Addendum to GPT-5 System Card: Sensitive Conversations

OpenAI

October 27, 2025

1 Introduction

When we launched GPT-5, we noted in the system card that we were working to establish better benchmarks and to continue to strengthen model safety in areas related to mental and emotional distress. On October 3, we deployed an update that reflected those efforts, improving ChatGPT's default model to better recognize and support people in moments of distress. In this effort, we worked with more than 170 mental health experts to help ChatGPT more reliably recognize signs of distress, respond with care, and guide people toward real-world support–reducing responses that fall short of our desired behavior by 65-80%.

We are publishing a related blog post that gives more information about this work, and this addendum to the GPT-5 system card to share baseline safety evaluations. These evaluations compare the August 15 version of ChatGPT's default model, also known as GPT-5 Instant, to the updated one launched October 3.

2 Observed Safety Challenges and Mitigations

2.1 Disallowed Content (Production Benchmarks)

We evaluate completions using LLM-based grading models. The primary metric is not_unsafe, checking that the model did not produce output that is disallowed under the relevant OpenAI policy.

With this launch, we created the following new evaluation sets:

- Emotional Reliance not_unsafe - tests that the model does not produce disallowed content under our policies related to unhealthy emotional dependence or attachment to ChatGPT
- Mental Health not_unsafe - tests that the model does not produce disallowed content under our policies in situations where there are signs that a user may be experiencing isolated delusions, psychosis, or mania

Our new evaluations for emotional reliance and mental health, like the existing evaluations in this Production Benchmarks set, were deliberately intended to be challenging. They were built around cases in which our existing models were not yet giving ideal responses, and this is reflected in the initial scores below. Error rates are not representative of average production traffic.

Note these are new evaluations and may evolve over time.

Table 1: Disallowed Content Evaluations - Production Benchmarks (higher is better)

Category	gpt-5-aug-15	gpt-5-oct-3
non-violent hate	0.800	0.853
personal-data	0.876	0.908
harassment/threatening	0.653	0.706
sexual/exploitative	0.785	0.910
sexual/minors	0.906	0.959
extremism	0.933	0.925
hate/threatening	0.780	0.791
illicit/nonviolent	0.720	0.800
illicit/violent	0.782	0.834
emotional reliance	0.507*	0.976
mental health	0.273*	0.926
self-harm/intent	0.874	0.933
self-harm/instructions	0.805	0.890

*These are new evaluations that were not available when the August 15 model launched. We have run them retrospectively for these launches.

2.2 Jailbreaks

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

- StrongReject [1]: inserts a known jailbreak into an example from the above safety refusal eval. We then run it through the same policy graders we use for disallowed content checks. We test jailbreak techniques on base prompts across several harm categories, and evaluate for not_unsafe according to relevant policy.

Table 2: Jailbreak evaluations

Category	metric	gpt-5-aug-15	gpt-5-oct-3
illicit/non-violent-crime prompts	not_unsafe	0.926	0.957
violence prompts	not_unsafe	0.942	0.968
abuse / disinformation / hate prompts	not_unsafe	0.967	0.981
sexual-content prompts	not_unsafe	0.954	0.969

2.3 Image Input

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 (higher is better)

Category	**gpt-5-aug-15**	**gpt-5-oct-3**
hate	0.982	0.990
extremism	0.984	0.986
illicit	0.990	0.986
attack planning	1.000	0.995
self-harm	0.994	0.994
harms-erotic	0.990	0.996

2.4 Hallucination

We evaluate hallucinations via SimpleQA, a diverse dataset of four-thousand fact-seeking questions with short answers and measures model accuracy for attempted answers.

We consider two metrics: accuracy (did the model answer the question correctly) and hallucination rate (checking how often the model hallucinated). Further details on hallucinations in GPT-5, including our work on newer evaluations and progress in our reasoning models, can be found in the original GPT-5 system card.

Table 4: Hallucination evaluation

Dataset	**Metric**	**gpt-5-aug-15**	**gpt-5-oct-3**
SimpleQA	accuracy (higher is better)	0.46	.44
	hallucination rate (lower is better)	0.49	.52

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.