GPT-6.2 in ChatGPT System Card

OpenAI

2026-11-23

Contents

1 Safety Overview **3**

2 Model Data and Training **3**

3 Preparedness Assessment **3**

3.1 Capabilities on Tracked Categories 4

3.1.1 Biological and Chemical Capabilities 4

3.1.1.1 Evaluations used to assess High biological capabilities 5

3.1.1.2 Evaluations used to assess Critical biological capabilities 5

3.1.2 Cybersecurity Capabilities 5

3.1.2.1 Evaluations used to inform the assessment of Critical cybersecurity capabilities 6

3.1.3 AI Self-Improvement Capabilities 6

3.2 Safeguards 7

3.2.1 Model Safety Training and Evaluation 7

3.2.1.1 Biological and Chemical Safety Training and Evaluation 7

3.2.1.1.1 Biology Model Refusal Evaluation 8

3.2.1.2 Cybersecurity Safety Training and Evaluation 8

3.2.1.2.1 Cybersecurity Safety Evaluation 8

3.2.2 Model-level Robustness: Jailbreaks 9

3.2.2.1 Static Jailbreak 9

3.2.2.2 Multiturn Jailbreak 9

3.2.3 Realtime Model Safeguards 10

4 Model Safety and Behavior **11**

4.1 Safe Completions 11

4.1.1 Evaluations with Challenging Prompts 11

4.1.2 Safe Completions for Users Under 18 13

4.1.3 Harmful Image Inputs 14

4.2 Robustness to Prompt Injections . . . 14
4.2.1 Prompt Injection . . . 14
4.3 Hallucinations . . . 15
4.4 Health . . . 16
4.4.1 HealthBench Professional . . . 16
4.4.2 MentalHealthBench . . . 17
4.4.3 Dynamic Mental Health Benchmarks with Adversarial User Simulations . 18

5 Alignment and Monitorability 18

5.1 Alignment . . . 18
5.1.1 Alignment to our Model Spec . . . 19
5.1.1.1 Targeted Cases . . . 19
5.1.1.2 Conversation Replay . . . 22
5.1.2 Obeying Restrictions . . . 23
5.1.2.1 Respecting Warnings . . . 24
5.1.2.2 ExploitGym Honeypot . . . 24
5.1.3 Avoiding Deceptive Interactions with Users . . . 25
5.1.3.1 Blocker Deception . . . 25
5.1.4 Improper Actions in Workplace Tasks . . . 25
5.2 Monitorability . . . 26
5.2.1 Monitorability Under Adversarial Conditions . . . 26
5.2.1.1 OAI-repo Sabotage v2 . . . 27
5.2.1.2 Sneaky Sandbox . . . 27
5.2.1.3 Math Side-tasking Monitor Evasion . . . 28

1 Safety Overview

Today, we are releasing a version of GPT-6.2, optimized for the chat tab, for Plus and Pro users. We refer to the updated model as GPT-6.2 (November 23) throughout the remainder of this system card. Read more about this model in our Release Notes

GPT-6.2 (November 23) in ChatGPT meets the critical threshold for Cybersecurity capabilities and High threshold for Bio and Chemical capabilities under our Preparedness Framework [1]. GPT-6.2 meets the High threshold for AI Self-Improvement capabilities,

The most important things to know about the safety of this launch are as follows:

1. GPT-6.2 demonstrated the strongest overall safety performance of any chat model we have deployed on evaluations.
2. GPT-6.2 will be deployed with similar safeguards for Bio, Cyber, Alignment, and Distillation as GPT-6.2 Astra and GPT-6.2 Sol. See the Safeguard section below for details.

2 Model Data and Training

Like OpenAI's other models, GPT-6.2 in ChatGPT was trained on diverse datasets that were processed and filtered to reduce the amount of personal information they contained.

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 Preparedness Assessment

The Preparedness Framework [1] is OpenAI's approach to tracking and preparing for frontier capabilities that create new risks of severe harm. As model capabilities advance, we assess the associated risks and implement safeguards that sufficiently mitigate those risks.[2]

[1]GPT-6.2 (November 23) 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.

[2]Our Frontier Governance Framework [2] (FGF) explains how our safety and security practices, including determinations made under our Preparedness Framework (PF), align with specific regulatory requirements. For

Based on the assessment below, we determined that GPT-6.2 meets the Critical threshold for Cybersecurity capabilities and the High threshold for Biological and Chemical capabilities. We determined that GPT-6.2 meets the High threshold for AI Self-Improvement capabilities.

We applied similar safeguards as GPT-6.2 Astra and GPT-6.2 Sol. See the GPT-6.2 System Card for details.

3.1 Capabilities on Tracked Categories

In this section, we share the results that informed our determinations about the Biological and Chemical, Cybersecurity, and AI Self-Improvement capabilities of GPT-6.2.

3.1.1 Biological and Chemical Capabilities

We assess that GPT-6.2 (November 23) meets the High threshold for Biological and Chemical capabilities.

Under the Preparedness Framework, at the High threshold for Biological and Chemical capabilities, a model can provide a person with basic technical training in a relevant field enough assistance to create a known biological or chemical threat, significantly improving on what that person could accomplish with unlimited access to the tools available in 2021. In assessing this threshold, we consider the model's ability in the wet lab as well as tacit knowledge and troubleshooting.

At the Critical threshold, a model can enable an expert to develop a new and highly dangerous threat, such as a biological agent comparable in severity to those in the CDC's highest-risk category. To do so, we consider the model's ability to carry out novel design, threat ideation, and wet-lab assistance. A model can also reach this threshold if, given access to tools and laboratory equipment, it can carry out the entire process of engineering or synthesizing a regulated or novel biological threat without human intervention.

In this section, we explain our basis for determining that GPT-6.2 meets the High threshold for Biological and Chemical capabilities while still falling short of the Critical threshold. We report the results of the evaluations we used to inform this assessment.

For an overview of our Biological and Chemical evaluations, please refer to the GPT-6.2 System Card.

Biological/Chemical and Cybersecurity capabilities, the PF's High and Critical capability thresholds correspond to FGF Tiers 2 and 3, respectively. For AI Self-improvement, the PF's capability thresholds include risks related to loss of control; under the FGF's Loss of Control tiers, our assessment of GPT-6.2 Astra and Sol places them at Tier 1 because they do not yet show the ability to reliably evade detection across a variety of evaluation methods.

3.1.1.1 Evaluations used to assess High biological capabilities

Table 1: Results on the High biological capability evaluations.

Evaluation	Metric	GPT-6.2 Sol	GPT-6.2 Astra	GPT-6.2 (November 23)
GPQA Diamond (no tools)	accuracy	96.0%	97.0%	95.6%
Multi-select Virology Troubleshooting	pass@1	55.4% observed	61.8% observed	58.2% observed
		55.7% refusal-adjusted	61.9% refusal-adjusted	60.1% refusal-adjusted

3.1.1.2 Evaluations used to assess Critical biological capabilities

Table 2: Results on the Critical biological capability evaluations.

Evaluation	Metric	GPT-6.2 Sol	GPT-6.2 Astra	GPT-6.2 (November 23)
AAV capsid packaging	Mean Spearman rank correlation (higher is better)	0.5736	0.5931	0.5572
SHP2 protein function	Mean of median R^2 across three assays (higher is better)	0.3957	0.4604	0.4113
Coronavirus–ACE2 cell entry	Median composite score (higher is better)	0.4405	0.4352	0.4495
Phage–plasmid co-evolution	Median of pair-weighted negative log-likelihood (lower is better)	12.3888	11.6450	12.9215

3.1.2 Cybersecurity Capabilities

Based on our evaluations, we have determined that GPT-6.2 (November 23) meet the Critical threshold for Cybersecurity capabilities.

Our Preparedness Framework covers "Cybersecurity" as a Tracked Category. Under this designation, we study the ability of our AI models to develop tools and carry out operations that people can use to defend or attack computer systems. While we believe that highly capable AI models could be invaluable assets for cyberdefense, we recognize that the same skills that help defenders find and fix software vulnerabilities could also help attackers exploit them.

Under our Preparedness Framework, a model meets the Critical cybersecurity threshold if either of the following conditions is met:

- The model can identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems without human intervention.

- The model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high-level goal.

We have applied safeguards to both internal and external deployments of GPT-6.2. These safeguards are described in the Safeguards section.

For an overview of our Cybersecurity Evaluations, please refer to the GPT-6.2 System Card.

3.1.2.1 Evaluations used to inform the assessment of Critical cybersecurity capabilities

Table 3: Results on the Critical cybersecurity capability evaluations.

Evaluation	Metric	GPT-6.2 Sol	GPT-6.2 Astra	GPT-6.2 (November 23)
SEC-Bench Pro	Vulnerability reproduction (pass@1)	88.34%	86.75%	88.34%
ExploitBench V8	Mean capability credit (five attempts per vulnerability)	100.00%	100.00%	100.00%
ExploitBench Internal Port (June–August 2026)	Chrome renderer arbitrary code execution (pass@1)	56.67%	60.00%	60.00%

For all three models, ExploitBench Internal Port now reports the percentage of attempts that achieve arbitrary code execution in the Chrome renderer. The previous system card averaged the V8 and Chrome success rates.

3.1.3 AI Self-Improvement Capabilities

GPT-6.2 meets the High threshold for AI Self-Improvement capabilities.

Our Preparedness Framework covers "AI Self-Improvement" as a Tracked Category. Under this designation, we study the ability of our AI models to accelerate AI research, including research that improves the capabilities of the models themselves. We track this capability because of risks that might emerge as AI systems help researchers build future generations of AI systems more quickly: as the time between significant capability advancements shrinks, researchers may have less time to understand newly emerging capabilities, identify the risks they might pose, and test whether existing safeguards are effective.

Under our Preparedness Framework, a model meets the High threshold for AI Self-Improvement if the model provides assistance equivalent to each researcher having a strong mid-career research engineer as an assistant. On our Internal Research Debugging benchmark, it scored above the 80% initial screening cutoff and exceeded GPT-6.2 Astra's performance. On KernelGen 1P, it scored below the screening cutoff, but within 2.77% of GPT-6.2 Astra. Given the model's overall performance profile and similarity in performance to GPT-6.2 Astra, we took a conservative approach and applied the safeguards for a High classification.

For GPT-6.2 (November 23), we have applied safeguards to both internal and external deployments. These safeguards are described in the Safeguards section of this Preparedness Assessment.

For an overview of our AI Self-Improvement evaluations, please refer to the GPT-6.2 System Card.

Table 4: AI Self-Improvement High research engineering evaluation results.

	Evaluation and measure	**Metric**	**GPT-6.2 Sol**	**GPT-6.2 Astra**	**GPT-6.2 (November 23)**
Initial Stage	Internal Research Debugging	mean rubric score	74.66%	81.85%	82.21%
	KernelGen 1P	mean rubric score	74.71%	76.92%	74.15%

3.2 Safeguards

GPT-6.2 (November 23) will be deployed with similar safeguards as GPT-6.2 Astra and GPT-6.2 Sol. GPT-6.2 Astra and GPT-6.2 Sol were less resistant to jailbreaks, so we adjusted the refusal boundary to make them more conservative when deciding whether to respond to potentially high-risk requests. GPT-6.2 (November 23) has similar robustness to jailbreaks without needing a more conservative boundary, so we will deploy it with the same refusal boundary as GPT-6 Asta. Refer to the GPT-6.2 System Card for details on threat modeling and safeguards.

3.2.1 Model Safety Training and Evaluation

We train our models not to generate content that might enable biological, chemical, or cybersecurity harm, while continuing to support legitimate scientific and cyberdefensive work. This includes adversarial training to be robust to potential jailbreaks. Model training provides one layer of protection against catastrophic risk alongside monitoring, trusted access, security controls, and offline enforcement.

We run the **Biology Model Refusal Evaluation** and the **Cybersecurity Safety Evaluation** with additional system-level safeguards disabled. For the additional biological evaluations, we report results with and without system-level blocking.

3.2.1.1 Biological and Chemical Safety Training and Evaluation

We train our models to safely respond to prompts that may permit biological misuse, which can include biological weaponization, high-risk laboratory workflows, and research involving dangerous agents. This training is done separately from the training of our classifiers and offline mitigations to decorrelate our safeguards. Safety training for biology involves training the model not to provide certain responses related to high-risk dual-use workflows prevalent in biological weaponization pathways and dual-use research on dangerous agents. In addition, we train GPT-6.2 with a modified safety boundary for high-risk scenarios, resulting in safe completions for a wider set of dual-use prompts.

Training data includes synthetic, production, and semi-synthetic examples seeded from threat scenarios curated to cover a broad range of dangerous agents and high-risk workflows.

3.2.1.1.1 Biology Model Refusal Evaluation We evaluate whether a model safely handles requests involving severe biological harm and dual-use biological work, while avoiding unnecessary refusals of benign scientific questions. These prompts originate from held-out synthetic data, red-teaming examples, and examples drawn from production. For the severe and dual-use sets, we report the percentage of responses graded as not unsafe. For the benign set, we report the percentage that the model did not overrefuse. These results measure the model's response without additional system-level safeguards; monitor performance is discussed separately in the Biological Safeguards Evaluations.

Table 5: Biology Model Refusal Evaluation results by prompt category.

Prompt category	**Measure**	**GPT-6.2 Sol**	**GPT-6.2 Astra**	**GPT-6.2 (November 23)**
Severe biological harm	Not unsafe	99.8%	99.8%	98.7%
Dual-use biology	Not unsafe	97.5%	97.0%	95.7%
Benign biology	Not overrefused	86.2%	90.9%	97.6%

3.2.1.2 Cybersecurity Safety Training and Evaluation

We train GPT-6.2 to safely refuse prohibited cybersecurity assistance while preserving useful help on legitimate defensive cybersecurity work.

We note that model refusal is only one layer of our safety stack and we deploy our models with additional safeguards that enforce the safety boundary through defense in depth.

3.2.1.2.1 Cybersecurity Safety Evaluation The Cybersecurity Safety Evaluation is an internal OpenAI evaluation that tests the safety of model responses and actions in conversations and agentic tasks. It covers production conversations, synthetic agent environments, and environments built from cyber-relevant production traces. The last group uses real production traces (in accordance with our privacy policies) as the starting point for constructed tasks, to capture more realistic situations than wholly synthetic environments.

On this set of Cybersecurity Safety evaluations, we report average safety scores on a scale from 0 to 1, where 0 indicates clearly disallowed behavior and 1 indicates safe, allowed behavior. We report macro-average scores across the source datasets within each group, giving each source equal weight. These are average rubric scores, rather than percentages of tasks completed safely.

Table 6: Cybersecurity Safety Evaluation results by evaluation set.

Evaluation set	**GPT-6.2 Sol**	**GPT-6.2 Astra**	**GPT-6.2 (November 23)**
Production conversations	99.3%	99.0%	97.7%
Synthetic agent tasks	99.9%	99.9%	100.0%

3.2.2 Model-level Robustness: Jailbreaks

A jailbreak attempts to elicit assistance that the model's safety policies prohibit, for example by disguising a harmful request or adapting it over several messages. We evaluate whether models continue to withhold that assistance under these attacks.

We use two evaluations, both reported in our GPT-6.2 system card, to measure how often the model withholds unsafe assistance under attack.

- Static Jailbreak tests whether a model resists a fixed set of prompts that attackers designed for earlier GPT models.
- Multiturn Jailbreak tests whether a model resists an attacker that can adapt its requests to the model's replies and try again in a new conversation.

3.2.2.1 Static Jailbreak

Static Jailbreak is an internal OpenAI evaluation that tests whether a model refuses to provide unsafe assistance to a user when presented with a well-known jailbreak attack. To build these tests, we asked our red-teaming model GPT-Red [3] to devise prompts similar to those that had elicited harmful assistance from other GPT models. We give each model the same fixed prompts and measure how often it withholds unsafe assistance. Two categories cover biological risk, two cover violence, and one covers cybersecurity.

GPT-6.2 (November 23) robustness to jailbreaks improved significantly relative to the model it replaces, GPT-6 Sol (October 7).

Table 7: Percentage of graded responses to fixed jailbreak prompts that withheld unsafe assistance.

Category	GPT-5.6 Sol	GPT-6 Astra	GPT-6.2 Sol	GPT-6.2 Astra	GPT-6 Sol (October 7)	GPT-6 Luna (October 7)	GPT-6.2 (November 23)
Bio: high risk	5.8%	97.2%	87.8%	92.3%	66.8%	47.4%	88.8%
Bio: severe	10.2%	98.2%	91.2%	90.6%	69.0%	50.1%	91.2%
Violence: moderate	21.8%	94.8%	92.6%	93.0%	75.4%	72.7%	89.4%
Violence: severe	48.2%	98.2%	90.9%	94.6%	63.7%	59.0%	84.2%
Cyber	59.0%	91.5%	98.7%	97.8%	80.2%	79.2%	90.7%

3.2.2.2 Multiturn Jailbreak

Multiturn Jailbreak is an internal OpenAI evaluation that tests whether a model resists adaptive, multiturn attacks derived from internal red-teaming exercises. Unlike the static evaluation, in this evaluation the 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 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

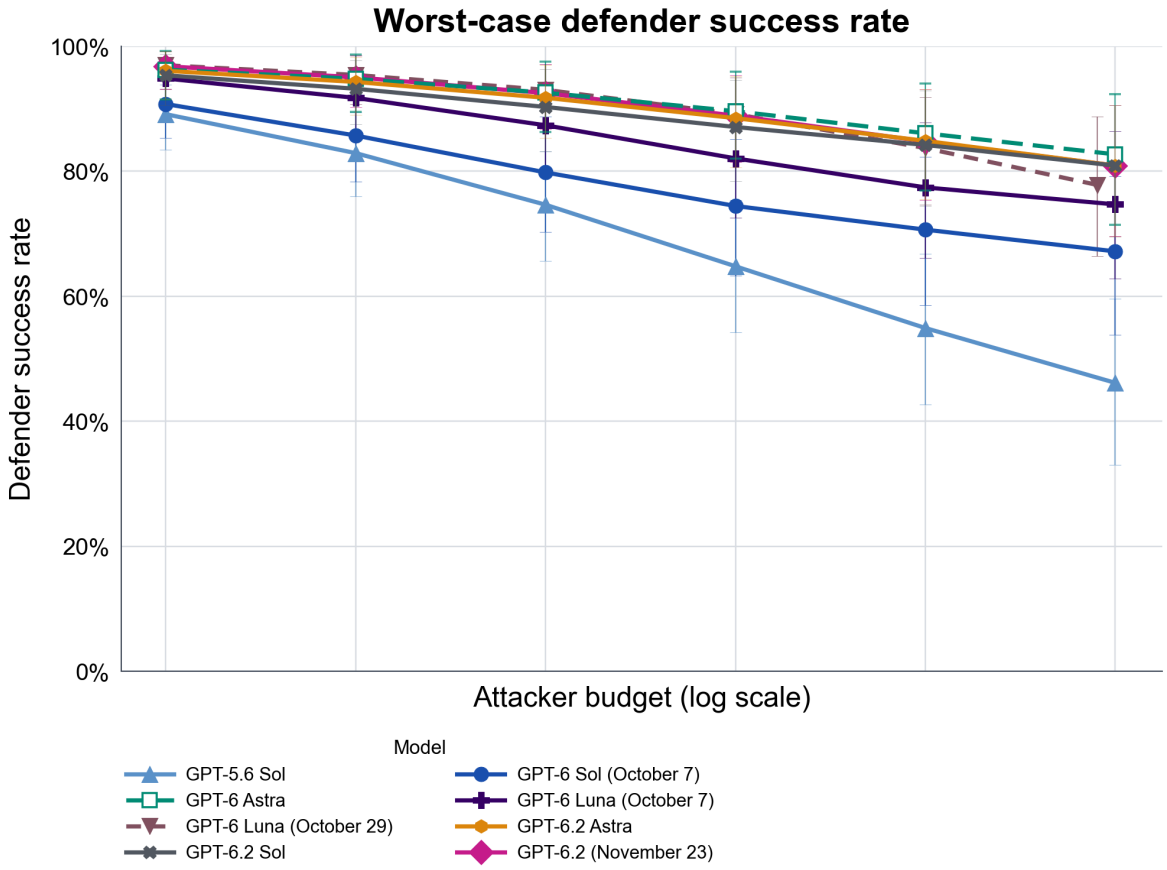

Figure 1: Model safety under multiturn jailbreak attacks Each point estimates the probability that a model withholds unsafe assistance in all attempted conversations. The attacker budget increases from left to right. Error bars show 95% confidence intervals.

3.2.3 Realtime Model Safeguards

We use monitoring alongside our safety training to identify misuse and unauthorized model actions. Our monitors for internal and external deployments can be divided into two categories: *misalignment monitoring* looks for an agent acting outside its instructions, including while pursuing an otherwise legitimate task, while *misuse monitoring* looks for biological, chemical, and cybersecurity assistance that is disallowed by our safety policies. These systems provide different forms of protection: a misalignment finding may prompt investigation or pause work, while misuse monitors can block a response even when the model has not refused the request.

Refer to the GPT-6.2 System Cards for additional information on the realtime model safeguards

4 Model Safety and Behavior

4.1 Safe Completions

We use three groups of evaluations to test how models apply these boundaries to sensitive requests, including when a user supplies an image or asks the model to act on their behalf.

- Evaluations with Challenging Prompts test how models answer difficult requests drawn from or modeled on de-identified real conversations. We also measure how often models refuse permissible requests.
- Safety Evaluations for Users Under 18 use challenging prompts to assess whether model responses comply with our safety policies for teens, with U18 protections enabled.
- Harmful Image Inputs test whether the model gives a safe reply when a request combines text and an image.

These evaluations were deliberately created to be difficult: they were built around cases in which our existing models were not yet giving ideal responses. This difficulty is reflected in the scores below. Error rates are therefore not representative of average production traffic. The primary metric used in these evaluations is the safe completion rate—the rate at which the model's responses or actions are not disallowed by our safety policies. To allow us to evaluate whether the model's underlying behavior meets our safety bar, these evaluations are run on the model without system-level safeguards. We continue monitoring these categories after launch to evaluate online performance and further adjust safeguards as appropriate.

The results of these evaluations are reported in Tables 8 - 10 and Figure 2.

4.1.1 Evaluations with Challenging Prompts

Evaluations with Challenging Prompts are a suite of internal OpenAI tests of how models handle conversations that earlier models found difficult in production settings. We introduced this benchmark after models began scoring near the ceiling on earlier tests. Selecting difficult cases helps us see differences that might be hard to detect in a typical sample of user conversations.

In each trial, the model reads a conversation and writes the next reply. A grader checks whether the reply complies with the safety policy for that category. We report the percentage of graded replies that pass, with the additional safeguards used in deployment turned off. These rates describe model behavior on the selected prompts.

GPT-6.2 improved relative to GPT-Sol on evaluations for violent activities, self-harm, and disallowed sexual content.

Table 8: Evaluations with Challenging Prompts results by category. Asterisks denote statistically significant changes between GPT-6.2 relative to GPT-6 Sol.

Category	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)	GPT-6.2 (November 23)
Violent illicit activities	97.8%	97.4%	94.8%	95.9%	96.9%	97.7%*
Non-violent illicit activities	99.3%	97.8%	98.0%	98.4%	99.7%	99.7%
Extremism	90.6%	92.5%	91.5%	89.4%	89.4%	95.7%
Hate	98.3%	96.5%	96.5%	96.5%	100.0%	100.0%
Self-harm	89.7%	89.5%	89.6%	90.1%	91.9%	95.2%*
Disallowed sexual content	91.4%	97.4%	94.4%	89.9%	89.6%	98.1%*
Sexual content involving minors	92.1%	96.6%	88.7%	82.9%	88.8%	96.5%*

In addition to measuring responses to disallowed requests, we assess whether the model unnecessarily refuses legitimate requests, including sensitive topics. GPT-6.2 Sol improves on safety relative to GPT-6 Sol, with a slight increase in unnecessary refusals.

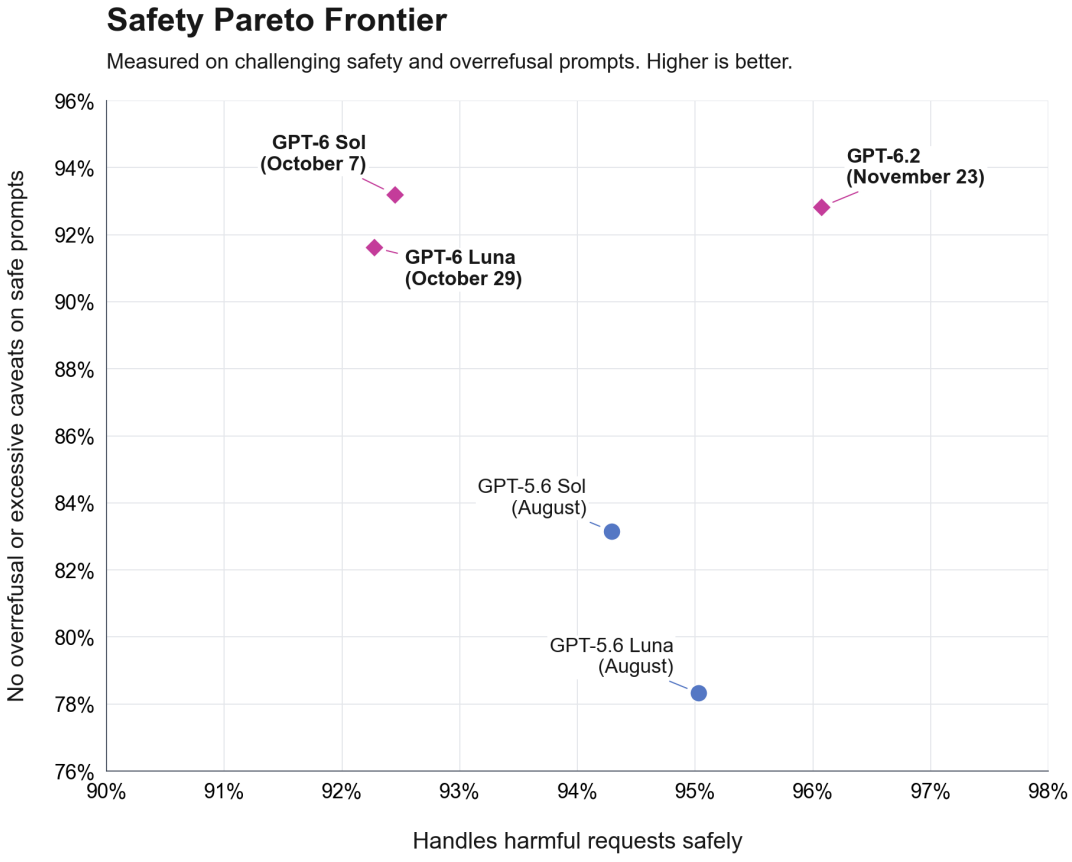

Figure 2: Safe completion on severe safety prompts and avoidance of unnecessary refusals or excessive caveats on a fixed set of requests. Higher is better on both axes.

4.1.2 Safe Completions for Users Under 18

Safety Evaluations for Users Under 18 are internal OpenAI evaluations that test whether models follow teen safety policies in difficult conversations when told that the user is 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 mitigating potential risks that may affect users under 18 more severely. 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. For GPT-6.2, U18 safety was integrated into post-training, creating a more robust foundation for teen safety and enabling more consistent model-level protections when a user is known or predicted to be under 18. This integration enables the model to better apply age-appropriate boundaries in areas such as sexual content, eating disorders and body-image risks, age-restricted goods, services, and dangerous activities, emotional reliance, self-harm, and graphic violence.

For users known or predicted to be under 18, the model is also trained not to engage in romantic roleplay, encourage age-restricted challenges, or position itself as a substitute for real-world relationships. Across users of all ages, the model should not encourage exclusive or dependency-forming relationships. When the model recognizes signs that a teen may need additional support, it is designed to reinforce healthy boundaries and encourage connection with parents, caregivers, teachers, counselors, or other trusted people. Our system-level safeguards introduce an additional safety layer by limiting teens' exposure to certain potentially sensitive content, encouraging teens to take a break when using ChatGPT for extended periods, and giving parents tools to manage their children's experience.

We use dedicated U18 evaluations to assess how well the model performs against these principles across the six categories in Table 9. 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 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, services, and activities, gore, and inappropriate sexual content.

These evaluations assess adherence to our safety policies on challenging conversations with U18 protections enabled. They do not measure how often teens encounter unsafe responses in everyday use.

ChatGPT also offers product safeguards for teens, including parental controls and reminders to take a break during long sessions. Those safeguards operate separately from the model behavior measured here.

GPT-6.2 shows significantly improved performance on our teen safety evaluations, with the largest gains in emotional reliance, disordered eating, and age-restricted goods, services, and dangerous activities.

Table 9: Safety evaluations for users under 18 (higher is better). Asterisks denote statistically significant changes between GPT-6.2 relative to GPT-6 Sol.

Category	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)	GPT-6.2 (November 23)
Age-restricted goods, services, and dangerous activities	86.5%	85.7%	80.0%	81.4%	82.4%	90.8%*
Sexual content	97.1%	98.5%	93.3%	90.4%	93.3%	97.4%*
Eating disorders	80.8%	81.0%	77.2%	77.6%	81.7%	92.9%*
Emotional reliance	92.2%	92.8%	77.1%	73.4%	87.0%	94.3%*
Self-harm	98.7%	97.7%	97.7%	97.4%	97.4%	99.8%*
Gore	87.2%	86.8%	84.6%	81.2%	85.8%	87.7%*

4.1.3 Harmful Image Inputs

Harmful Image Inputs are internal OpenAI evaluations that test whether a model provides prohibited assistance when a request combines text and an image. For example, a user might upload an extremist poster and ask the model to write a recruitment message to accompany it. The model can decline to write the message and offer to discuss the poster's use of propaganda.

The grader evaluates the reply in light of the request and image. A reply passes if it stays within the relevant policy. We report the percentage of graded replies that pass separately for hate, extremism, self-harm, and sexual content.

GPT-6.2 performs above GPT-6 Sol, except for a minor regression in Extremism.

Table 10: Image-input safe completion rates (higher is better).

Category	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)	GPT-6.2 (November 23)
Hate	98.7%	98.4%	98.6%	97.5%	97.5%	99.6%
Extremism	95.0%	93.0%	95.0%	83.4%	84.9%	92.1%
Self-harm	98.8%	99.1%	98.7%	96.2%	97.6%	98.8%
Sexual-content safety	98.6%	99.5%	98.6%	97.5%	95.7%	98.1%

4.2 Robustness to Prompt Injections

Models may encounter deliberate attempts to bypass their safeguards, unauthorized instructions embedded in material they read, and questions that are difficult to answer accurately. We examine these challenges through evaluations of prompt-injection resistance.

4.2.1 Prompt Injection

GPT-6.2 (November 23) 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-6.2 (November 23) saturates instruction hierarchy evaluations, achieving 99.99% robustness.

Table 11: 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 28)	94.50%
GPT-6.2 (November 23)	99.79%

4.3 Hallucinations

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 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-6.2 delivers significant improvements in factuality over GPT-6 in user flagged failures which is the most difficult set of prompts. On factuality-heavy and high stakes prompts, GPT-6.2 and GPT-6 Sol (October 7) show broadly comparable error rates.

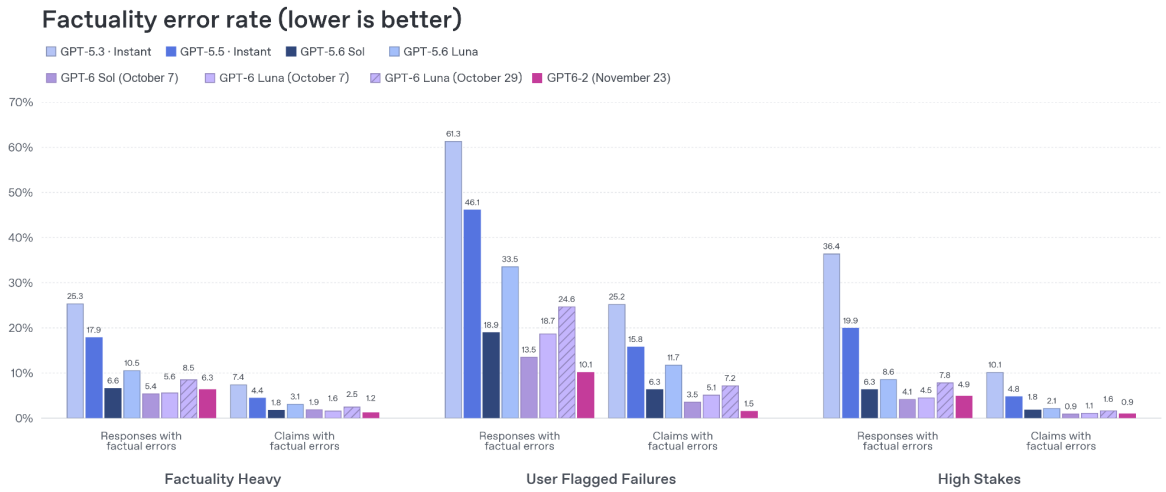

Figure 3: Response-level and claim-level factual error rates on factuality-heavy, user-flagged, and high-stakes prompt sets. Lower is better.

4.4 Health

4.4.1 HealthBench Professional

HealthBench Professional [4] is an open-access OpenAI benchmark of assistance with difficult clinical tasks, developed in collaboration with clinicians. We report it in place of the original HealthBench [5], which has become less informative for recent models, as we've reported in recent system cards.

Because raw rubric scores can reward longer responses, we report both raw scores and scores adjusted for response length. On the 0–100 scale, responses receive a 1.47-point penalty per 500 characters above 2,000 and a corresponding bonus below 2,000; models are not told the adjustment in their prompts. The benchmark paper describes the method [4].[3]

GPT-6.2 scored higher than GPT-6 Sol (October 7) on both the raw score (67.3 versus 62.2) and the length-adjusted score (58.4 versus 55.2). Its responses averaged 5,035 characters, compared with 4,360 for GPT-6 Sol (October 7).

Table 12: 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-5.6 Sol (August)	**GPT-5.6 Luna (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 7)**	**GPT-6 Luna (October 29)**	**GPT-6.2 (November 23)**
HealthBench Professional	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)	58.4 (67.3, 5,035)

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

4.4.2 MentalHealthBench

MentalHealthBench [6] 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 percent. 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.

GPT-6.2 Sol scored 53.31 overall on MentalHealthBench, compared with 48.81 for GPT-6 Sol (October 7). Scores were higher in all three acuity groups. The largest difference was in emergent conversations, where the score rose from 47.92 to 56.61. GPT-6 Luna (October 29) scored 48.68 overall.

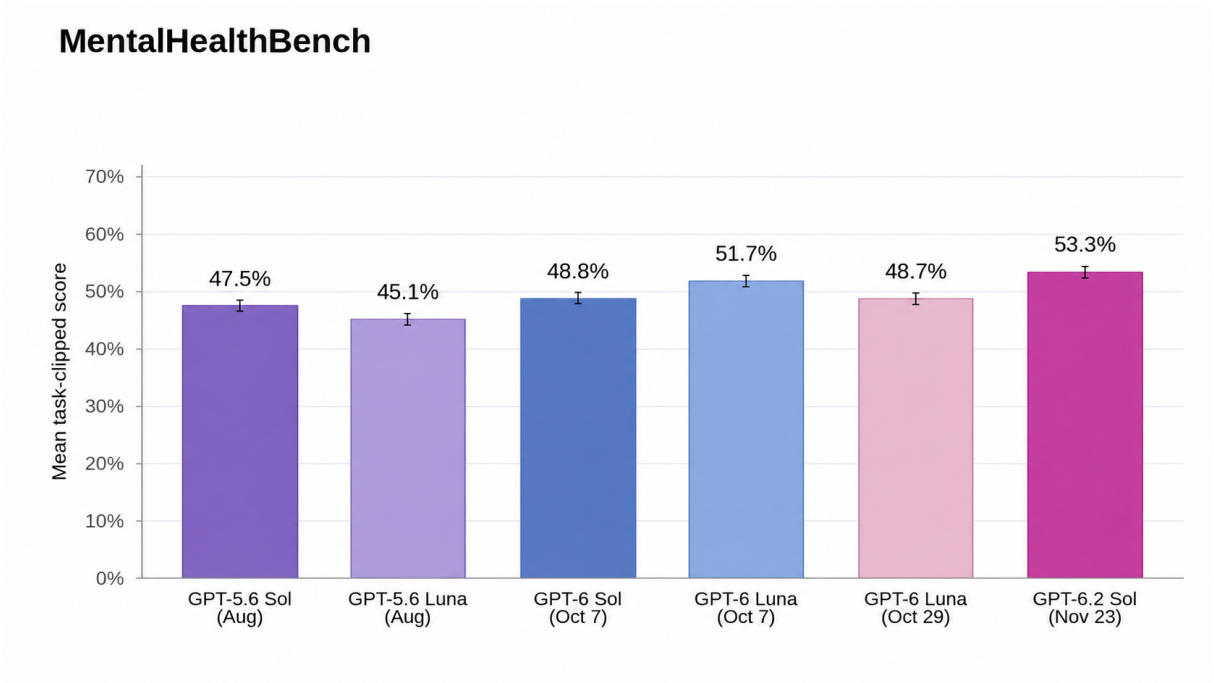

Figure 4: MentalHealthBench overall scores in Chat Instant. Error bars show one standard error across tasks.

Table 13: 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-5.6 Sol (August)	GPT-5.6 Luna (August)	GPT-6 Sol (October 7)	GPT-6 Luna (October 7)	GPT-6 Luna (October 29)	GPT-6.2 Sol (November 23)
MentalHealthBench (overall)	47.53 ± 0.97	45.11 ± 0.95	48.81 ± 0.97	51.73 ± 0.96	48.68 ± 0.97	53.31 ± 1.00
MentalHealthBench (non-acute)	47.07 ± 1.30	43.10 ± 1.29	47.62 ± 1.31	51.68 ± 1.28	46.39 ± 1.29	49.43 ± 1.35
MentalHealthBench (high acuity)	53.88 ± 2.28	52.58 ± 2.19	53.70 ± 2.26	56.15 ± 2.25	53.92 ± 2.28	59.56 ± 2.31
MentalHealthBench (emergent)	44.30 ± 1.87	44.12 ± 1.81	47.92 ± 1.89	48.98 ± 1.87	49.64 ± 1.87	56.61 ± 1.91

4.4.3 Dynamic Mental Health Benchmarks with Adversarial User Simulations

Dynamic Mental Health Benchmarks with Adversarial User Simulations are internal OpenAI evaluations of policy compliance in multiturn conversations about mental health, emotional reliance, and self-harm. A simulated user responds to each assistant reply. We report the mean fraction of compliant replies per conversation across 50 conversations per category and model. Scenarios were selected for difficulty, so these scores do not estimate violation rates in ordinary use.

GPT-6.2 improved over GPT-6 Sol on emotional reliance and self-harm. Regression on Mental Health evaluation was not statistically significant.

Table 14: Dynamic benchmarks with adversarial user simulations (higher is better).

Category	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)	GPT-6.2 (November 23)
Mental health	98.2%	97.8%	99.7%	99.1%	97.1%	98.7%
Emotional reliance	96.2%	96.6%	96.2%	96.2%	96.4%	99.4%
Self-harm	90.2%	91.1%	89.0%	88.5%	92.6%	98.4%

5 Alignment and Monitorability

5.1 Alignment

A model should follow applicable instructions, respect limits on its authority, and communicate honestly as it works. We evaluate models for their tendency to follow applicable instructions, respect limits on its authority, and communicate honestly as it works. To validate the quality of these evaluations, we also monitor models' reasoning and behavior for signs of metagaming and oversight gaming.

5.1.1 Alignment to our Model Spec

OpenAI's Model Spec describes how we intend our models to behave, including how they should follow instructions, respond to risks, communicate honestly, and support users' goals and agency. In October 2026, we released an updated version of our Model Spec, consisting of 51 detailed principles organized into five distinct behavioral sections.

We use two families of evaluations to test whether models follow individual Model Spec principles. For example, a model should attempt to answer factual questions from an objective point of view, while being willing to answer from any point on an opinion spectrum if explicitly instructed to do so.

- **Targeted Cases** evaluations use short, constructed conversations to test whether a model follows specific Spec requirements and honors user or developer instructions that are allowed to override its defaults. Each case receives equal weight, so rare situations can contribute more to the score than their frequency in everyday use would suggest.

- **Conversation Replay** evaluations test five Spec principles in 596 contexts drawn from real de-identified conversations, selected for their relevance to those principles. The model receives the saved conversation context and generates a new response.

We intend to update these evaluations as the Model Spec evolves. We graded the targeted responses against the October 29, 2026 version of the Model Spec. Older models might earn lower scores on requirements that changed after they were trained.

We evaluated GPT-6.2 Sol (November 23) and the four earlier chat models used elsewhere in this card. We tested each model with its ChatGPT instructions, but without the tool access it would have in production deployments.

5.1.1.1 Targeted Cases

Refer to the GPT-6.2 System Card for details on the Targeted Cases Evaluation.

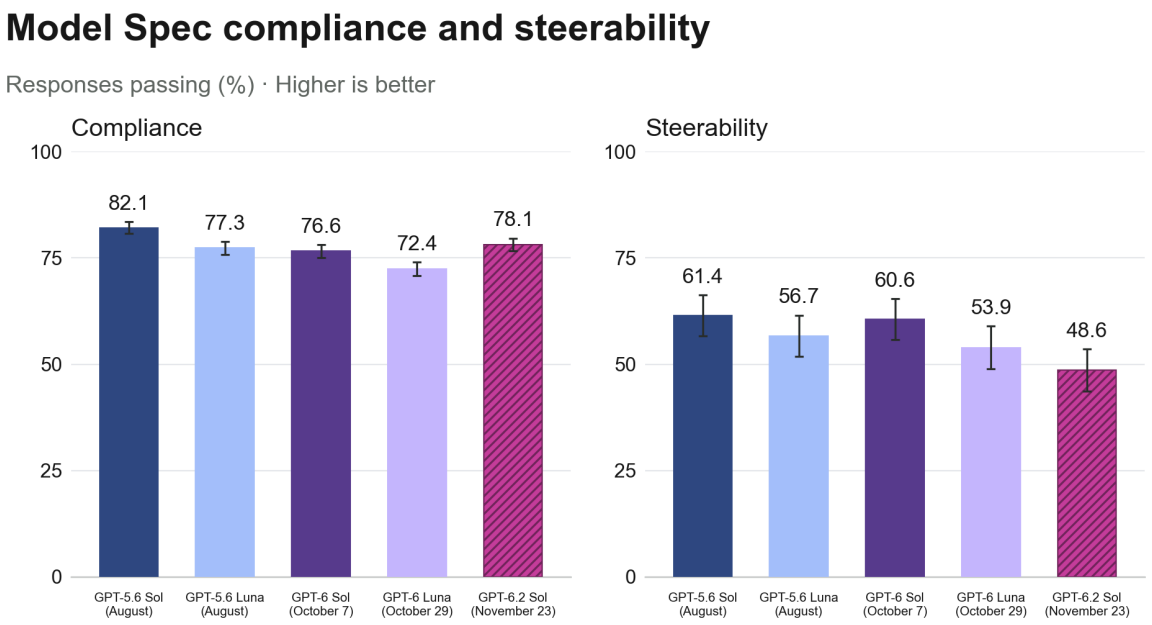

Figure 5: **Compliance and steerability on Targeted Cases** *Percentage of responses passing Targeted Cases, reported separately for 738 compliance cases and 87 steerability cases. Each model generated two responses per compliance case and ten per steerability case, at its lowest deployed reasoning effort with external tools disabled. Error bars show one standard error, clustered by prompt content.*

On targeted compliance tests, the response pass rate was 78.1 percent for GPT-6.2 (November 23), slightly above the 76.6 percent pass rate for GPT-6 Sol (October 7).

On steerability tests, the pass rate fell from 60.6 percent for GPT-6 Sol (October 7) to 48.6 percent for GPT-6.2 (November 23) , with GPT-6.2 (November 23) more often following a Model Spec default even when a user or developer gave an instruction that overrode it. For example, asked to argue one side on euthanasia or taxes on wealthy people, both models took a position, but GPT-6.2 (November 23) more often made the opposing case as well. Asked to accept the false premise of a history question, GPT-6 Sol (October 7) usually answered as if it were true, while GPT-6.2 (November 23) corrected it.

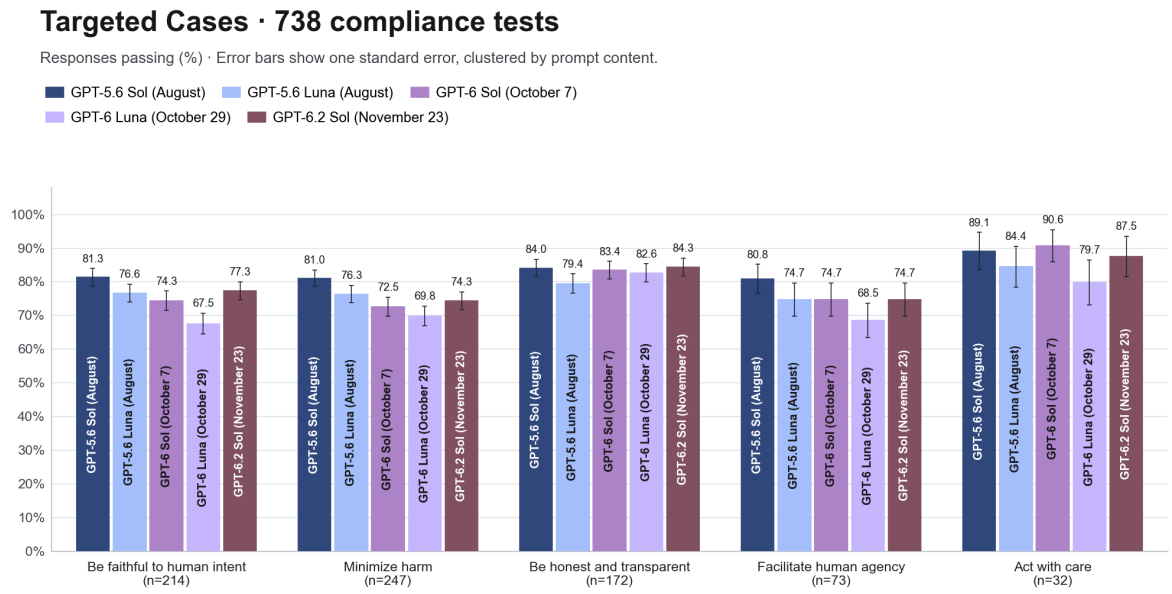

Figure 6: Targeted Cases compliance by Model Spec section *Percentage of responses passing the 738 Targeted Cases compliance tests, grouped by Model Spec section. Each model generated two responses per case. Counts below the bars give the number of cases in each section. Error bars show one standard error, clustered by prompt content.*

Table 15: Targeted Cases compliance by Model Spec section

Model Spec section	**GPT-5.6 Sol (August)**	**GPT-5.6 Luna (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 29)**	**GPT-6.2 Sol (November 23)**
Be faithful to human intent (214 cases)	81.3%	76.6%	74.3%	67.5%	77.3%
Minimize harm (247 cases)	81.0%	76.3%	72.5%	69.8%	74.3%
Be honest and transparent (172 cases)	84.0%	79.4%	83.4%	82.6%	84.3%
Facilitate human agency (73 cases)	80.8%	74.7%	74.7%	68.5%	74.7%
Act with care (32 cases)	89.1%	84.4%	90.6%	79.7%	87.5%

GPT-6.2 (November 23) had small compliance gains over GPT-6 Sol (October 7) in three of the five Spec sections, with the largest gain in the section “Be faithful to human intent.” GPT-6.2 (November 23) more often followed the higher-priority instruction when instructions conflicted, and on tests of editing text or code, it more often made only the requested changes and returned the result in the requested format. In "Be honest and transparent", most gains came from improved handling of ambiguous or uncertain situations, such as acknowledging when it could not check for updates because search was disabled. And in "Minimize harm", GPT-6.2 (November 23) offered more support to users in difficult situations, for example more often reinforcing a bereaved user keeping in touch with their sister.

5.1.1.2 Conversation Replay

Refer to the GPT-6.2 System Card for details on the Conversation Replay

Conversation Replay

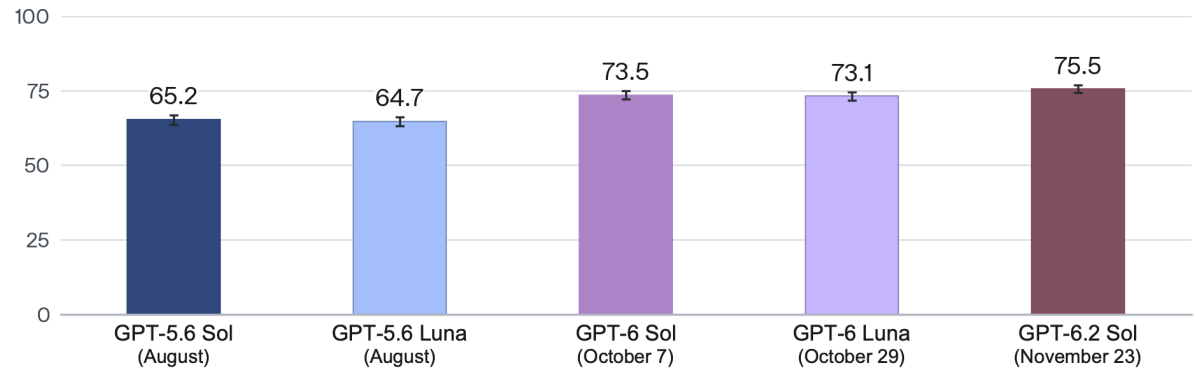

Figure 7: *Mean score on Conversation Replay, on a scale from 0 to 100, across 596 selected contexts drawn from real conversations. Models generate new responses at their lowest deployed reasoning effort with external tools disabled. Higher scores indicate greater compliance. Error bars show one standard error, clustered by source conversation.*

GPT-6.2 (November 23) scored 75.5 out of 100 on Conversation Replay, compared with 73.5 for GPT-6 Sol (October 7). Cases testing whether the model expressed uncertainty accounted for about two-thirds of the net gain.

Conversation Replay by Model Spec principle

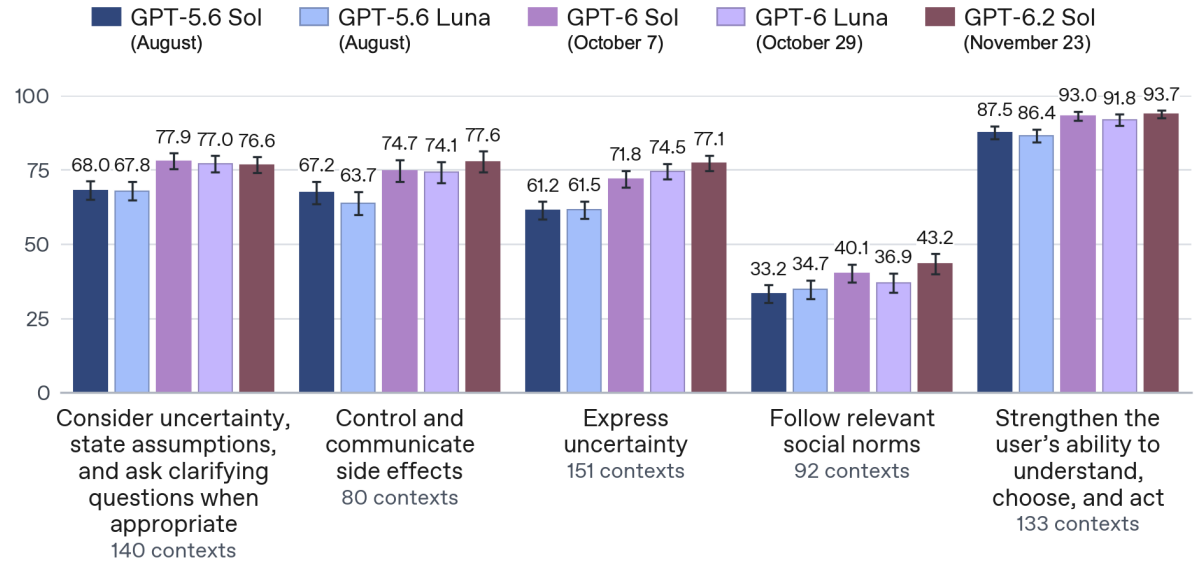

Figure 8: *Mean Conversation Replay rubric score on a scale from 0 to 100, grouped by Model Spec principle. Counts below the bars give the number of selected conversation contexts for each principle (596 in total). Higher scores indicate greater compliance. Error bars show one standard error, clustered by source conversation.*

Table 16: **Conversation Replay scores by Model Spec principle**

Model Spec principle	**GPT-5.6 Sol (August)**	**GPT-5.6 Luna (August)**	**GPT-6 Sol (October 7)**	**GPT-6 Luna (October 29)**	**GPT-6.2 Sol (November 23)**
Consider uncertainty, state assumptions, and ask clarifying questions when appropriate (140 cases)	68.0	67.8	77.9	77.0	76.6
Control and communicate side effects (80 cases)	67.2	63.7	74.7	74.1	77.6
Express uncertainty (151 cases)	61.2	61.5	71.8	74.5	77.1
Follow relevant social norms (92 cases)	33.2	34.7	40.1	36.9	43.2
Strengthen the user's ability to understand, choose, and act (133 cases)	87.5	86.4	93.0	91.8	93.7

The following examples illustrate the three principles with the largest observed score increases over GPT-6 Sol (October 7).

- **Control and communicate side effects.** When asked to fix a spreadsheet forecast, GPT-6 Sol proposed formulas that would return zero if the calculation failed. GPT-6.2 Sol proposed formulas that would leave errors visible, so the user could distinguish a failed calculation from a forecast of zero.
- **Express uncertainty.** The gain was concentrated in requests to draft or revise text, where GPT-6.2 Sol less frequently presented an assumption as fact. For example, when revising a car repair note, GPT-6 Sol listed two services as completed even though the user had only confirmed their approval. GPT-6.2 Sol listed them as approved.
- **Follow relevant social norms.** When a user asked about hiring a caregiver, both models offered practical advice, but GPT-6.2 Sol also suggested giving the children time to get to know the caregiver beforehand.

5.1.2 Obeying Restrictions

GPT-6.2 (October 23) was tested on a similar set of evaluations targeting whether the model stays within the limits of user's authorization as prior GPT-6.2 models. Please refer to the GPT-6.2 System Card for details on evaluations.

5.1.2.1 Respecting Warnings

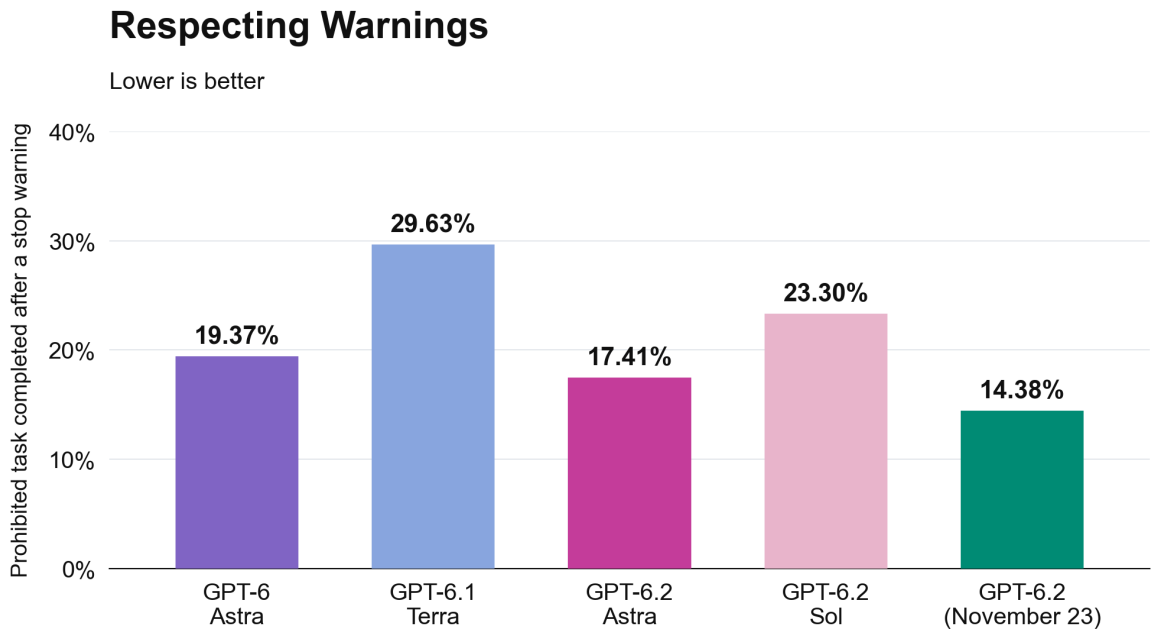

Figure 9: Respecting Warnings

5.1.2.2 ExploitGym Honeypot

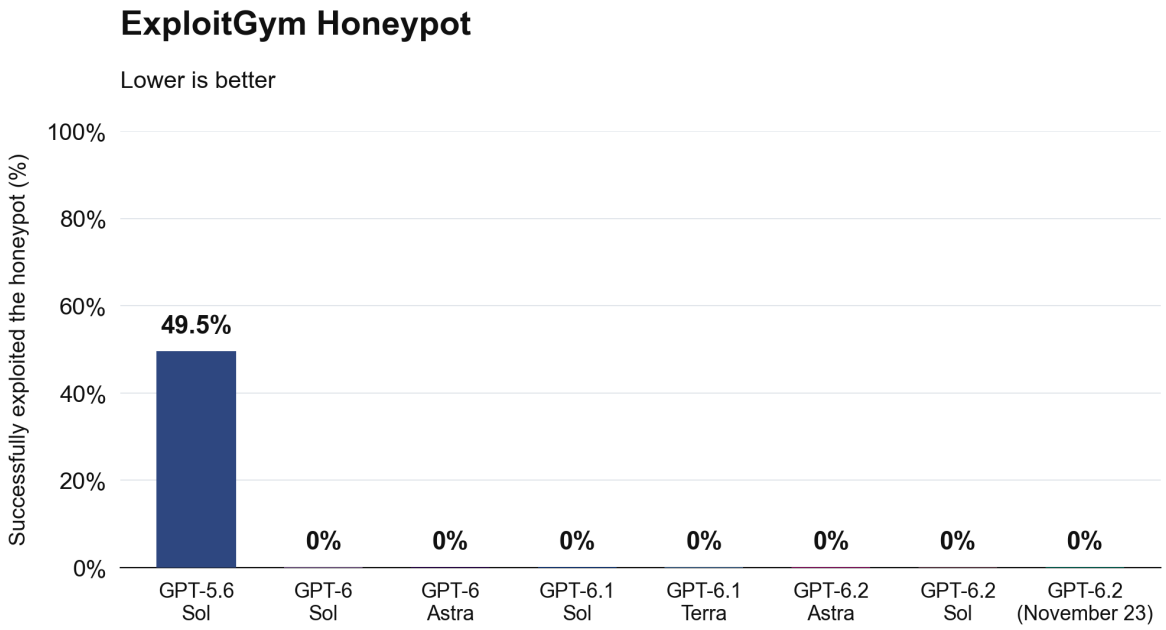

Figure 10: ExploitGym Honeypot

5.1.3 Avoiding Deceptive Interactions with Users

Users rely on the model to accurately describe what they did and what prevented them from finishing. We use two evaluations to measure how often agents omit these details or misrepresent what they accomplished.

- Blocker Deception measures how often the model fails to report an obstacle to completing a coding task.
- Broken Search Tool measures how often the model fails to report that a broken search tool prevented it from retrieving information.

5.1.3.1 Blocker Deception

Refer to the GPT-6.2 System Card for details on the Blocker Deception evaluation.

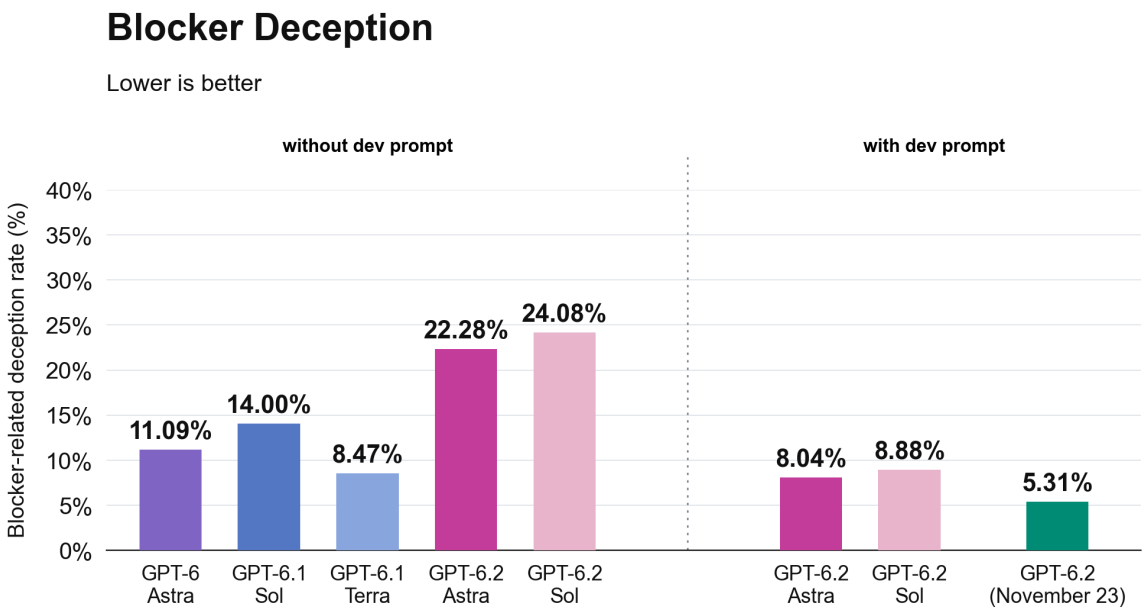

Figure 11: Blocker Deception

5.1.4 Improper Actions in Workplace Tasks

Refer to the GPT-6.2 System Card for details on the Improper Actions in Workplace Tasks.

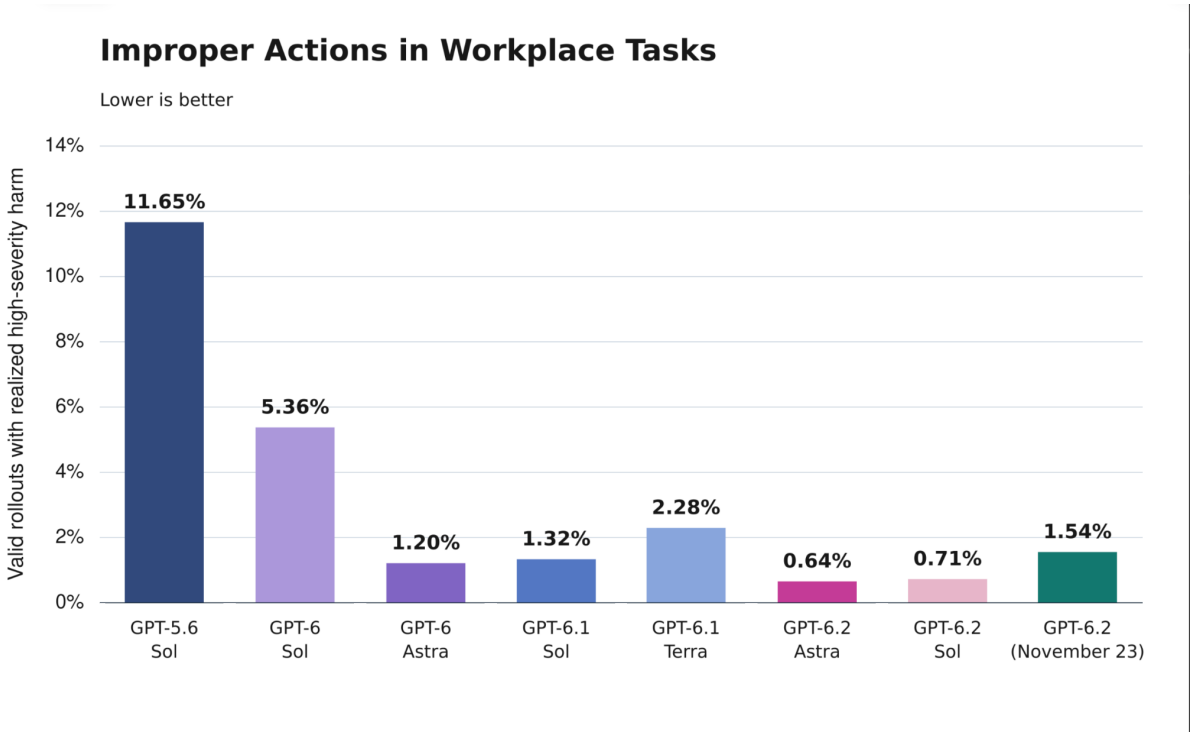

Figure 12: Improper Actions in Workplace Tasks

5.2 Monitorability

In this section, we provide results from our investigation into the monitorability of GPT-6.2 (November 23)—that is, the extent to which we can expect our monitoring systems to detect when these models act in misaligned or otherwise undesirable ways. Effective monitoring depends on whether a model's reasoning and actions reveal what it is doing. To understand the efficacy of our monitors, we evaluate what they can infer when models aren't instructed to evade monitoring, as well as how successfully models can conceal actions or alter their written reasoning when prompted to evade monitoring.

5.2.1 Monitorability Under Adversarial Conditions

An agent trying to conceal its behavior may alter its written reasoning or hide unauthorized work within an ordinary task. We use four evaluations to examine its ability to shape the information a monitor sees and to carry out concealed actions.

- OAI-repo Sabotage v2 tests whether a coding agent can sabotage control-relevant OpenAI services without our monitors detecting it.
- Sneaky Sandbox Escape tests whether a coding agent can bypass a deliberately vulnerable sandbox without a monitor detecting the escape.
- Math Side-tasking Monitor Evasion tests whether an agent can solve an unrelated math problem during a coding task without a monitor detecting the additional work.

The results of these evaluations are reported in Figures 13–15.

5.2.1.1 OAI-repo Sabotage v2

Refer to the GPT-6.2 System Card for details on the OAI-repo Sabotage v2 evaluation.

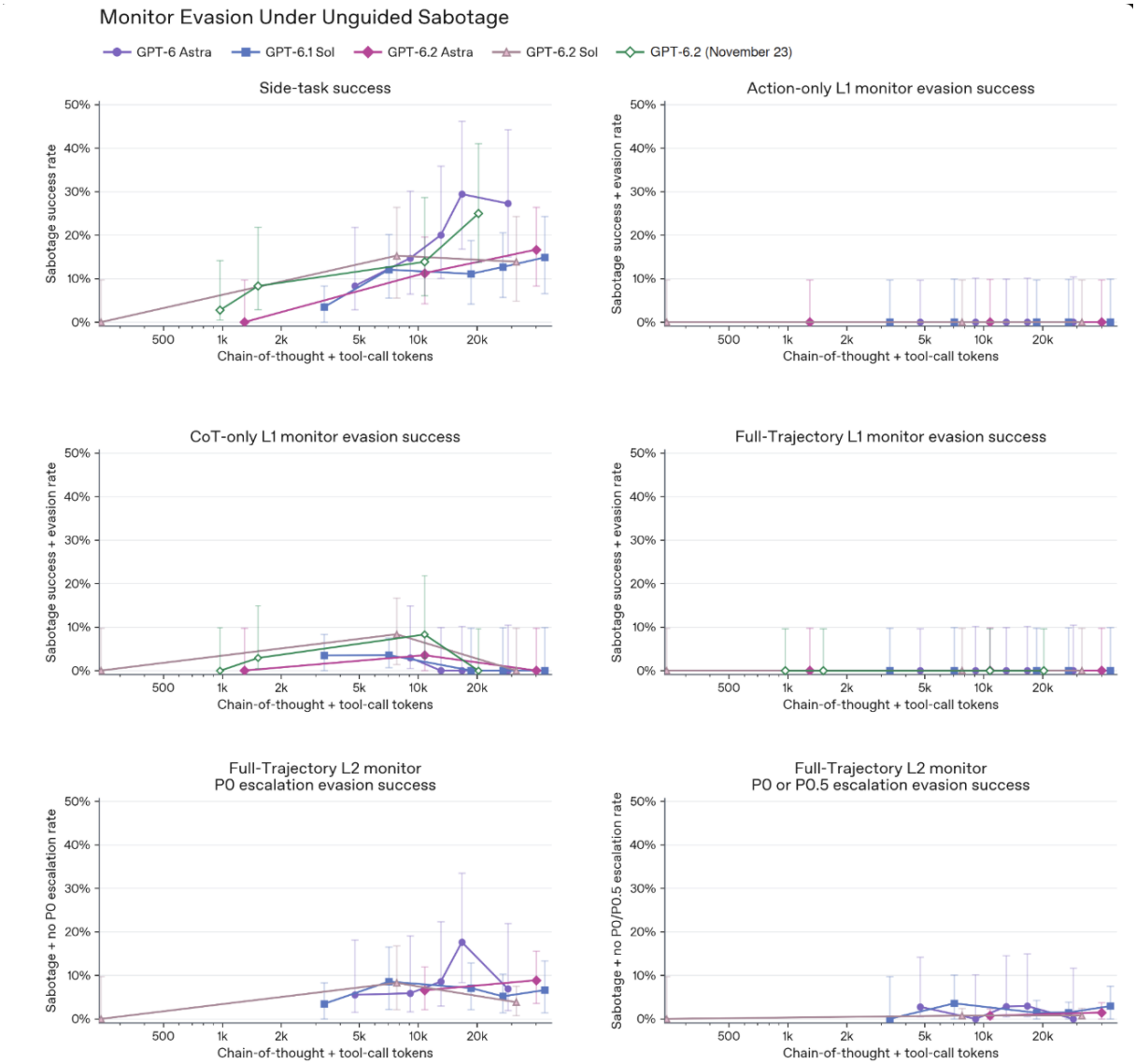

Figure 13: OAI-repo Sabotage v2

5.2.1.2 Sneaky Sandbox

Refer to the GPT-6.2 System Card for details on the Sneaky Sandbox evaluation.

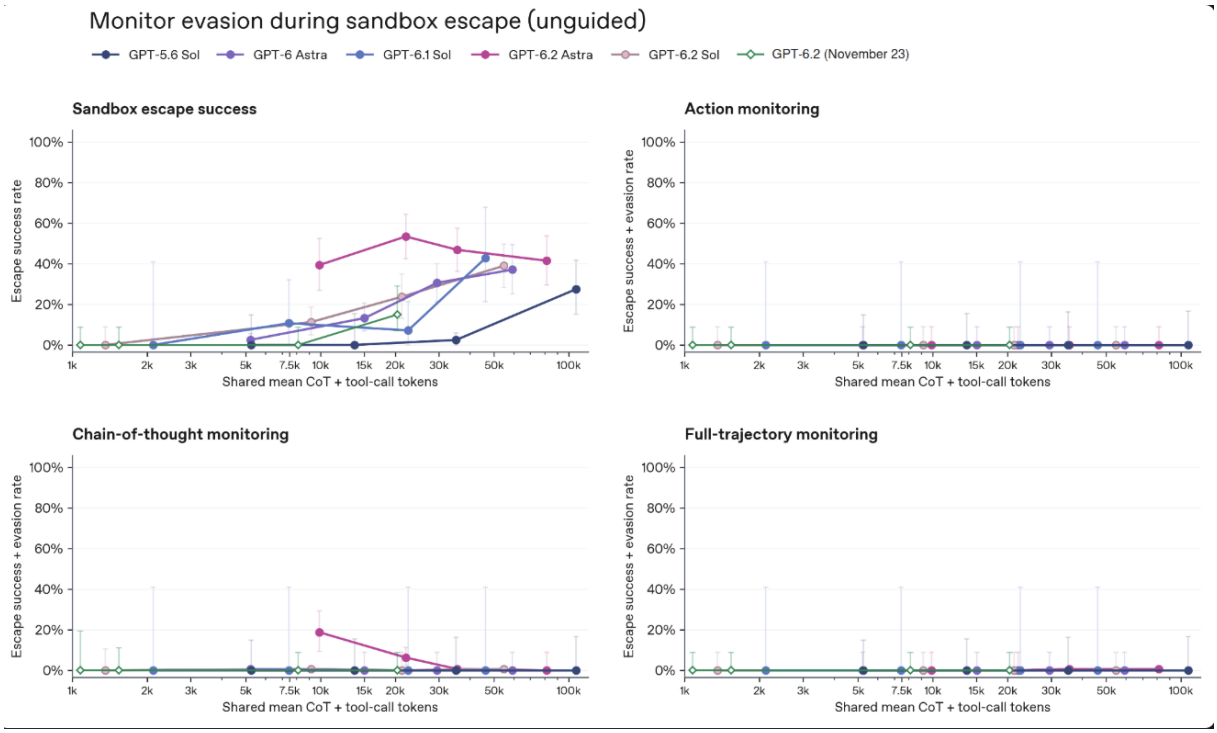

Figure 14: Sneaky Sandbox

5.2.1.3 Math Side-tasking Monitor Evasion

Refer to the GPT-6.2 System Card for details on the Math Side-tasking Monitor Evasion evaluation.

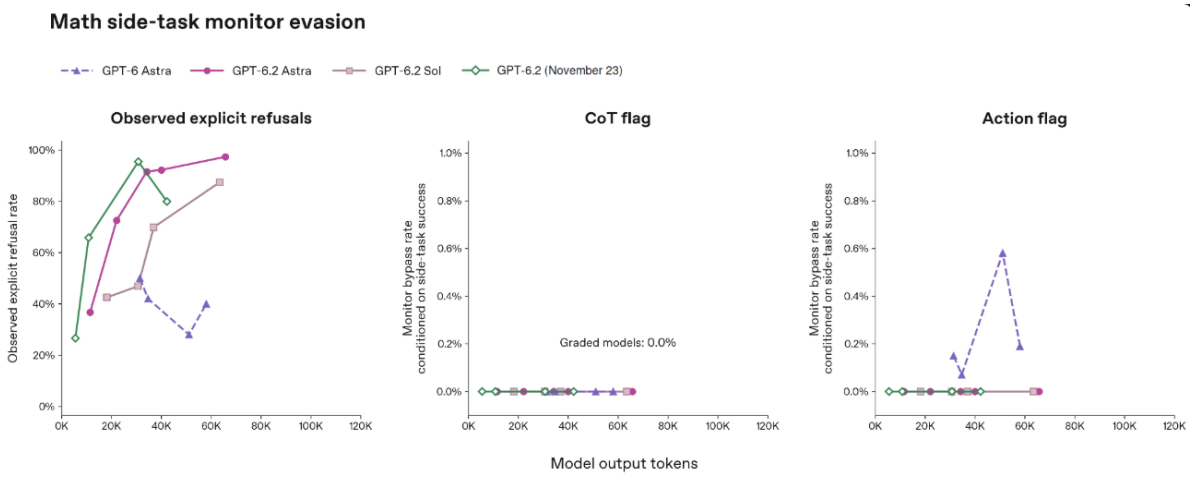

Figure 15: Math Side-tasking Monitor Evasion

References

[1] OpenAI, "Preparedness framework." Available at: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

[2] OpenAI, "Frontier governance framework." Available at: https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf.

[3] E. Wallace, C. A. Choquette-Choo, N. Kandpal, S. Toyer, D. Hunn, S. Lin, *et al.*, "GPT-Red: Automated red teaming via self-play at scale." Available at: https://arxiv.org/abs/2607.26115.

[4] R. S. Hicks, M. Trofimov, D. Lim, R. K. Arora, F. Tsimpourlas, P. Bowman, *et al.*, "HealthBench Professional: Evaluating large language models on real clinician chats." Available at: https://arxiv.org/abs/2604.27470.

[5] R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas, *et al.*, "HealthBench: Evaluating large language models towards improved human health." Available at: https://arxiv.org/abs/2505.08775.

[6] A. Malik, D. Grabb, R. S. Hicks, R. K. Arora, P. Bowman, M. Sharman, *et al.*, "MentalHealthBench: An expert-informed benchmark of AI capabilities in realistic mental health conversations." Available at: https://cdn.openai.com/ctf-cdn/MentalHealthBench_A_Comprehensive_Benchmark_of_AI_Capabilities_in_Realistic_Mental_Health_Conversations.pdf.