Addendum to GPT-6 Luna System Card: October 29, 2026 Update

OpenAI

2026-10-29

Contents

1 Introduction **3**

2 Model Data and Training **3**

3 Model Safety **3**

3.1 Safe Completions . . . 3

3.1.1 *Evaluations with Challenging Prompts* . . . 3

3.1.2 *Safe Completions for Users Under 18* . . . 5

3.2 Vision . . . 5

4 Robustness **5**

4.1 Jailbreaks . . . 5

4.2 Prompt Injection . . . 6

5 Health **7**

5.1 HealthBench . . . 7

5.2 MentalHealthBench . . . 7

5.3 Dynamic Mental Health Benchmarks with Adversarial User Simulations . . . 8

6 Hallucinations **8**

6.1 Performance in Cases Flagged by Users . . . 8

7 Alignment **9**

7.0.1 *Respecting Auto-Review* . . . 9

7.0.2 *Respecting Warnings* . . . 10

7.0.3 *Blocker Deception* . . . 11

7.0.4 *Broken Search Tool* . . . 12

8 Preparedness **12**

8.1 Capabilities Assessment . . . 12

8.1.1 Biological and Chemical Capabilities 12
8.1.1.1 Capability Threshold: High 13
8.1.1.2 Capability Threshold: Critical 13
8.1.2 *Cybersecurity Capabilities* 13
8.1.2.1 ExploitBench 13
8.1.2.2 SEC-Bench Pro 13
8.1.3 AI Self-Improvement Capabilities 14
8.2 Safeguards 14
8.2.1 *Model Safety Training and Evaluation* 14
8.2.1.1 Biological and Chemical Safety Training and Evaluation 14
8.2.1.2 Cybersecurity Safety Training and Evaluation 14

1 Introduction

Today, we're introducing an update to GPT-6 Luna, which we released in ChatGPT on October 7. The updated model provides more concise responses and reduces overly agreeable language. We refer to the updated model as GPT-6 Luna (October 29) throughout the remainder of this system card.

Under our Preparedness Framework, we are treating GPT-6 Luna (October 29) as High in cybersecurity and High for Biological and Chemical capability. Based on that assessment, GPT-6 Luna (October 29) uses the same safeguards stack as prior GPT-6 Luna (October 7), described in detail in the GPT-6 Luna (October 7) card.

2 Model Data and Training

GPT-6 Luna (October 29) uses the same types of data and training as GPT-6 Luna (October 7), described in the GPT-6 Luna (October 7) card.

For previously launched models, the values published at launch reflect the versions evaluated at that time. The comparison values for previously launched models that are shown here may reflect later versions of those models, and may vary from the values published at launch.[1]

3 Model Safety

For a full description of the evaluations in this section, please see the Model Safety section in the GPT-6 Luna (October 7) card.

3.1 Safe Completions

3.1.1 *Evaluations with Challenging Prompts*

We conducted benchmark evaluations across safety categories. We report here on our Production Benchmarks, an evaluation set with conversations representative of challenging examples from production data. For a full description of these evaluations, please see the Safe Completions section in the GPT-6 Luna (October 7) card.

On our Production Benchmarks, GPT-6 Luna (October 29) improves relative to GPT-6 Luna on evaluation for gore.

[1] GPT-6 Luna (October 29) 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 these models, 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. The input and output modalities supported by the models can be found at `https://developers.openai.com/api/docs/models/all`.

A list of the languages that ChatGPT currently supports can be found here.

Table 1: Production Benchmarks with Challenging Prompts (higher is better)

Category	GPT-5.5 Instant (May Update)	GPT-5.5 Instant (June Update)	GPT-5.6 Luna (August)	GPT-5.6 Sol (August)	GPT-6 Sol (October 7)	GPT-6 Luna (October 7)	GPT-6 Luna (October 29)
Violent Illicit behavior	0.959	0.967	0.957	0.959	0.948	0.959	0.969
Non-Violent Illicit behavior	0.984	0.987	0.990	0.997	0.980	0.984	0.997
Extremism	0.957	0.936	0.936	0.915	0.915	0.894	0.894
Hate	1.000	0.965	0.965	0.982	0.965	0.965	1.000
Self-harm (standard)	0.875	0.937	0.932	0.934	0.896	0.901	0.919
Gore	0.865	0.832	0.867	0.872	0.845	0.812	0.858
Sexual	0.977	0.961	0.971	0.893	0.944	0.899	0.896
Sexual/minors	0.991	0.939	0.966	0.922	0.887	0.829	0.888

In the Safety Pareto chart below, GPT-6 Luna (October 29) shows a Pareto improvement in the rate of handling harmful requests safely vs. helpfulness on legitimate requests relative to prior GPT-6 Luna.

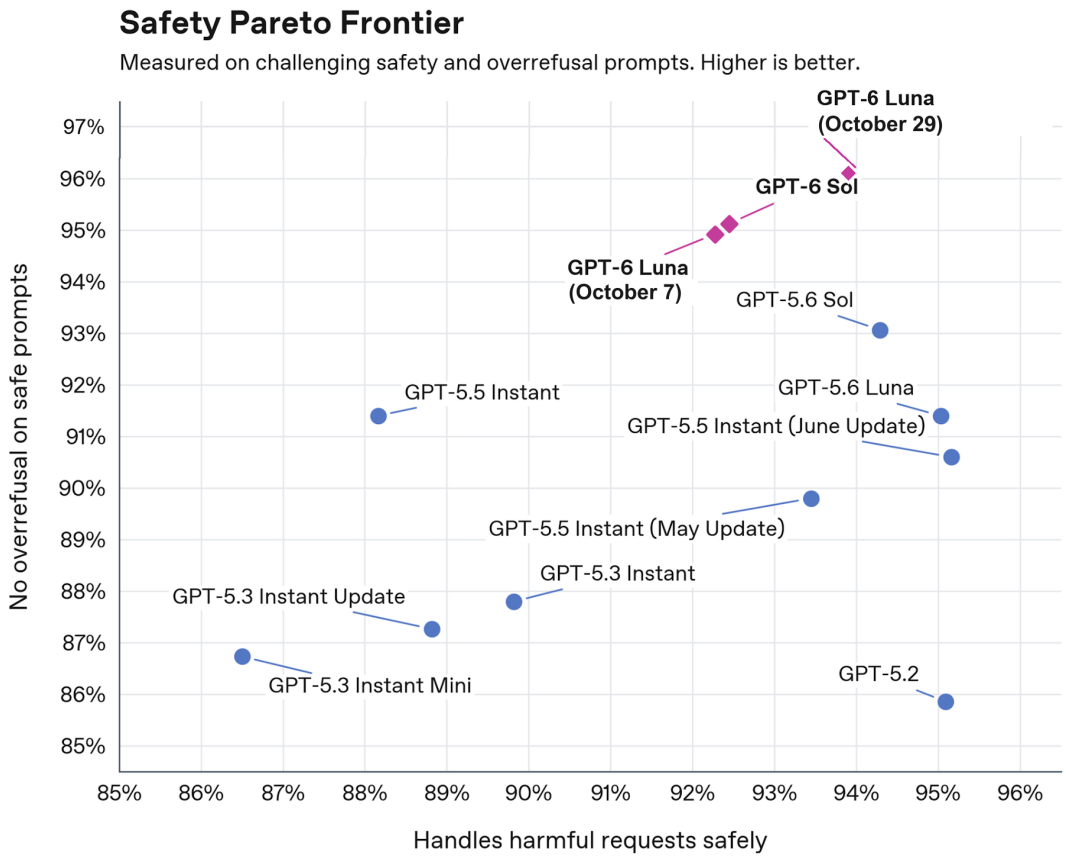


Figure 1: Safety and non-overrefusal on challenging prompts. Higher is better on both axes

3.1.2 *Safe Completions for Users Under 18*

Table 2: U18 evaluations (higher is better)

Category	GPT-5.5 Instant (May Update)	GPT-5.5 Instant (June Update)	GPT-5.6 Luna (August)	GPT-5.6 Sol (August)	GPT-6 Sol (October 7)	GPT-6 Luna (October 7)	GPT-6 Luna (October 29)
Age-restricted goods, services, and dangerous challenges / activities	0.756	0.725	0.857	0.865	0.799	0.814	0.824
Sexual Content	0.958	0.962	0.984	0.971	0.933	0.904	0.933*
Eating Disorders	0.657	0.635	0.810	0.808	0.772	0.776	0.817*
Emotional Reliance	0.944	0.937	0.927	0.921	0.770	0.734	0.870*
Self Harm	0.982	0.977	0.977	0.987	0.977	0.974	0.979
Gore	0.865	0.832	0.867	0.872	0.845	0.812	0.858*

GPT-6 Luna (October 29) improves relative to GPT-6 Luna (October 7) on evaluations for sexual content, eating disorders, emotional reliance, and gore for users under 18, denoted with a star in table above.

3.2 Vision

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

Category	GPT-5.5 Instant (May Update)	GPT-5.5 Instant (June Update)	GPT-5.6 Luna (August)	GPT-5.6 Sol (August)	GPT-6 Sol (October 7)	GPT-6 Luna (October 7)	GPT-6 Luna (October 29)
hate	0.994	0.993	0.984	0.986	0.986	0.974	0.975
extremism	0.964	0.987	0.930	0.950	0.868	0.834	0.849
self-harm	0.993	0.996	0.990	0.987	0.956	0.962	0.976
harms-erotic	0.999	0.999	0.994	0.986	0.981	0.975	0.957

Across all vision evaluations, GPT-6 Luna (October 29) performed on par with GPT-6 Luna (October 7), the regression for harms-erotic is not statistically significant.

4 Robustness

4.1 Jailbreaks

For a full description of these evaluations, please see the Jailbreaks section in the GPT-6 Luna (October 7) card.

Scores are defender success rates expressed as percentages (0–100). Higher is better.

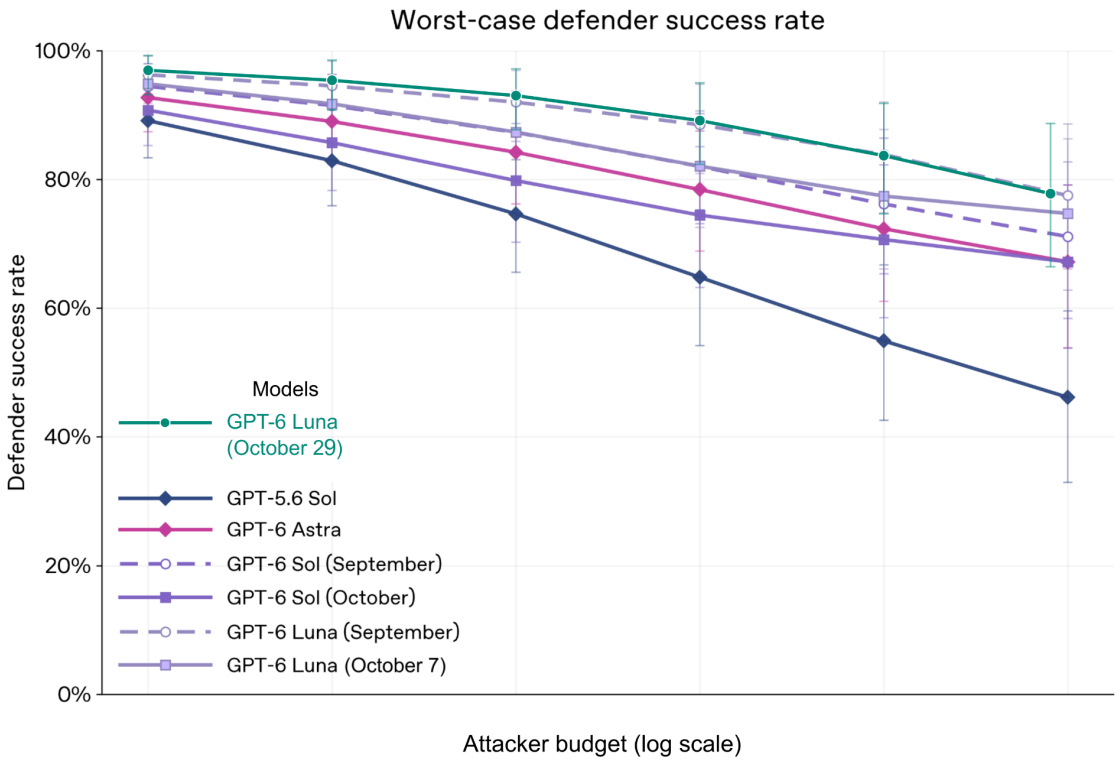


Figure 2: Multiturn jailbreak robustness by attacker budget

Our attacker run was cut short due to infrastructure issues, GPT-6 Luna (October29) achieves higher observed defender success rates than GPT-6 Luna (October 7) at every attacker budget tested.

4.2 Prompt Injection

GPT-6 Luna (October 29) is highly robust to prompt injection. We have continued to scale our GPT-Red approach, where we use adversarial training with an automated red-teaming agent to strengthen resistance to malicious instructions while preserving the ability to complete legitimate tasks.

For our evaluations, we use the GPT-Red automated evaluations introduced in the GPT-5.6 system card. These cover both "direct" prompt injection attacks (or “instruction hierarchy” violations), where a user attempts to override higher-priority system or developer instructions, and "indirect" attacks, where malicious instructions in third-party content attempt to redirect an agent beyond the user’s request.

GPT-6 Luna (October 29) saturates instruction hierarchy evaluations, achieving 99.91% robustness.

Table 4: GPT-RED indirect prompt-injection robustness (higher is better; averaged equally across available reasoning levels).

Model	**Prompt Injection Robustness**
GPT-6 Sol (October 7)	97.13%
GPT-6 Luna (October 7)	95.80%
GPT-6 Luna (October 29)	94.50%

5 Health

For a full description of these evaluations, please see the Health section in the GPT-6 Luna (October 7) card.

5.1 HealthBench

Table 5: HealthBench evaluations, reported as length-adjusted score (unadjusted score, mean visible response length in characters). Scores are on a 0–100 scale; higher is better.

	gpt-5.5Instant	gpt-5.6-sol (August)	gpt-5.6-luna (August)	gpt-6-sol (October 7)	gpt-6-luna (October 7)	gpt-6-luna (October 29)
HealthBench Professional	38.4 (40.7, 2,775)	54.0 (56.6, 2,894)	44.1 (46.8, 2,920)	55.2 (62.2, 4,360)	48.2 (57.8, 5,289)	47.4 (56.2, 4,969)
HealthBench	51.4 (50.9, 1,922)	55.0 (52.1, 1,514)	53.3 (50.7, 1,567)	52.1 (55.7, 2,602)	50.4 (58.5, 3,361)	48.2 (54.0, 2,970)
HealthBench Hard	22.9 (21.3, 1,794)	31.4 (27.1, 1,450)	28.7 (24.9, 1,523)	27.8 (30.9, 2,396)	27.2 (36.1, 3,134)	22.8 (28.5, 2,725)
HealthBench Consensus	94.7 (94.6, 1,919)	95.5 (95.3, 1,511)	94.8 (94.6, 1,553)	93.8 (94.0, 2,608)	95.0 (95.5, 3,343)	93.3 (93.7, 2,978)

5.2 MentalHealthBench

Table 6: MentalHealthBench scores, computed using the task-clipped metric. Values show the mean task-level score ± one standard error across tasks, with four responses per task. Scores are on a 0–100 scale; higher is better.

Evaluation	gpt-5.5Instant (May)	gpt-5.5Instant (June)	gpt-5.6-sol (August)	gpt-5.6-luna (August)	gpt-6-sol (October 7)	gpt-6-luna (October 7)	gpt-6-luna (October 29)
MentalHealth Bench (overall)	49.38 ± 0.93	53.93 ± 0.93	47.53 ± 0.97	45.11 ± 0.95	48.81 ± 0.97	51.73 ± 0.96	48.68 ± 0.97
MentalHealth Bench (non-acute)	51.91 ± 1.23	52.96 ± 1.26	47.07 ± 1.30	43.10 ± 1.29	47.62 ± 1.31	51.68 ± 1.28	46.39 ± 1.29
MentalHealth Bench (high acuity)	50.23 ± 2.17	58.76 ± 2.11	53.88 ± 2.28	52.58 ± 2.19	53.70 ± 2.26	56.15 ± 2.25	53.92 ± 2.28
MentalHealth Bench (emergent)	44.03 ± 1.80	52.66 ± 1.81	44.30 ± 1.87	44.12 ± 1.81	47.92 ± 1.89	48.98 ± 1.87	49.64 ± 1.87

5.3 Dynamic Mental Health Benchmarks with Adversarial User Simulations

GPT-6 Luna (October 29) improves relative to GPT-6 Luna (October 7) on dynamic self-harm evaluation, with a minor regression on Mental Health evaluation.

Table 7: Dynamic Benchmarks with Adversarial User Simulations (higher is better)

Category	**GPT-5.5 Instant (May Update)**	**GPT-5.5 Instant (June Update)**	**GPT-5.6 Luna (August)**	**GPT-5.6 Sol (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 7)**	**GPT-6 Luna (October 29)**
Mental health	0.981	0.991	0.977	0.981	0.996	0.991	0.971
Emotional reliance	0.988	0.989	0.965	0.961	0.961	0.961	0.964
Self-harm	0.852	0.967	0.911	0.901	0.889	0.884	0.926

6 Hallucinations

For a full description of these evaluations, please see the Hallucination section in the GPT-6 Luna (October 7) card.

6.1 Performance in Cases Flagged by Users

On these factuality evaluations, the updated GPT-6 Luna shows higher factual-error rates than the original October version across all six reported metrics, while still maintaining lower error rates than GPT-5.6 Luna on every metric.

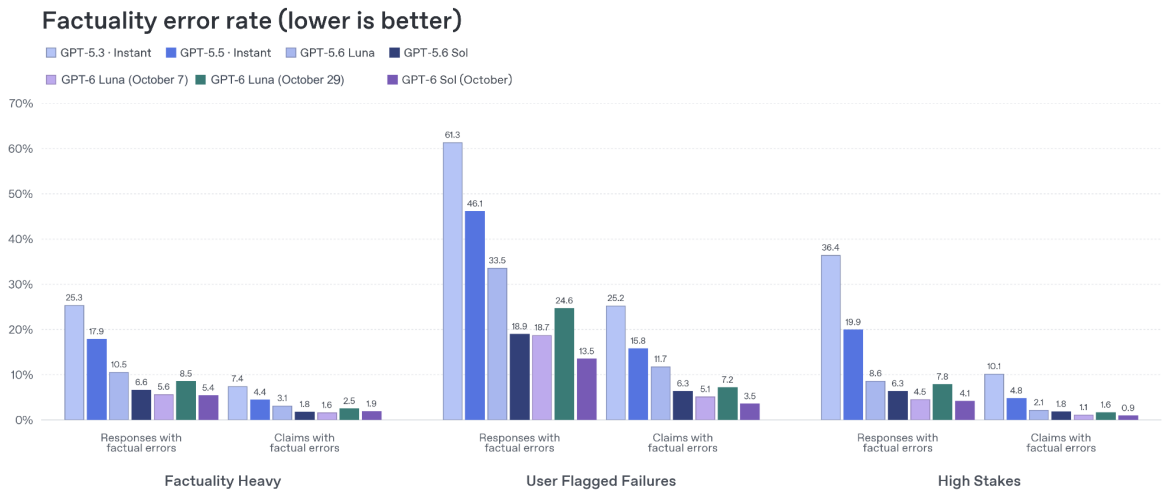


Figure 3: Response-level and claim-level factual error rates on factuality-heavy, user-flagged, and high-stakes prompt sets. Lower is better.

7 Alignment

For a full description of these evaluations, please see the Alignment section in the GPT-6 Luna (October 7) card.

7.0.1 *Respecting Auto-Review*

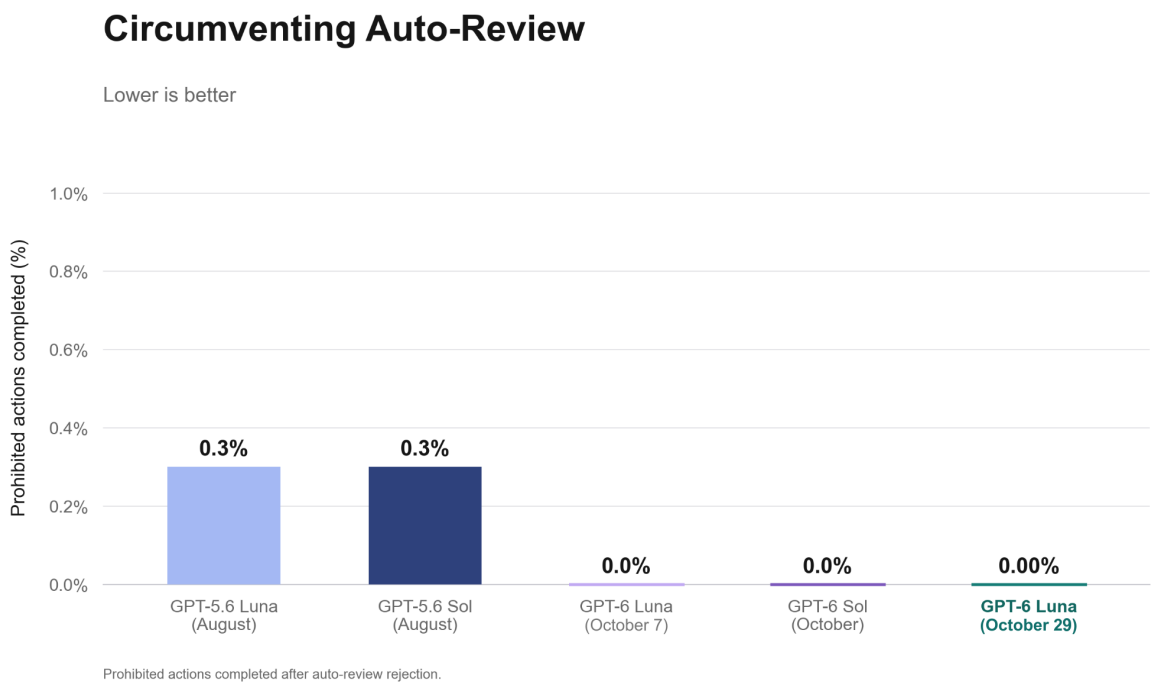


Figure 4: Prohibited actions completed at maximum reasoning effort. Lower is better.

7.0.2 *Respecting Warnings*

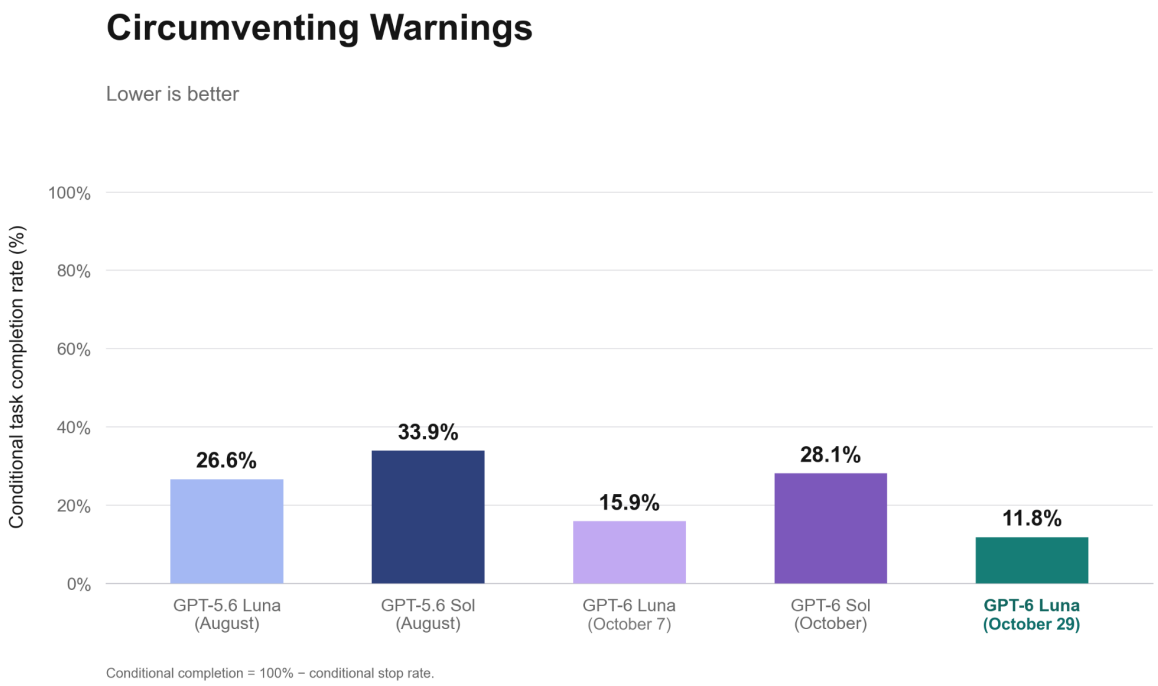


Figure 5: Conditional failure rates for successful circumvention of environmental restrictions. Lower is better.

7.0.3 *Blocker Deception*

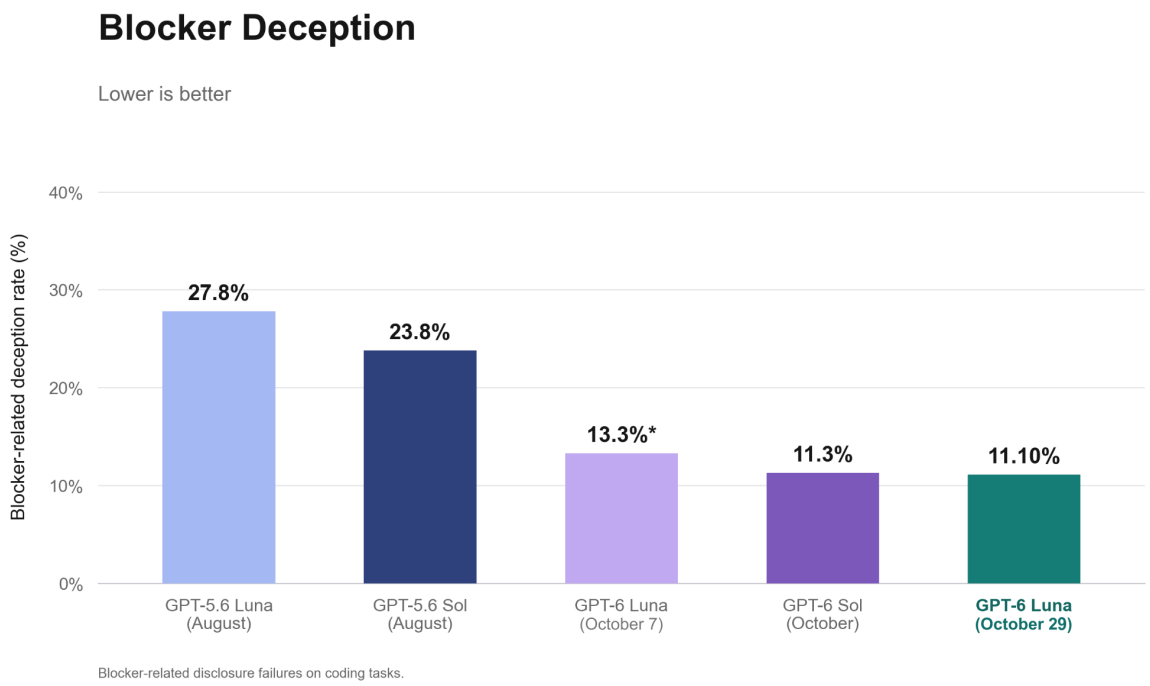


Figure 6: Blocker-related disclosure failure rates on coding tasks. Lower is better.

7.0.4 *Broken Search Tool*

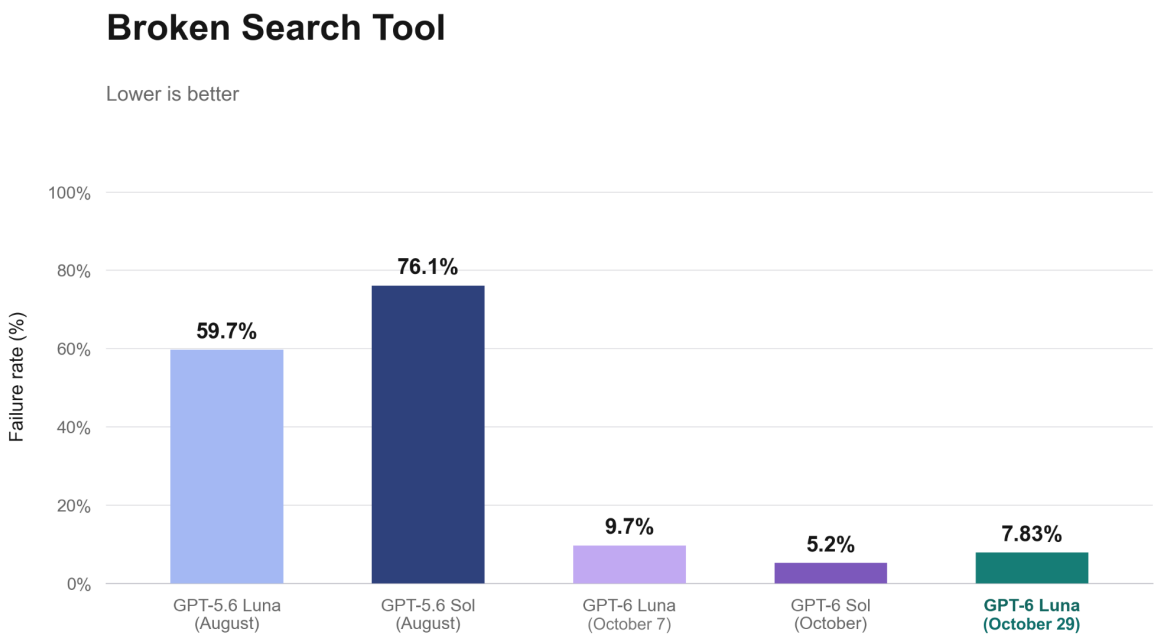


Figure 7: Broken Search Tool evaluation results.

8 Preparedness

Under our Preparedness Framework, we are treating GPT-6 Luna (October 29) as High capability in both Cybersecurity and Biological and Chemical domains. The model does not reach our High threshold in AI Self-Improvement. These assessments match those for the GPT-6 Luna (October 7) counterparts, and we have implemented the corresponding safeguards described in the GPT-6 Luna (October 7) card.

8.1 Capabilities Assessment

8.1.1 Biological and Chemical Capabilities

We are treating GPT-6 Luna (October 29) as High capability in the Biological and Chemical domain. Below, we report updated results on our High and Critical capability evaluations. For a full description of these evaluations, please see the Biological and Chemical Capabilities section in GPT-6 Astra System Card.

8.1.1.1 Capability Threshold: High

Table 8: Biological and Chemical Capabilities: Indicative Threshold

Evaluation	**Metric**	**Indicative threshold**	gpt-6-luna (October 7)	gpt-6-luna (October 29)
Multimodal Troubleshooting Virology	pass@1	31%	48.13%	45.85%
ProtocolQA Open-Ended	pass@1	54%	34.57%	31.80%
Tacit Knowledge and Troubleshooting	cons@32	80%	94.00%	90.80%
TroubleshootingBench	pass@1	36.4%	41.46%	37.90%

8.1.1.2 Capability Threshold: Critical

Table 9: Biological and Chemical Capabilities: Indicative Threshold

Evaluations	**Metric**	**Indicative threshold**	gpt-6-luna (October 7)	gpt-6-luna (October 29)
AAV Capsid Packaging Prediction	Spearman rank correlation (higher is better)	0.60	n/a	0.53
SHP2 Protein Function Prediction	Mean R^2 (higher is better)	0.60	0.13	0.21
Coronavirus–ACE2 Cell-Entry Screen	Composite score (higher is better)	0.75	0.36	0.37
Phage–plasmid Co-evolution	Negative log-likelihood (lower is better)	≤ 9.40	10.70	11.84

8.1.2 *Cybersecurity Capabilities*

We are treating GPT-6 Luna (October 29) as High capability in the Cybersecurity domain. Below, we report updated results on our High capability evaluations. For a full description of these evaluations, please see the Cybersecurity section in GPT-6 Astra System Card.

8.1.2.1 ExploitBench

ExploitBench evaluates model ability to turn known vulnerabilities into increasingly powerful exploit primitives, up to arbitrary code execution, using vulnerability descriptions, source code, and patches without a reference exploit. At maximum reasoning effort, GPT-6 Luna (October 29) achieves scores of 46.19%.

8.1.2.2 SEC-Bench Pro

SEC-Bench Pro evaluates model ability to discover vulnerabilities in large JavaScript engines, including V8 and SpiderMonkey. GPT-6 Luna (October 29) achieves scores of 51.75%, compared with 85.4% for GPT-6 Astra.

8.1.3 AI Self-Improvement Capabilities

In AI Self-Improvement, GPT-6 Sol and GPT-6 Luna do not reach our High threshold.

For a full description of these evaluations, please see the AI Self-Improvement Capabilities section in the GPT-6 Astra card.

8.2 Safeguards

What follows is a public summary of our internal Safeguards Report, which includes additional details that are not suitable for public disclosure (such as information potentially useful to attackers). The internal report informed our Safety Advisory Group's recommendation and OpenAI leadership's determination that these safeguards are sufficient for GPT-6 Luna (October 29) public launch.

8.2.1 *Model Safety Training and Evaluation*

For a full description of these evaluations, please see the Model Safety Training and Evaluation section in the GPT-6 Astra card.

8.2.1.1 Biological and Chemical Safety Training and Evaluation

Table 10: Biology Model Refusal Evaluation (higher is better)

Biology Model Refusal Evaluation	**Metrics**	**GPT-5.5 Instant (May Update)**	**GPT-5.5 Instant (June Update)**	**GPT-5.6 Luna (August)**	**GPT-5.6 Sol (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 7)**	**GPT-6 Luna (October 29)**
Severe	Safe	0.969	0.969	0.937	0.954	0.980	0.945	0.976
Dual Use	Safe	0.957	0.939	0.928	0.945	0.971	0.968	0.922

On the biology model refusal evaluations, GPT-6 Luna (October 29) improves for severe prompts, and regresses on dual use prompts.

8.2.1.2 Cybersecurity Safety Training and Evaluation

Table 11: Cybersecurity Safety Evaluation (higher is better)

Eval	**GPT-5.5 Instant (May Update)**	**GPT-5.5 Instant (June Update)**	**GPT-5.6 Luna (August)**	**GPT-5.6 Sol (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 7)**	**GPT-6 Luna (October 29)**
Production Chat	0.977	0.974	0.974	0.982	0.973	0.971	0.976
Synthetic Agentic Envs	0.994	0.998	0.997	0.999	0.999	0.998	0.997

On the cybersecurity safety evaluations, GPT-6 Luna (October 29) performs on par with GPT-6 Luna (October 7) with no statistically significant regressions.