaluation of health performance and safety. HealthBench comprises 5,000 realistic (potentially multi-turn) health conversations. Model responses are evaluated with example-specific rubrics. We report results on three variants, HealthBench, HealthBench Hard, and HealthBench Consensus.

Table 3: HealthBench

Metric	**gpt-5.2-instant**	**gpt-5.3-instant**
HealthBench	55.4%	54.1%
Hard	26.8%	25.9%
Consensus	95.8%	95.3%
Length	2101 chars	2140 chars

The major wins and losses below refer only to consensus-criteria deltas greater than 2.0% versus GPT-5.2 instant.

Relative to GPT-5.2 instant, GPT-5.3 Instant scores 54.1% on HealthBench (-1.3%), 25.9% on Hard (-0.9%), and 95.3% on Consensus (-0.5%); at 2140 chars on average (+1.9%), it has slightly worse performance at essentially matched length. On consensus criteria, its main strengths are better context-seeking when important information is missing (+4.4%) and better hedging behavior in irreducible-uncertainty settings (+4.0%). On consensus criteria, its main weaknesses are poorer emergency-triage behavior, especially around when to seek context before referral (-10.1%), and lower accuracy when local healthcare context matters but remains unclear (-5.5%).

References

[1] OpenAI, "Introducing gpt-5," Aug. 2025. Accessed: 2025-12-10.

[2] OpenAI, "Pioneering an AI clinical copilot with Penda health," July 2025. Accessed: 2025-12-10.

[3] OpenAI, "Introducing healthbench," May 2025. Accessed: 2025-12-10.