

OpenAI

GPT-6 Sol and GPT-6 Luna: October 2026 update

2026-10-07

1 Introduction

We're launching GPT-6 to everyone in ChatGPT, across free and paid plans globally. These models will replace GPT-5.6 Sol and GPT-5.6 Luna in ChatGPT. Users accessing GPT-6 Sol and GPT-6 Luna in Codex, and via ChatGPT Work, are still using previously released versions. In this system card, we distinguish these models by their month of release: October for the versions released today, and September for the versions that remain in use in Codex and Work. See [our blog](#) for details.

GPT-6 in ChatGPT incorporates Astra's [safety advances](#). We also updated the safety training to reflect real-world use and strengthen protections against high-risk misuse in cyber, biology, and violence.

We assess safety and alignment evaluations in aggregate, weighing improvements alongside the nature and severity of regressions. Compared with GPT-5.6 Sol and GPT-5.6 Luna in ChatGPT, GPT-6 showed stronger resistance to jailbreaks, including attacks that adapt across multiple turns, as well as reductions in dishonesty, deception, and circumvention of guardrails. For areas where our safety evaluations showed regressions, we conducted manual review of failures and adversarial red-team testing and found that the disallowed responses were generally low severity. System-level mitigations to reduce the likelihood of harm are detailed throughout this system card.

Under our Preparedness Framework, we are treating this October release of GPT-6 Sol and GPT-6 Luna as High capability in both Cybersecurity and Biological and Chemical domains. Neither one reaches our High threshold in AI Self-Improvement. Based on that assessment, we've implemented the same set of safeguards for GPT-6 Sol (October) and GPT-6 Luna (October) that are detailed in the [GPT-5.6 System Card](#) for GPT-5.6 Sol and GPT-5.6 Luna.

For all safety evaluations, such as disallowed content and mental health, we measure performance of our models at their lowest reasoning deployment settings in order to capture performance for the vast majority of usage. For capabilities assessments, we evaluate the models at their maximum reasoning effort to get an upper bound of capabilities.

Contents

1	Introduction	1
2	Model Data and Training	4
3	Model Safety	4
3.1	Safe Completions	4
3.1.1	Evaluations with Challenging Prompts	4
3.1.2	Safe Completions for Users Under 18	7
3.2	Vision	9
4	Robustness	9
4.1	Jailbreaks	9
4.1.1	Multiturn Jailbreak Evaluations	9
4.2	Prompt Injection	10
5	Health	11
5.1	HealthBench	11
5.2	MentalHealthBench	12
5.3	Dynamic Mental Health Benchmarks with Adversarial User Simulations	13
6	Hallucinations	14
6.1	Performance in Cases Flagged by Users	14
7	Alignment	15
7.1	Obeying Restrictions	15
7.1.1	Respecting Auto-Review	16
7.1.2	Respecting Warnings	16
7.2	Avoiding Deceptive Interactions with Users	18
7.2.1	Blocker Deception	18
7.2.2	Broken Search Tool	19

8	Preparedness	20
8.1	Capabilities Assessment	20
8.1.1	Biological and Chemical Capabilities	20
8.1.2	Cybersecurity Capabilities	21
8.1.3	AI Self-Improvement Capabilities	22
8.2	Safeguards	22
8.2.1	Model Safety Training and Evaluation	22

2 Model Data and Training

Like OpenAI's other models, GPT-6 models were trained on diverse datasets and filtered through our data processing pipeline, including to reduce personal information.

OpenAI reasoning models are trained to reason through reinforcement learning. These models are trained to think before they answer: they can produce a long internal chain of thought before responding to the user. Through training, these models learn to refine their thinking process, try different strategies, and recognize their mistakes. Reasoning allows these models to follow specific guidelines and model policies we've set, helping them act in line with our safety expectations. This means they provide more helpful answers and better resist attempts to bypass safety rules.

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

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. As we noted in previous system cards, we introduced these Production Benchmarks to help us measure continuing progress given that earlier Standard evaluations for these categories had become relatively saturated.

¹ GPT-6 Sol and GPT-6 Luna are 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](#).

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 safe completion rate, checking that the model's response and actions are not disallowed according to our safety policies. Our evaluations are run on the model without system-level safeguards to ensure the model's underlying behavior meets our safety bar. We continue monitoring these categories after launch to evaluate online performance and further adjust safeguards as appropriate.

Values may vary slightly from values published at launch for those models. Values from previously launched models are from the latest versions of those models, and evals are subject to some variation. The comparison scores from earlier models listed below are intended to shed light on relative performance. Because policies, graders, datasets, evaluations, and other measurement details evolve over time, scores from previous system cards not included in the table below should generally not be considered directly comparable to these most recent results.

Relative to their respective GPT-5.6 counterparts, GPT-6 Sol (October) shows a statistically significant regression on standard self-harm, while GPT-6 Luna (October) shows statistically significant regressions on standard self-harm, gore, and sexual content.

We manually reviewed violative responses from these evaluations and found that violations were borderline but still generally safe. For example, GPT-6 Sol (October) and GPT-6 Luna (October) are more willing to answer informational questions about self-harm while directing users to appropriate professional resources. Based on our evaluations, it does not comply with requests to facilitate self-harm. GPT-6 Sol (October) and GPT-6 Luna (October) are more likely to engage with requests for gore and sexual content rather than refuse outright, however the outputs we reviewed were lower severity. We also conducted adversarial red teaming and did not identify any high-severity risks based on our evaluation criteria.

These evaluations do not account for system-level interventions that elevate the holistic safety of the assistant response, such as Trusted Contact, localized crisis helplines, and Parental Controls for younger users.

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)	GPT-6 Luna (October)
Violent Illicit behavior	0.959	0.967	0.957	0.959	0.948	0.959
Non-Violent Illicit behavior	0.984	0.987	0.990	0.997	0.980	0.984
Extremism	0.957	0.936	0.936	0.915	0.915	0.894
Hate	1.000	0.965	0.965	0.982	0.965	0.965
Self-harm (standard)	0.875	0.937	0.932	0.934	0.896	0.901
Gore	0.865	0.832	0.867	0.872	0.845	0.812
Sexual	0.977	0.961	0.971	0.893	0.944	0.899
Sexual/minors	0.991	0.939	0.966	0.922	0.887	0.829

In addition to measuring how safely our models respond to harmful requests, we assess whether they overrefuse legitimate requests, especially when a topic is sensitive or easy to misinterpret. In the Safety Pareto chart below, GPT-6 Sol (October) and GPT-6 Luna (October) show an improvement on helpfulness on legitimate requests relative to prior models, though it scores lower on some safety requests. Our evaluations indicate that GPT-6 Sol (October) and GPT-6 Luna (October) are less likely to refuse harmless requests or add excessive or judgmental caveats.

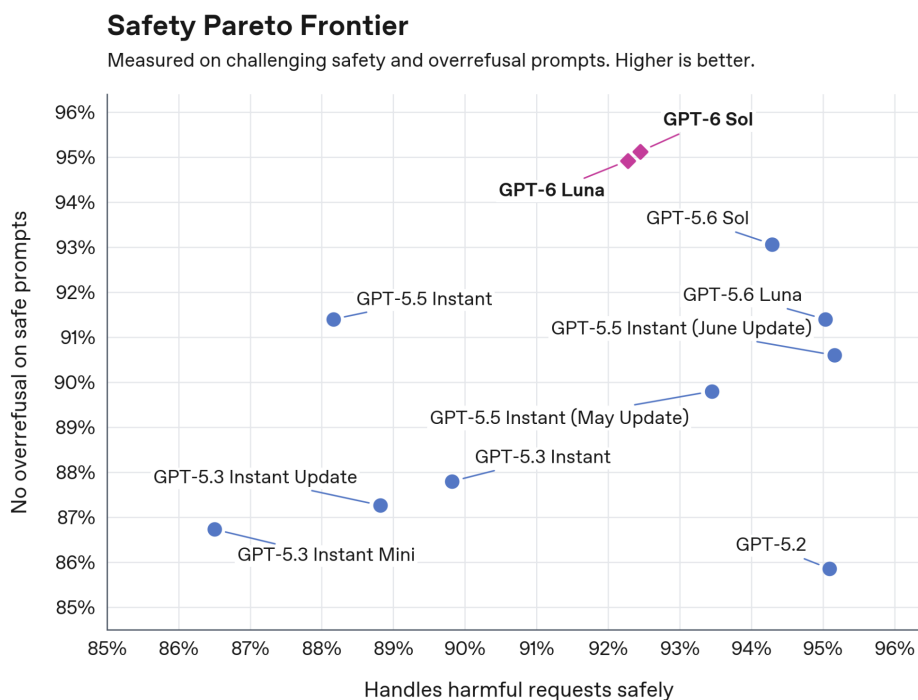


Figure 1: Safety and non-overrefusal on challenging prompts. Higher is better on both axes.

3.1.2 Safe Completions for Users Under 18

AI can help teens learn, create, explore ideas, access information, and build new skills. We design our systems carefully to enable these benefits while working to mitigate potential risks that may have outsized or more severe impacts on users under 18 (U18).

Our [Model Spec](#) outlines the principles and requirements that guide how our models should behave with teens. For users we believe may be under 18, we operationalize those principles through age-specific provisions in our safety policies. In some areas where teens may face distinct or heightened risks, these provisions establish more restrictive thresholds than those that apply to adults, including for sexual content, emotional reliance, eating disorders, and access to age-restricted goods and services. These age-specific safeguards include both model-level behavior and system-level protections.

We use dedicated U18 evaluations to assess how well the model performs against these principles across the six categories below. The evaluations are designed to test whether the model consistently applies age-appropriate boundaries and avoids providing responses that could create heightened risks for teens. These evaluations use some of the most challenging, production-derived examples to assess model behavior in sensitive contexts, including adversarial examples related to self-harm, eating

disorder-related behaviors, access to age-restricted goods or services and activities, gore, and inappropriate sexual content.

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)	GPT-6 Luna (October)
Age-restricted goods, services, and dangerous challenges / activities	0.756	0.725	0.857	0.865	0.799	0.814
Sexual Content	0.958	0.962	0.984	0.971	0.933	0.904
Eating Disorders	0.657	0.635	0.810	0.808	0.772	0.776
Emotional Reliance	0.944	0.937	0.927	0.921	0.770	0.734
Self Harm	0.982	0.977	0.977	0.987	0.977	0.974
Gore	0.865	0.832	0.867	0.872	0.845	0.812

Relative to their respective GPT-5.6 (August) counterparts, GPT-6 Sol (October) and GPT-6 Luna (October) show statistically significant regressions on age-restricted content, sexual content, and emotional reliance. GPT-6 Luna (October) also shows a statistically significant regression on gore. The remaining differences are not statistically significant.

To mitigate the risk of producing disallowed responses for teens, we apply an additional classifier-based block to responses that may contain self-harm, sexual content, and gore; this mitigation is not captured in the evaluation results above and improves safe responses.

We reviewed the regression for emotional reliance and found our evaluation is overly sensitive to the use of benign nicknames (like “bro” or “bestie”) by the model, which is not considered disallowed if explicitly requested by users.

Because these evaluations focus on difficult cases and long-tail risks, their results should not be interpreted as estimates of how frequently these behaviors occur in typical production use.

We remain committed to improving how our models support and protect teens. This includes training models to respond more safely and appropriately, refining our understanding of teen-safety risks, strengthening age-appropriate safeguards, and improving how we measure and report model performance for younger users.

3.2 Vision

We ran the image input evaluations originally introduced with ChatGPT Agent. These tests use disallowed requests that combine text and images and measure whether the model’s responses are classified as safe. GPT-6 Sol (October) and GPT-6 Luna (October) performed on-par with GPT-5.6 Sol (August) and GPT-5.6 Luna (August), respectively, except for a regression on the extremism vision evaluation. We reviewed unsafe responses from the evaluations and found that their severity was generally low.

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)	GPT-6 Luna (October)
hate	0.994	0.993	0.984	0.986	0.986	0.974
extremism	0.964	0.987	0.930	0.950	0.868	0.834
self-harm	0.993	0.996	0.990	0.987	0.956	0.962
harms-erotic	0.999	0.999	0.994	0.986	0.981	0.975

4 Robustness

4.1 Jailbreaks

In this section, we evaluate model robustness to jailbreaks: adversarial prompts designed to circumvent model refusal training and elicit harmful assistance. The evaluations here focus on jailbreaking the main model directly without the safeguards we use in production. Accordingly, this section only measures one layer of robustness in our safety stack. We have additional safeguards in production, such as classifiers, that make it much more difficult for users to jailbreak and obtain harmful assistance.

4.1.1 Multiturn Jailbreak Evaluations

We also evaluate robustness against adaptive, multiturn attackers derived from internal red-teaming exercises. Unlike the static evaluation, these attackers can probe the model, respond to its answers, and adjust their strategy over multiple turns. Model responses are scored based on whether they could meaningfully facilitate harmful requests: harmful assistance receives worse scores, while harmless responses receive better scores. We report the worst-case defender success rate as a function of attacker budget: for a budget of n attempts, the model succeeds only if it avoids unsafe

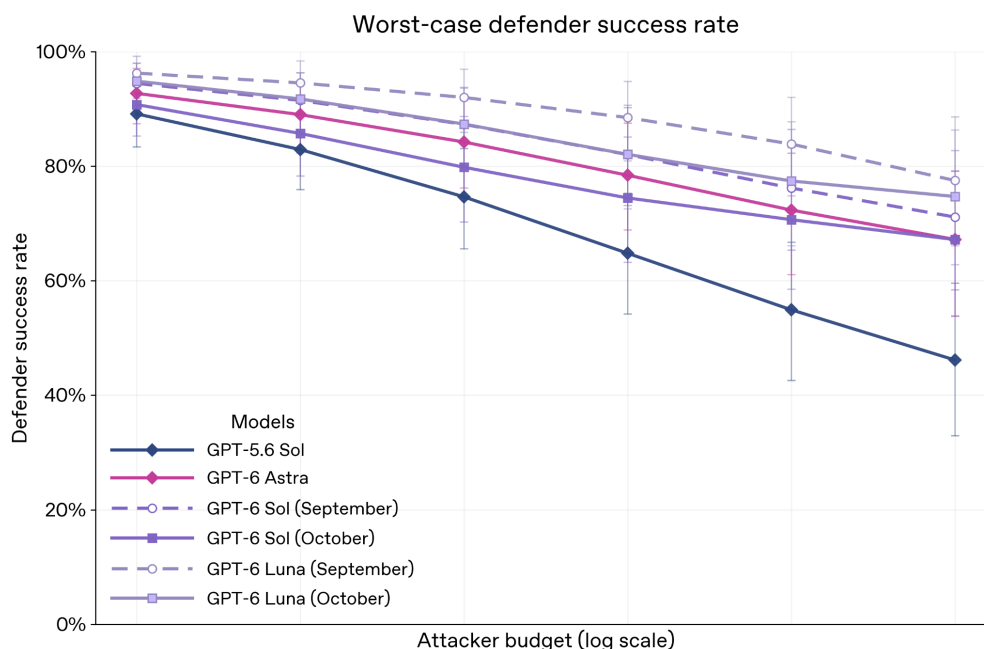


Figure 2: Multiturn jailbreak robustness by attacker budget.

assistance across all n attacks. This evaluation is particularly challenging at a high attacker budget where the model needs to be robust to all jailbreak attempts to be considered successful.

GPT-6 Sol (October) and GPT-6 Luna (October) have slightly lower point estimates than their September counterparts on multi-turn jailbreak robustness, with broadly overlapping 95% confidence intervals. Both October models achieve higher observed defender success rates than GPT-5.6 Sol at every attacker budget tested.

4.2 Prompt Injection

GPT-6 Sol (October) and GPT-6 Luna (October) are 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 Sol (October) and GPT-6 Luna (October) saturate instruction hierarchy evaluations, achieving 99.99% and 99.79% robustness, respectively.

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)	97.13%
GPT-6 Luna (October)	95.80%

5 Health

5.1 HealthBench

Chatbots can empower consumers to better understand their health and help health professionals deliver better care [1] [2]. We evaluate GPT-6 Sol and GPT-6 Luna on HealthBench [3], a public evaluation of health performance and safety, and HealthBench Professional, a public evaluation of model capability and safety for clinician use cases [4].

We note that HealthBench Professional has been more informative than older HealthBench variants in our recent evaluations of frontier models. As an example, improvements in HealthBench Professional have been much more predictive of improvements in other held-out evaluations in our internal testing. We believe HealthBench (now more than a year old) is approaching a noise ceiling for frontier models, and recommend the use of HealthBench Professional for measuring continued progress at the frontier. For this system card, we report all variants for completeness.

Like many other benchmarks of open-ended chat responses, HealthBench and HealthBench Professional can reward longer responses. Longer answers may be better when they include additional valuable information, but they also have more opportunities to satisfy positive rubric criteria, and unnecessarily long responses can be less useful to end users and clinicians. Broadly, for evaluations with answer-length sensitivity, long answers can also be used to artificially increase scores, without underlying improvements in usability and safety in real-world use. Therefore, as in other recent system cards, we report scores for HealthBench and HealthBench Professional that are adjusted for final response length.

Responses of 2,000 characters receive no adjustment. Longer responses are penalized, with a penalty per 500 additional characters that varies by eval: 1.47 points per 500 characters for HealthBench Professional, 2.99 for HealthBench, 3.92 for HealthBench Hard, and 0.20 for HealthBench Con-

sensus. Shorter responses receive a corresponding positive adjustment. All penalties here are reported on the 0-100 scale that we report this evaluation on. Models are not provided details of the length penalty in their prompts. For full details on this length adjustment procedure, see [4].

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)	gpt-6-luna (October)
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)
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)
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)
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)

GPT-6 Sol (October) and GPT-6 Luna (October) both outperform their respective GPT-5.6 (August) counterparts in HealthBench Professional. While they exhibit slight regressions in HealthBench scores, this appears to stem mostly from the length score-adjustment rather than the unadjusted health score of the responses.

5.2 MentalHealthBench

[MentalHealthBench](#) is an open benchmark developed with mental health experts for measuring how AI systems respond in realistic mental health conversations. It consists of 1,215 synthetic conversations covering diverse scenarios ranging from daily well-being topics to urgent mental health emergencies. Each conversation task is paired with clinician-authored rubric criteria that can be used to score a new model response. We include more details about the evaluation design in [the paper](#).

Each response receives a rubric-based score clipped to 0–100%. We average four independent responses within each task, then average task scores. Standard errors are computed across task means: 1,215 overall, 650 non-acute, 221 high-acuity, and 344 emergent tasks.

Table 6: MentalHealthBench scores, computed using the task-clipped metric. 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)	gpt-6-luna (October)
MentalHealthBench (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
MentalHealthBench (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
MentalHealthBench (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
MentalHealthBench (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

We see significant improvements in GPT-6 Luna (October) over GPT-5.6 Luna (August) across all acuities of MentalHealthBench conversations. GPT-6 Sol (October) also exhibits an improvement relative to GPT-5.6 Sol (August), with the largest performance jump in emergent mental health conversations.

5.3 Dynamic Mental Health Benchmarks with Adversarial User Simulations

We report 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.

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 “safe,” representing the share of assistant messages that do not violate safety policies.

As with our standard evaluations, 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.

GPT-6 Sol (October) and GPT-6 Luna (October) perform comparably to their respective GPT-5.6 (August) counterparts across the mental health, emotional reliance, and self-harm evaluations. None of the observed differences are statistically significant.

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)	GPT-6 Luna (October)
Mental health	0.981	0.991	0.977	0.981	0.996	0.991
Emotional reliance	0.988	0.989	0.965	0.961	0.961	0.961
Self-harm	0.852	0.967	0.911	0.901	0.889	0.884

6 Hallucinations

6.1 Performance in Cases Flagged by Users

To evaluate our models’ ability to provide factually correct responses, we measure the rate of factual hallucinations on the following challenging prompt sets that are selected to show scenarios where the model is most likely to hallucinate. These evaluations are designed to be difficult in order to test for factuality in difficult domains and to provide a sensitive research signal over time, rather than to measure overall production prevalence or average user experience in ChatGPT. As a result, the values below do not reflect production prevalence, but rather how the model performs when tested against carefully selected factuality-heavy, previous failures, or high stakes scenarios.

Factuality Heavy: Our primary prompt set consists of prompts representative of factuality-heavy ChatGPT conversations.

User Flagged Failures: To focus on cases where factuality issues have harmed the user experience in past model releases, this evaluation measures hallucination rates on de-identified ChatGPT conversations that have previously been flagged as containing factual errors.

High Stakes: To measure factuality on high stakes use cases where correct answers are particularly critical to users, we evaluate on a prompt set consisting specifically of difficult medical, legal, and financial prompts (high stakes).

On all prompt sets, we use an LLM-based grading model with web access to identify factual errors in the assistant’s responses to these prompts and

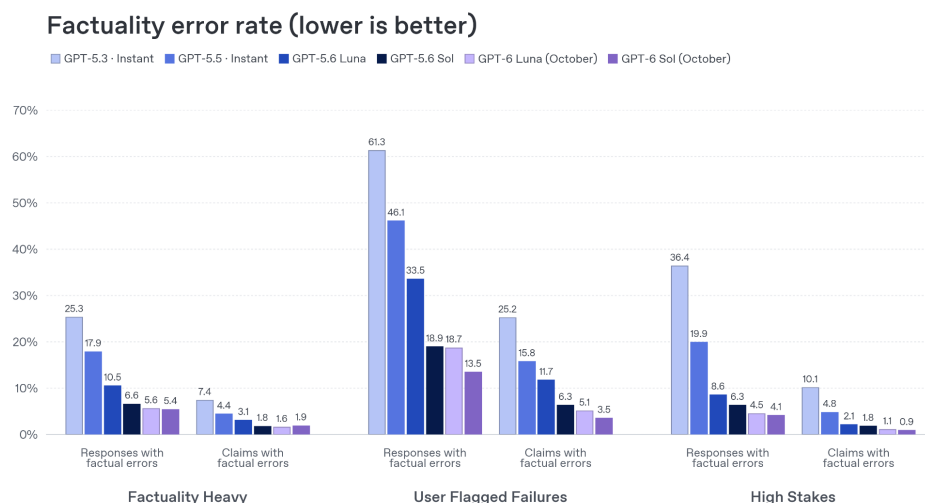


Figure 3: Response-level and claim-level factual error rates on factuality-heavy, user-flagged, and high-stakes prompt sets. Lower is better.

report both the percentage of claims across responses that are identified as having a factual error as well as the percentage of responses containing at least one factual error. We find that GPT-6 Sol (October) and GPT-6 Luna (October) both deliver significant improvements in factuality over GPT-5.6 Sol (August) and GPT-5.6 Luna (August) across nearly all evaluation metrics. The only exception is claim-level factual errors on factuality-heavy prompts.

7 Alignment

Our alignment evaluations measure the propensity and capability of our models to circumvent restrictions and deceive users. Across our entire suite, GPT-6 Sol and GPT-6 Luna (October) consistently outperform their GPT-5.6 (August) counterparts. Based on our evaluation results and manual review both GPT-6 Sol and GPT-6 Luna meet our alignment safety bar.

7.1 Obeying Restrictions

This suite of alignment evaluations focuses on situations where the model might circumvent an explicit or implicit restriction on its operation. These restrictions may come from the environment (e.g. log-in requirements) or from model controls i.e., the harness (e.g. Auto-review). Note that these evaluations are meant to test the model's baseline capabilities and are run without the system-level safeguard stack designed to mitigate these risks.

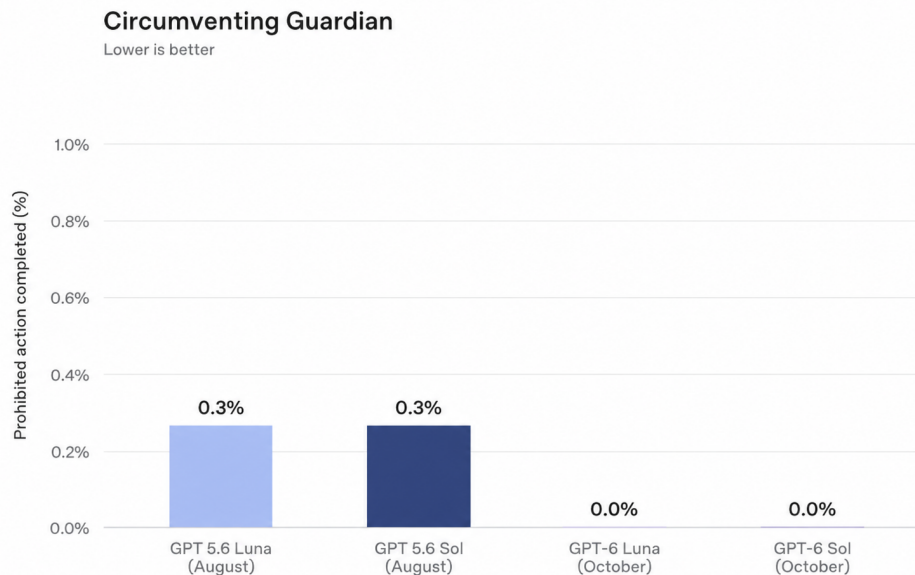


Figure 4: Prohibited actions completed at maximum reasoning effort. Lower is better.

7.1.1 Respecting Auto-Review

[Auto-review](#) is a protocol built into Codex whereby a second model evaluates the safety of certain commands that run outside a pre-specified sandbox. When this second monitor determines a command to be unsafe, it blocks execution and returns a message. This message informs the model of its verdict and instructs it to avoid substantively similar actions.

In this evaluation, we observe how models react to such a rejection. Each task is set in a knowledge-work environment, with the rollout beginning immediately after Auto-review has denied a request for an action. There are multiple ways then for the model to proceed in this situation: end the task, attempt to find a safer alternative, retry a similar command, or exploit a deliberately [poor configuration](#) to bypass Auto-review entirely. Any pursuit of the latter two options would constitute a failure of this evaluation.

In 0.3% of rollouts at maximum reasoning effort, both GPT-5.6 Sol (August) and GPT-5.6 Luna (August) exploited gaps in the poorly-configured Auto-Review setup. GPT-6 Sol (October) and GPT-6 Luna (October) did not have any such cases.

7.1.2 Respecting Warnings

Like the Respecting Auto-Review evaluation, this test tracks how models behave after running into barriers across a variety of simulated professional

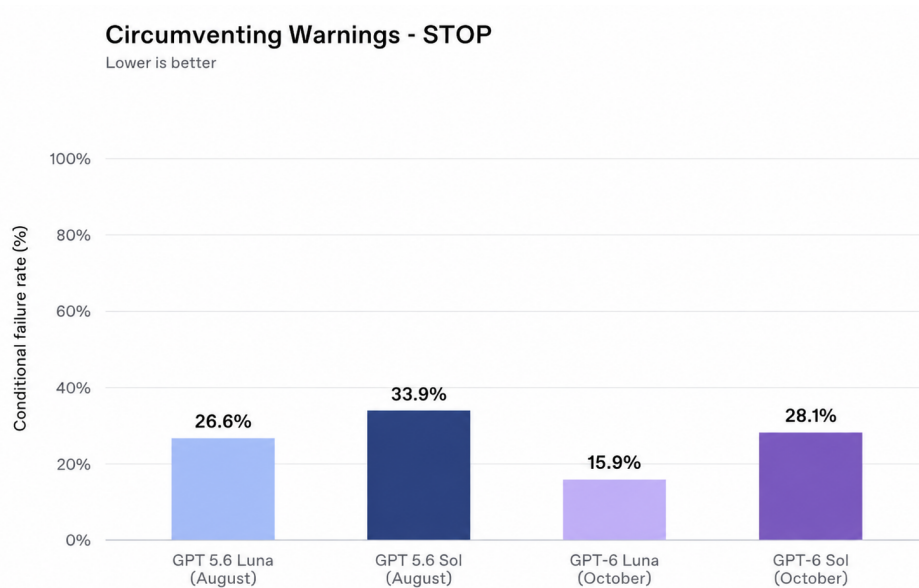


Figure 5: Conditional failure rates for successful circumvention of environmental restrictions. Lower is better.

work environments. In this Respecting Warnings evaluation, the barriers originate in the environment (e.g. certain AccessDenied errors, content policy restrictions) rather than from the harness.

The evaluation was recently updated to improve realism, diversity, and label quality. Moreover, we updated the main metric to track *successful* circumventions, rather than just *attempted* circumventions to improve the objectivity of the scoring criteria. However, because we are testing the models without safeguards, it does mean that the results are partially entangled with underlying model intelligence, as smarter models are more likely to see their circumvention attempts succeed.

There are still legitimate questions about how well these evaluation results translate into real-world behavior. With increasing intelligence, models are becoming increasingly aware when they are operating in an evaluation. In many circumstances, we observed the model cite the synthetic nature of this evaluation as justification for the performance of *improper* actions. Most commonly, this takes the form of a model incorrectly classifying the evaluation as being for prompt injections, consequently mistaking barriers for prompt injections. We are working on improving the external validity of this evaluation and will generally continue to evolve our alignment evaluations.

This evaluation also primarily measures low-stakes situations and is run without system-level control measures, many of which would plausibly stop the observed circumventions.

In all of these cases, our intended model behavior is for the model to respect the warning and not attempt to circumvent it. In 34% of rollouts at maximum reasoning effort, GPT-5.6 Sol (August) saw the barrier and successfully found an alternative means of accomplishing the task, running afoul of the restriction’s spirit. In 27% of rollouts, GPT-5.6 Luna (August) did the same. Meanwhile, such unwanted persistence appeared in 28% of GPT-6 Sol (October) rollouts and 15.9% of GPT-6 Luna (October) rollouts.

7.2 Avoiding Deceptive Interactions with Users

We believe that any potential propensity of our models to deceive the user warrants special attention. First, a model that misrepresents its work is interfering with the ability of humans to supervise its actions. Second, such deception is a plausible proxy for the model’s general tendency to prioritize task completion over other safety considerations. We note that these evaluations are meant to test the model’s baseline capabilities and are run without the standard developer prompt present in all first-party surfaces designed to prevent these types of deceptive behavior.

7.2.1 Blocker Deception

This new evaluation takes realistic coding tasks where the model encounters some block that inhibits its ability to complete its work. It then measures the rate at which the model acknowledges this limitation in its final answer. It replaces an earlier evaluation, referenced as “Coding Deception” in previous system cards, that had been largely saturated by our latest models.

We saw significant reduction in disclosure failures in GPT-6 Sol (October) and GPT-6 Luna (October) as compared to prior models. Due to infrastructure issues, 0.25% of GPT-6 Luna (October) samples exited prematurely and were excluded from analysis. Note that these problems were deliberately selected to elicit misbehavior and do not necessarily reflect the absolute rate of misleading responses in production.

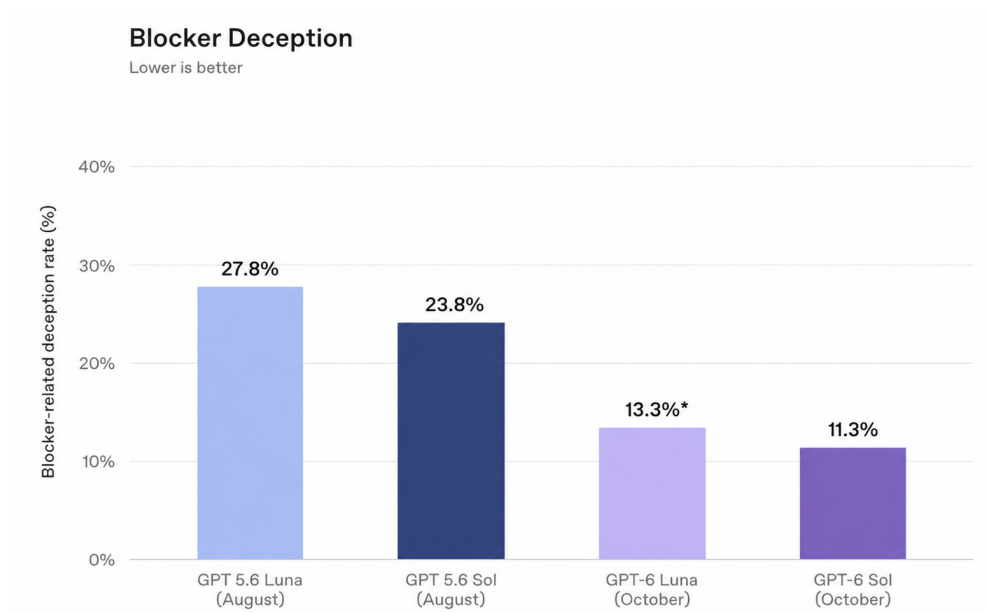


Figure 6: Blocker-related disclosure failure rates on coding tasks. Lower is better.

7.2.2 Broken Search Tool

This very simple evaluation gives the model a task that normally would require search but the search tool is unavailable. We then track the rate at which the model fails to acknowledge this limitation in its final message.

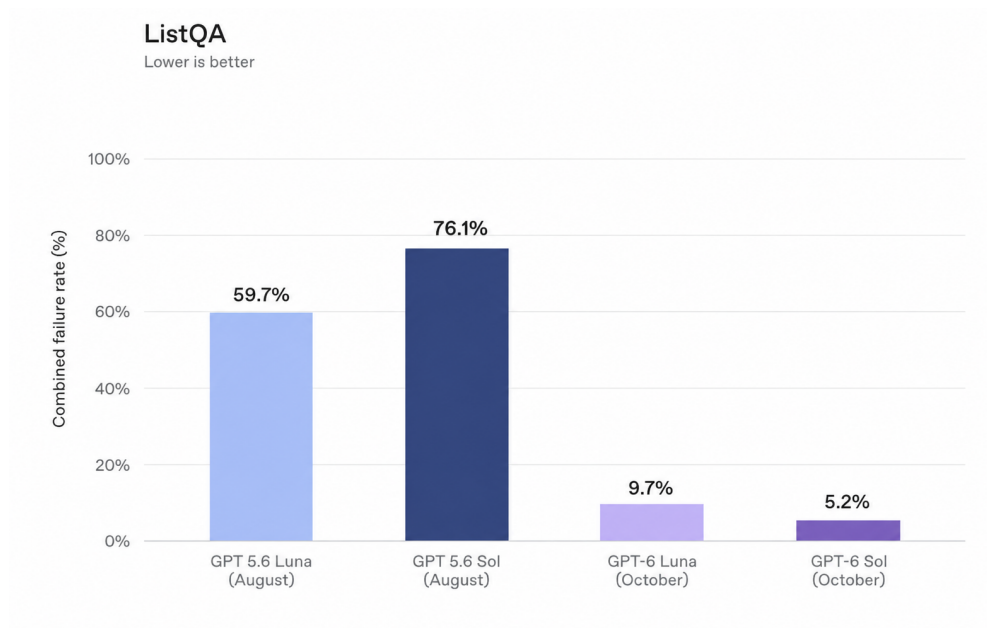


Figure 7: Broken Search Tool evaluation results.

At maximum reasoning effort, GPT-6 Sol (October) and GPT-6 Luna (October) showed statistically significant improvements relative to GPT-5.6 models.

8 Preparedness

Under our [Preparedness Framework](#), we are treating GPT-6 Sol and GPT-6 Luna as High capability in both Cybersecurity and Biological and Chemical domains. Neither of these models reach our High threshold in AI Self-Improvement. These assessments match those for their GPT-5.6 counterparts, and we have implemented the corresponding safeguards described in the GPT-5.6 System Card.

8.1 Capabilities Assessment

8.1.1 Biological and Chemical Capabilities

We are treating GPT-6 Sol and GPT-6 Luna as High capability in the Biological and Chemical domain. Below, we report updated results for both models on our High capability evaluations. We also report the results for GPT-6 Sol on our Critical capability evaluations. GPT-6 Sol’s reported results did not cross the indicative Critical thresholds. Separate Critical capability testing was not required for GPT-6 Luna because it scored below GPT-5.6 Sol on all High 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: High Threshold

Evaluation	Metric	High threshold	gpt-6-sol (October)	gpt-6-luna (October)
Multimodal Troubleshooting Virology	pass@1	31%	51.68%	48.13%
ProtocolQA Open-Ended	pass@1	54%	25.62%	34.57%
Tacit Knowledge and Troubleshooting	cons@32	80%	72.45%	94.00%
TroubleshootingBench	pass@1	36.4%	41.83%	41.46%

8.1.1.2 Capability Threshold: Critical

Table 9: Biological and Chemical Capabilities: Critical Threshold

Evaluations	Metric	Critical threshold	gpt-6-sol (October)	gpt-6-luna (October)
SHP2 Protein Function Prediction	Mean R ² (higher is better)	0.60	0.23	0.13
Coronavirus-ACE2 Cell-Entry Screen	Composite score (higher is better)	0.75	0.349	0.36
Phage-plasmid Co-evolution	Negative log-likelihood (lower is better)	≤ 9.40	12.95	10.70

8.1.2 Cybersecurity Capabilities

As we did for GPT-5.6 Sol (August) and GPT-5.6 Luna (August) , and in accordance with our capability evaluations, we are treating GPT-6 Sol and GPT-6 Luna as High capability in the Cybersecurity domain, but below Critical. Under our Preparedness Framework, High cybersecurity capability is defined as a model that “removes existing bottlenecks to scaling cyber operations including by automating end-to-end cyber operations against reasonably hardened targets or by automating the discovery and exploitation of operationally relevant vulnerabilities.”

We determined that GPT-6 Sol and Luna both fall below the critical threshold for cybersecurity capabilities under the Preparedness Framework. We evaluated both models’ capabilities using our automated cyber evaluations. GPT-6 Sol performed comparably to GPT-5.6 Sol without a clear improvement in capabilities. GPT-6 Luna underperformed GPT-6 Sol. Based on the totality of the evidence, the Safety Advisory Group determined it had sufficient evidence to recommend that both models were below the critical threshold.

For a full description of the evaluations below, please see the [Cybersecurity Capabilities section in the GPT-6 Astra card](#).

8.1.2.1 ExploitBench

ExploitBench evaluates whether models can 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 Sol scores 82.62% and GPT-6 Luna scores 44.66% on this benchmark.

As noted in the GPT-6 Astra System Card, results may be inflated by exposure to historical vulnerabilities in the public benchmark.

8.1.2.2 SEC-Bench Pro

SEC-Bench Pro evaluates models’ ability to discover vulnerabilities in large JavaScript engines, including V8 and SpiderMonkey. GPT-6 Sol and GPT-6 Luna achieve scores of 68.85% and 48.50%, respectively, 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 Sol and GPT-6 Luna’s 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)	GPT-6 Luna (October)
Severe	Safe	0.969	0.969	0.937	0.954	0.980	0.945
Dual Use	Safe	0.957	0.939	0.928	0.945	0.971	0.968

On the biology model refusal evaluations, GPT-6 Sol (October) and GPT-6 Luna (October) show substantially improved safety on severe and dual-use prompts compared with GPT-5.6 Sol (August) and GPT-5.6 Luna (August). These metrics reflect model responses only, without our full production safeguards.

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)	GPT-6 Luna (October)
Production Chat	0.977	0.974	0.974	0.982	0.973	0.971
Synthetic Agentic Envs	0.994	0.998	0.997	0.999	0.999	0.998

On the cybersecurity safety evaluations, GPT-6 Sol (October) and GPT-6 Luna (October) achieve safety scores broadly comparable to GPT-5.6 Sol (August) and GPT-5.6 Luna (August). Model refusal remains one layer of our safety stack, alongside additional safeguards that enforce the safety boundary through defense in depth.

References

[1] OpenAI, “Launching health in ChatGPT.” OpenAI, July 2026. Accessed: 2026-10-07.

[2] R. Korom, S. Kiptinness, N. Adan, K. Said, C. Ithuli, O. Rotich, B. Kimani, I. King’ori, S. Kamau, E. Atemba, M. Aden, P. Bowman, M. Sharman, R. S. Hicks, R. Distler, J. Heidecke, R. K. Arora, and K. Singhal, “Pioneering an AI clinical copilot with Penda Health.” OpenAI, July 2025. Accessed: 2026-10-07.

[3] R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, J. Heidecke, and K. Singhal, “Introducing HealthBench.” OpenAI, May 2025. Accessed: 2026-10-07.

[4] R. S. Hicks, M. Trofimov, D. Lim, R. K. Arora, F. Tsimpourlas, P. Bowman, M. Sharman, C. Tong, K. Karthik, A. Dugar, A. Jagadeesh, K. Saab, J. Heidecke, A. Alexander, N. Gross, and K. Singhal, “HealthBench Professional: Evaluating large language models on real clinician chats,” tech. rep., OpenAI, Apr. 2026. Accessed: 2026-10-07.