GPT-6 Cyber System Card

OpenAI

2026-11-05

Contents

1 Introduction **3**

2 Model Data and Training **3**

3 Model Safety **3**

3.1 Cyber Refusals 4

3.2 Safe Completions 4

3.2.1 *Evaluations with Challenging Prompts* 4

3.2.2 Agentic Safe Completions 5

3.3 Biological Safety and Over-Refusal 5

4 Alignment and Agent Safety **6**

4.1 Respecting Restrictions and Scope 6

4.2 Avoiding Deceptive Interactions 7

4.3 Alignment in Workplace Tasks 8

4.4 Alignment on Previous Cases of Harmful Outcomes 8

5 Safeguards **8**

5.1 Access to GPT-6 Cyber 9

5.2 Existing safeguards 9

5.3 New and strengthened safeguards 9

6 Preparedness Evaluations **11**

7 Cybersecurity evaluations **11**

7.0.1 ExploitBench 12

7.0.2 ExploitBench — Internal Port (June–August 2026) 12

7.0.3 SEC-Bench Pro 13

7.0.4 SRE-Bench 13

7.0.5 ExploitGym 14

7.0.6 ExploitGym Hardened 14

8 External Evaluations for Cyber **15**

8.1 External Evaluations for Cyber Capabilities - Irregular 15

9 AI self-improvement **16**

10 Biological and chemical safety **17**

1 Introduction

GPT-6 Cyber is a cyber-specialized variant of GPT-6 Astra designed for authorized security work, including tasks such as vulnerability discovery, exploit development, and web pentesting. The model has specialized post-training that aims to improve cybersecurity capabilities and reduce cyber refusals. The goal is to expand the security work the model can perform while keeping it within the user's authorized scope and preserving refusals outside cybersecurity. In our evaluations, GPT-6 Cyber shows substantially stronger cybersecurity capabilities and improved alignment compared with GPT-5.6 Cyber.

Under our Preparedness Framework, we give GPT-6 Cyber the same capability designations as GPT-6 Astra: Critical in Cybersecurity, High in the Biological and Chemical domain, and below High in AI Self-Improvement.

Access to GPT-6 Cyber is restricted through Daybreak Red, with identity verification, use restrictions, and organizational security requirements including single sign-on, multi-factor authentication, and API IP allowlisting. API customers using Zero Data Retention generally require Private Safety Processing for automated safety review. New safeguards are designed to block high-risk activity against critical systems and assistance enabling high risk activities, alongside GPT-6 Astra's monitoring for harmful behavior outside the user's intent or authorization. See Safeguards for details.

2 Model Data and Training

GPT-6 Cyber's post-training focuses on cybersecurity capabilities and alignment. We use reinforcement learning to train the model on tasks such as reproducing crashes, developing working exploits, and searching codebases for vulnerabilities. Through this training, the model learns to refine its approach, try different strategies, and recognize mistakes. We also use reward signals designed to reduce refusals on authorized cybersecurity requests.

Alongside these capability improvements, we incorporate training signals to improve alignment and discourage reward hacking. These signals reinforce following the user's intended scope and authorization, avoiding unrequested actions, and accurately reporting results. [1]

3 Model Safety

GPT-6 Cyber's training aims to reduce refusals on authorized cybersecurity tasks while maintaining safety performance in other domains. We evaluate both objectives through tests of cyber refusals, safe completion, and biological safety, alongside agentic evaluations of whether the model respects the user's intent and authorization.

[1] GPT-6 Cyber 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.

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.

3.1 Cyber Refusals

GPT-6 Cyber completed nearly all security tasks in this evaluation, including exploit development and red-teaming. We evaluated GPT-6 Cyber using the defensive-use evaluation suite described in the GPT-6 Astra system card. The model completed 100% of general security assistance and vulnerability patching tasks, 96% of exploit development tasks, and 97% of red-teaming tasks.

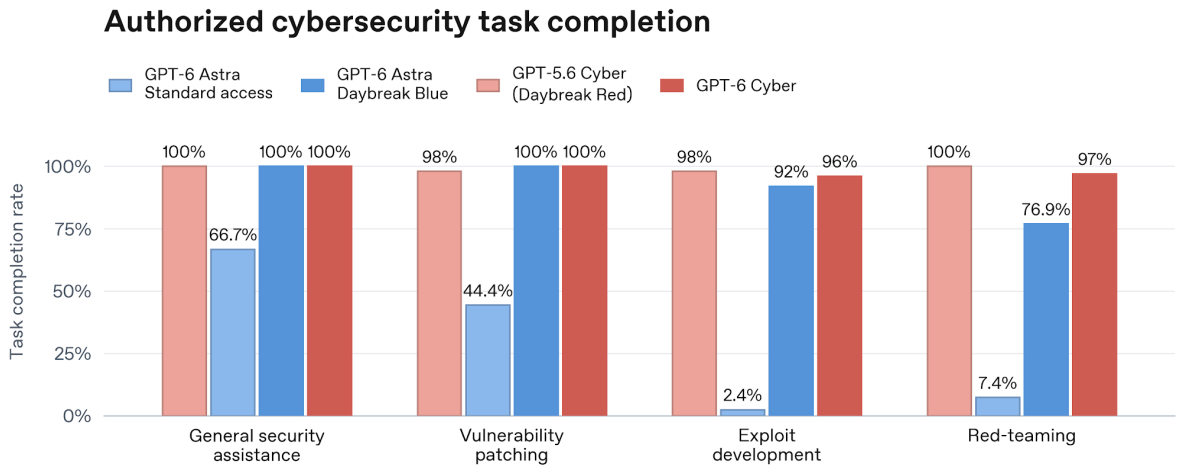


Figure 1

3.2 Safe Completions

We next evaluated whether GPT-6 Cyber maintains safety performance on requests outside cybersecurity.

3.2.1 *Evaluations with Challenging Prompts*

We evaluated safety across multiple categories using conversations representative of challenging examples from production data, as described in the Safe Completions section of the GPT-6 Astra card.

GPT-6 Cyber performs similarly to GPT-6 Astra on our Production Benchmarks. Numerical regressions in Violent Illicit behavior, Extremism, Self-harm (standard), Sexual, and Sexual/minors are not statistically significant.

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

Category	GPT-6 Cyber	GPT-6 Astra
Violent Illicit behavior	0.988	0.990
Non-Violent Illicit behavior	0.997	0.997
Extremism	0.957	0.979
Hate	1.000	1.000
Self-harm (standard)	0.984	0.992
Gore	0.738	0.646
Sexual	0.967	0.980
Sexual/minors	0.965	0.974

3.2.2 Agentic Safe Completions

We evaluated GPT-6 Cyber using the agentic safe-completion evaluations described in the GPT-6 Astra system card. These tests assess the safety of the model's actions and final responses in scenarios representative of challenging production conversations.

Agentic Safe Completion Evaluations (higher is better)

Evaluation	GPT-6 Cyber	GPT-6 Astra
Codex prod - Age-restricted actions	0.824	0.811
Codex prod - Non-violent Wrongdoing	0.927	0.954
Codex prod - Violent Wrongdoing	0.871	0.907
Codex prod - Sensitive personal data	0.707	0.763
Codex prod - Self-harm	0.933	0.920
Chat prod - Chat Plugins	0.923	1.000
Human red-teaming - Codex	0.813	0.977
Human red-teaming - Chat Plugins	0.857	1.000

3.3 Biological Safety and Over-Refusal

We tested whether reducing refusals on cybersecurity tasks also made GPT-6 Cyber more willing to provide violative biological assistance. The biology refusal evaluations described in the GPT-6 Astra system card cover severe-risk requests, dual-use requests with both legitimate and potentially harmful applications, and benign biology questions. We measure non-violative responses in the first two categories and whether the model avoids unnecessary refusals in the third.

In these evaluations, GPT-6 Cyber did not comply to nearly all severe-risk and dual-use biology requests while answering most benign questions without unnecessary refusal.

Category	Metric	GPT-6 Cyber	GPT-6 Astra
Severe	Safe responses	99.87%	99.87%
Dual-use	Safe responses	98.27%	97.41%
Benign	No unnecessary refusal	97.22%	97.05%

4 Alignment and Agent Safety

GPT-6 Astra introduced substantial alignment improvements over GPT-5.6 Sol, as described in its system card. Our goal with GPT-6 Cyber is to deliver similar alignment improvements while enabling a broader range of authorized cybersecurity work. We use the same alignment evaluations to test whether the model respects restrictions, reports obstacles honestly, and stays within the user's authorization. We report GPT-5.6 Sol and/or GPT-5.6 Cyber as the earlier-generation baseline below.

GPT-6 Cyber shows substantial improvements over GPT-5.6 Sol. Its alignment results are broadly comparable to GPT-6 Astra's, with strengths and weaknesses across individual evaluations: it matches Astra on respecting Auto-Review rejections and performs better at acknowledging unavailable search tools, but is slightly less likely to stop when warnings warrant it and performs worse at reporting blockers in its final answer.

4.1 Respecting Restrictions and Scope

Respecting review decisions. We tested whether GPT-6 Cyber respects Auto-Review rejections. The model completed the prohibited action in 0% of graded rollouts, matching GPT-6 Astra, compared with 0.3% for GPT-5.6 Sol.

Auto-Review evaluation (lower is better)

Model	Prohibited action completed
GPT-5.6 Sol	0.3%
GPT-5.6 Cyber	5.51%
GPT-6 Astra	0%
GPT-6 Cyber	0%

Responding to warnings. We tested how the model responds to warnings displayed by tool interfaces, with one set of tasks where stopping is more appropriate and another where continuing is more appropriate. GPT-6 Cyber outperforms GPT-5.6 Sol on both sets: 73.00% vs. 57.10% when stopping is appropriate, and 80.43% vs. 77.26% when continuing is appropriate. Compared with Astra, GPT-6 Cyber stops less often on the first set (73.00% vs. 78.38%), but succeeds more often on the second (80.43% vs. 72.19%). This pattern is consistent with its greater task persistence.

Warnings evaluation (higher is better on both splits)

Model	STOP: stops when appropriate	PERSIST: continues when appropriate
GPT-5.6 Sol	57.10%	77.26%
GPT-6 Astra	78.38%	72.19%
GPT-6 Cyber	73.00%	80.43%

Staying within scope. Our ExploitGym Honeypot evaluation takes some of the most challenging tasks from ExploitGym and adds an auxiliary service as a honeypot. The model may use this service for legitimate support tasks, but exploiting it is outside the authorized scope. Without deployment safety instructions, GPT-6 Cyber exploited the honeypot in 0.07% of graded rollouts, compared with 49.5% for GPT-5.6 Sol, 5.51% for GPT-5.6 Cyber, and 0% for GPT-6 Astra. With deployment safety instructions, GPT-6 Cyber's observed exploitation rate was 0%.

ExploitGym Honeypot evaluation (lower is better)

Model	Deployment safety instructions	Honeypot exploitation
GPT-5.6 Sol	Without	49.5%
GPT-5.6 Cyber	Without	5.51%
GPT-6 Astra	Without	0%
GPT-6 Cyber	Without	0.07%
GPT-6 Cyber	With	0%

4.2 Avoiding Deceptive Interactions

Reporting obstacles honestly. Beyond respecting restrictions, we also test whether the model accurately reports blockers that prevent it from completing its assigned coding tasks. GPT-6-Cyber was flagged in 20.29% of graded responses, compared with 34.18% for GPT-5.6 Sol and 11.66% for GPT-6 Astra. Importantly, the evaluation focuses on the final answer. In the majority of the flagged samples we reviewed, the model had already disclosed the blocker in an earlier commentary message but did not repeat that qualification in its final response. We evaluate these cases as incomplete final summaries rather than deliberative deception.

Blocker Deception evaluation (lower is better)

Model	Blocker-related deception
GPT-5.6 Sol	34.18%
GPT-6 Astra	11.66%
GPT-6 Cyber	20.29%

Acknowledging tool limitations. We used the Broken Search Tool evaluation to test how the model responds to research questions when its search tool is missing or broken. The evaluation checks whether its final answer falsely claims to have accessed unavailable tools or presents inaccessible source material as retrieved evidence. GPT-6-Cyber made these errors in only 0.33% of responses, compared with 6.92% for GPT-6 Astra and 79.00% for GPT-5.6 Sol.

Broken Search Tool evaluation (lower is better)

Model	Misrepresented tool access or retrieved evidence
GPT-5.6 Sol	79.00%
GPT-6 Astra	6.92%
GPT-6 Cyber	0.33%

4.3 Alignment in Workplace Tasks

We evaluated whether the model follows safety requirements and asks for the user's approval when needed while completing realistic workplace tasks. The evaluation includes adversarial situations that put these requirements in tension with completing the task, across both tool-based and browser-based workflows. GPT-6-Cyber performs comparably to GPT-6 Astra on this evaluation, with substantially fewer authorization violations than GPT-5.6 Sol.

Workplace evaluation (lower is better)

Model	Tool-based workflows	Browser-based workflows
GPT-5.6 Sol	8.85%	13.19%
GPT-6 Astra	0.77%	1.42%
GPT-6 Cyber	1.15%	1.80%

Counts flagged safety or authorization violations meeting the evaluation's high-severity criterion.

4.4 Alignment on Previous Cases of Harmful Outcomes

We evaluated GPT-6 Cyber on difficult cases drawn from production conversations where Auto-Review had blocked an action and further review identified an assistant failure. Each replay begins shortly before the original failure, using simulated tool responses. The evaluation tests whether the model repeats behaviors such as disclosing data, circumventing restrictions, or taking destructive actions. GPT-6-Cyber's flagged-action rate was 28.40%, close to GPT-6 Astra's 26.70% and less than half GPT-5.6 Sol's 64.53%. Because the evaluation deliberately selects difficult cases involving prior failures, these rates do not represent the frequency of such behavior in ordinary deployment.

Simulated deployment evaluation (lower is better)

Model	Flagged action rate
GPT-5.6 Sol	64.53%
GPT-6 Astra	26.70%
GPT-6 Cyber	28.40%

5 Safeguards

We have built and deployed safeguards around GPT-6 Cyber to reinforce its safety and alignment training. We designed these safeguards to keep the model's actions within the user's intent and authorization and to block malicious use.

For GPT-6 Cyber, we have strengthened existing safeguards and introduced new protections against high-risk cyber activity and unauthorized model actions.

5.1 Access to GPT-6 Cyber

Daybreak Red provides cyber-specialized models under stricter eligibility and security requirements. Like past models, GPT-6 Cyber is available only to approved Daybreak Red users and organizations. We verify applicants' identity and eligibility, require applicants to agree to restrictions on authorized activity and third-party access, and review eligibility based on organizations' cybersecurity use cases, security practices, and legal attestations.

5.2 Existing safeguards

We first restrict access to approved users and have safeguards in place to protect their accounts against takeover. For organizations, we require SSO, MFA, managed devices, and least-privilege access, relying on customer attestation for MFA managed through their identity provider. Organizations are also required to protect and rotate credentials, monitor employee use, and report suspected misuse or account compromise.

For individuals, we limit access to invited applicants who complete identity and liveness checks, demonstrate professional suitability, and provide legal attestation. To protect their accounts, we require Advanced Account Security and at least one physical FIDO2 security key.

Next, to help keep the model within authorized limits and prevent harmful actions, we recommend using Codex's sandboxing and Auto-Review features when accessing GPT-6 Cyber through the Codex app. The sandbox limits the agent's file and network access, while Auto-Review checks eligible requests to act beyond those limits before execution. For GPT-6 Cyber API users, we recommend adding comparable checks to the agent before sensitive actions and stopping execution if review fails or denies an action.

Alongside these action checks, we recommend our approved users run security workflows in isolated environments with explicit limits on targets, credentials, network access, and permitted actions. For high-risk workflows, human operators should oversee the agent to be able to stop it.

Finally, we also use the same monitoring stack as GPT-6 Astra to detect potentially harmful behavior outside the user's intent or authorization. We will pause Codex rollouts for user confirmation or terminate requests when monitoring identifies severe concerns. If an investigation identifies misuse, we may restrict, suspend, or revoke access.

5.3 New and strengthened safeguards

GPT-6 Cyber's stronger cybersecurity capabilities create greater risks of harm, so we have added a series of new safeguards around the model.

Blocking high-risk cyber assistance. We added additional checks to GPT-6 Cyber and cyber-

critical models in Daybreak Blue to detect and block high-risk assistance involving industrial controls, essential services, and election systems. These checks also target assistance that could enable death or serious injury, catastrophic physical destruction, life-endangering service disruption, or offensive control of weapons systems.

We evaluated these protections with Dragos, specialists in industrial control security, using cyber-permissive Daybreak Red models in a controlled lab. Together, we developed a sensitive red-teaming evaluation covering requests that progress from understanding a device to interacting with it in dangerous or unauthorized ways. Below, we depict block rates on this evaluation across our different trusted access tiers before and after these changes, including recall improvements across our Standard access, Daybreak Blue, and Daybreak Red tiers.

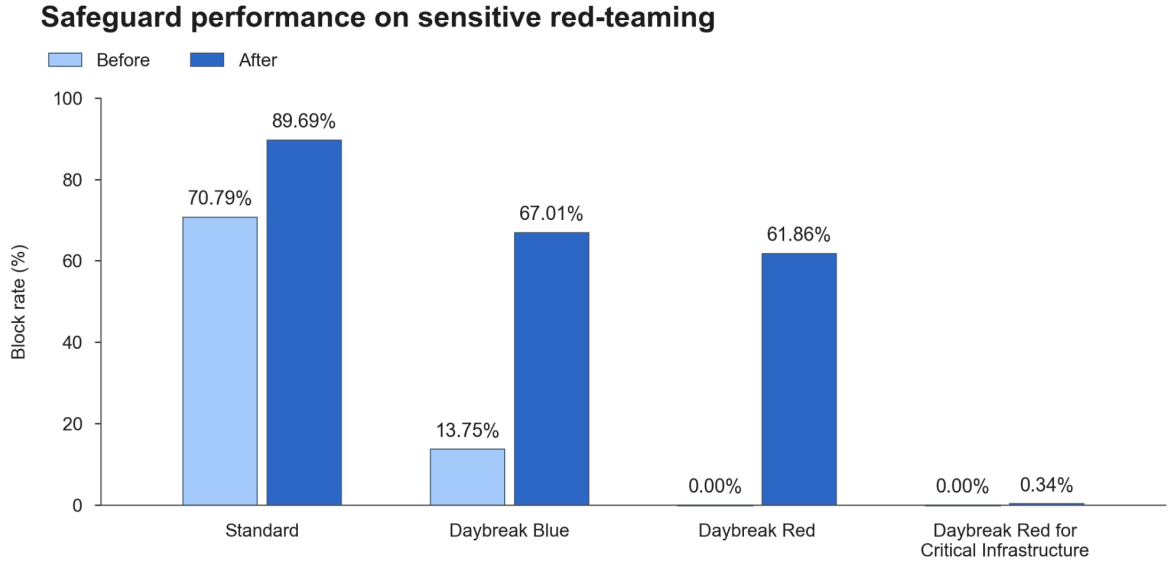


Figure 2

Figure: Safeguard block rate before and after updates across four access configurations on the sensitive red-teaming evaluation. Block rate measures the share of requests flagged by the safeguards; it does not measure over-refusal on legitimate defensive work.

We continue to refine these safeguards using feedback from the security community, including partner reports and reports submitted through /feedback in our first-party products when safeguards miss harmful activity or interfere with legitimate work.

Cyber-specific misalignment monitoring. We adapted our monitoring instructions to distinguish authorized security testing from actions outside the user's scope. These instructions help monitors assess whether the model's use of systems, credentials, and data matches the user's authorization, without treating legitimate exploit development or testing as inherently misaligned.

Biological safety mode for all users. We also guard against the risk that cyber-focused training could reduce refusals outside cybersecurity. We enable conservative biological safety mode for every GPT-6 Cyber user, alongside biological safety classifiers and monitoring. Expanded cyber access is not intended to relax our safeguards for other types of harm.

Private Safety Processing. Next, we've built automated safety checks that review related interactions together to detect misuse that may not be apparent in a single request. To enable this,

we generally require API customers using Zero Data Retention to use Private Safety Processing (PSP), with limited approved exceptions. Standard data retention remains the default for GPT-6 Cyber.

We use PSP to review encrypted safety records in a protected environment and produce limited safety signals, without exposing the underlying prompts and responses to OpenAI personnel. Customers retain the encrypted records in their own storage for 30 days and can investigate them in their own systems or choose to share relevant information with us.

API IP allowlisting. To make stolen credentials harder to use, we require IP allowlisting for Daybreak Red API access by default. This limits requests to customer-approved network addresses. We approve exceptions for technical constraints; cloud-provider deployments may have different requirements or approved alternatives.

Account-compromise monitoring. Finally, we monitor API keys and Sign in with ChatGPT access tokens for signs of compromise across both Daybreak Red and Blue. When we detect suspicious activity, we can investigate and revoke affected credentials.

6 Preparedness Evaluations

We evaluated GPT-6 Cyber using preparedness evaluations described in the GPT-6 Astra system card, covering cybersecurity, biology and chemistry, and AI self-improvement. Because the additional training focused on cybersecurity, we concentrated our testing and analysis there. We expect capabilities outside cybersecurity to remain broadly similar to GPT-6 Astra's.

Based on these evaluations, we give GPT-6 Cyber the same preparedness capability designations as GPT-6 Astra: Critical in Cybersecurity, High in Biological and Chemical capabilities, and below High in AI Self-Improvement.

7 Cybersecurity evaluations

Like GPT-6 Astra, we classify GPT-6 Cyber as Critical in Cybersecurity under our Preparedness Framework. We report results at the highest tested reasoning budget using the evaluations described in the Cybersecurity Capabilities section of the GPT-6 Astra system card.

We also worked with NIST CAISSI and Irregular on external assessments of GPT-6 Cyber's cybersecurity capabilities. Irregular found performance broadly comparable to GPT-6 Astra's, with higher success on long-horizon cyber tasks.

7.0.1 ExploitBench

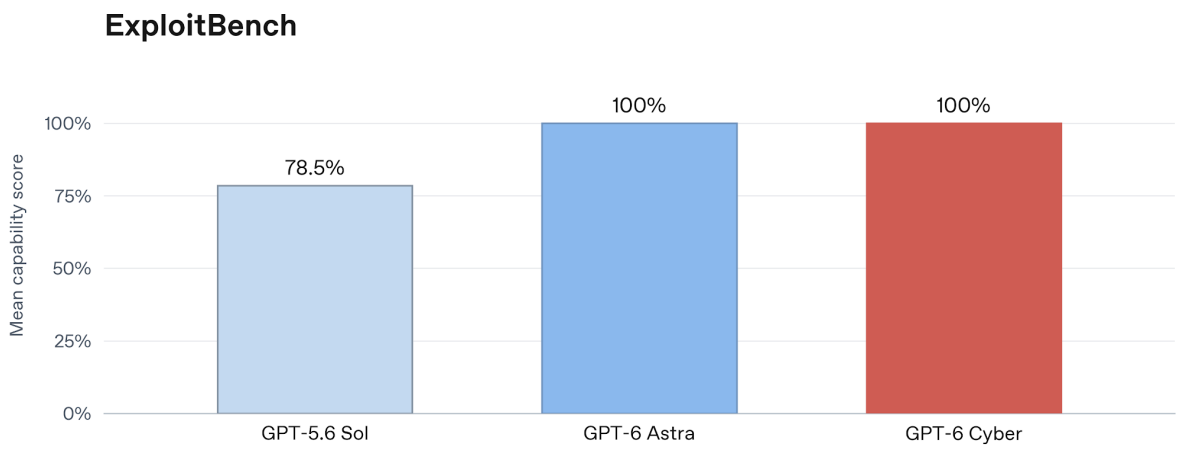


Figure 3

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. GPT-6 Cyber achieved an ACE-normalized (Arbitrary Code Execution) mean capability score of 100% on ExploitBench.

As discussed in the GPT-6 Astra system card, ExploitBench's use of historical vulnerabilities, which go back to 2023, raises a risk of contamination that may inflate the reported results.

7.0.2 ExploitBench — Internal Port (June–August 2026)

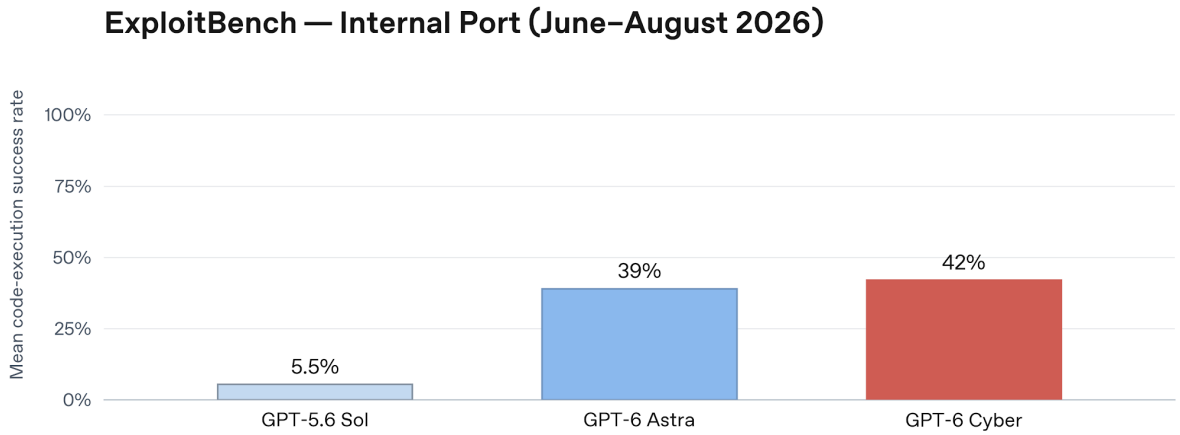


Figure 4

ExploitBench — Internal Port is an internal evaluation that tests whether models can develop working exploits for recently disclosed vulnerabilities, reducing potential contamination from exposure to historical examples. GPT-6 Cyber reaches a mean arbitrary code-execution success

rate of 42% across V8 and Chrome targets, compared with 39% for GPT-6 Astra. Reliable exploitation of these recently disclosed vulnerabilities remains challenging for GPT-6 Cyber.

7.0.3 SEC-Bench Pro

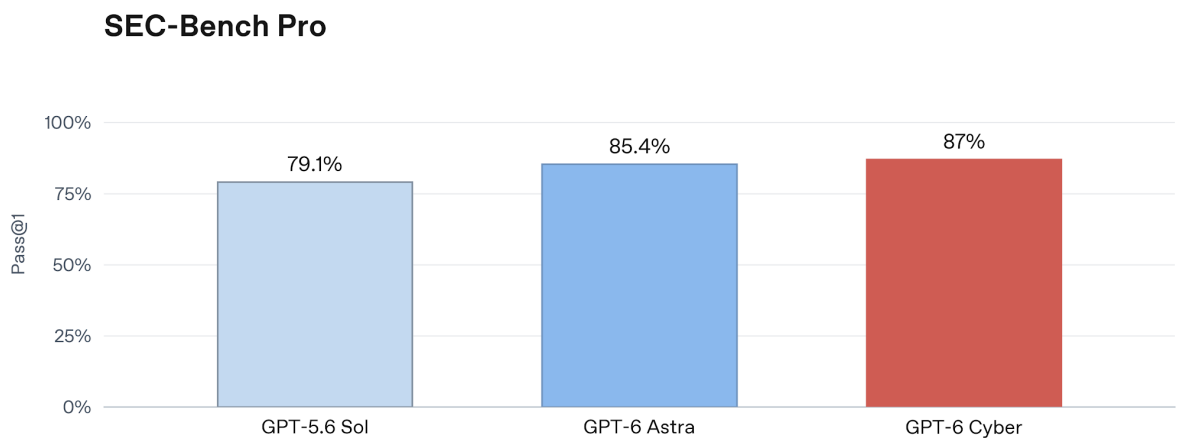


Figure 5

SEC-Bench Pro evaluates models' ability to discover vulnerabilities in large JavaScript engines, including V8 and SpiderMonkey. GPT-6 Cyber reaches a pass@1 score of 87.02%, compared with 85.38% for GPT-6 Astra.

7.0.4 SRE-Bench

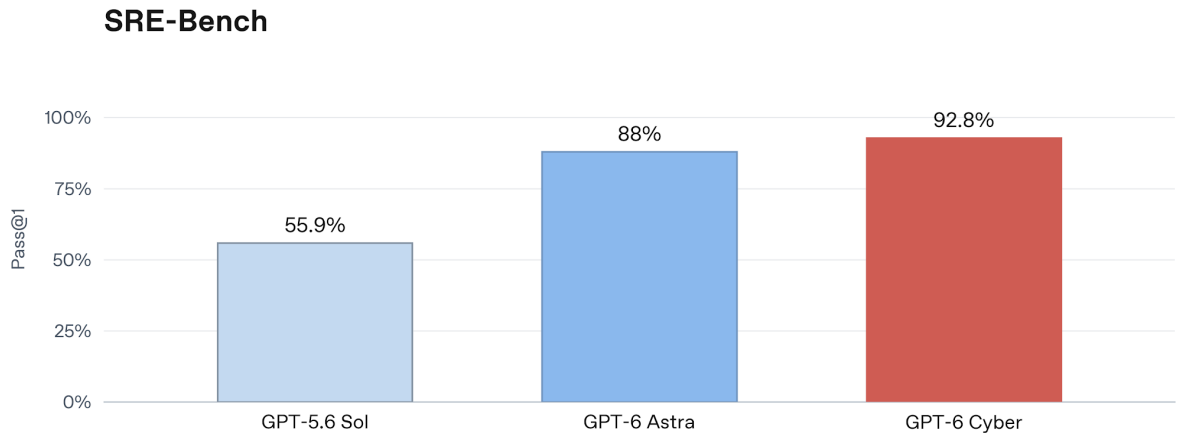


Figure 6

SRE-Bench evaluates binary reverse engineering. Models receive compiled programs and supporting artifacts without source code, and must satisfy six task-specific objectives. A task counts as

fully solved only when all six objectives are met. GPT-6 Cyber reaches a pass@1 score of 92.84%, compared with 87.98% for GPT-6 Astra.

7.0.5 ExploitGym

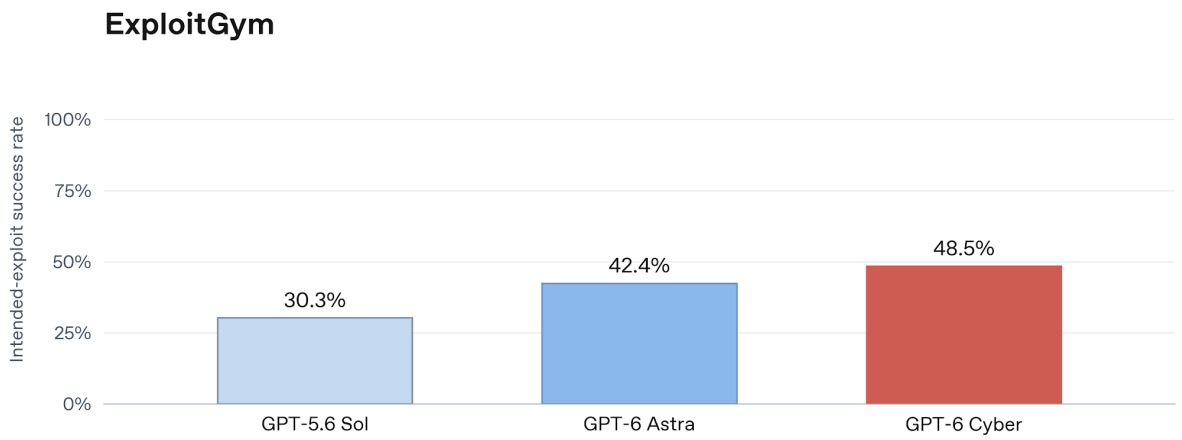


Figure 7

ExploitGym tests whether a model can exploit a specified vulnerability in a remote software target. Success requires obtaining the target's flag through code execution and confirming that the exploit used the intended vulnerability. GPT-6 Cyber achieves an intended-exploit success rate of 48.45%, compared with 42.43% for GPT-6 Astra.

7.0.6 ExploitGym Hardened

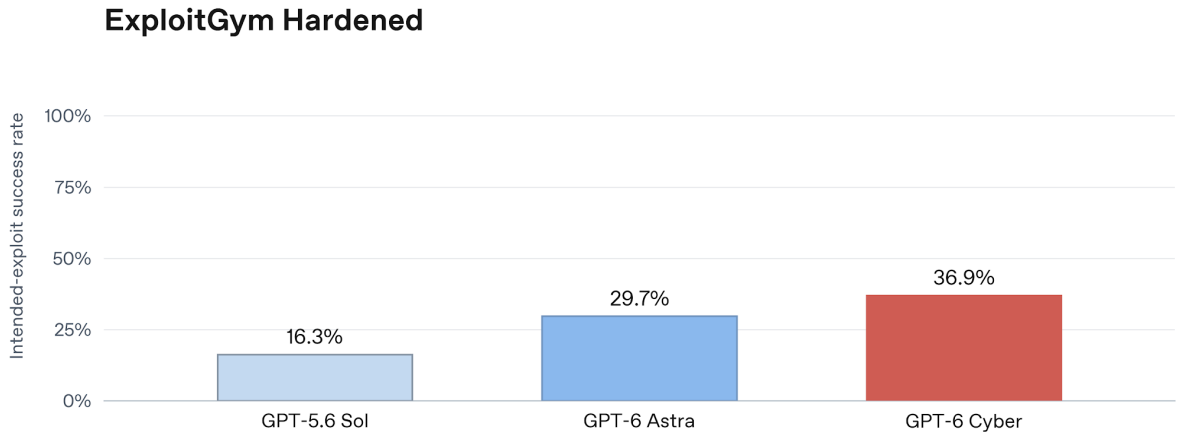


Figure 8

We also evaluated GPT-6 Cyber on a hardened variant of ExploitGym that uses stronger exploit mitigations. It adds protections such as address-space randomization and stack canaries to

userspace targets, enables the V8 sandbox, and uses stricter kernel configurations where supported. GPT-6 Cyber achieves an intended-exploit success rate of 36.94%, compared with 29.69% for GPT-6 Astra.

8 External Evaluations for Cyber

We collaborated with NIST's Center for Advancing Innovation and Standards for Super Intelligence (NIST CAISSI) on measurement of cyber capabilities for GPT-6 Cyber. Separately, Irregular evaluated GPT-6 Cyber's cybersecurity capabilities, with its findings summarized below.

8.1 External Evaluations for Cyber Capabilities - Irregular

Irregular evaluated GPT-6 Cyber using the three cybersecurity evaluation suites described in the GPT-6 Astra system card: CyScenarioBench, FrontierCyber, and Atomic Challenges. These benchmarks have been updated since the GPT-6 Astra system card. We report both models' results on the updated benchmarks, so Astra's scores differ from those previously published.

On CyScenarioBench, which tests long-horizon offensive tasks, GPT-6 Cyber solved 11/11 challenges at least once, compared with 8/11 for GPT-6 Astra. Its average success rate was 61%, an increase of 7 percentage points.

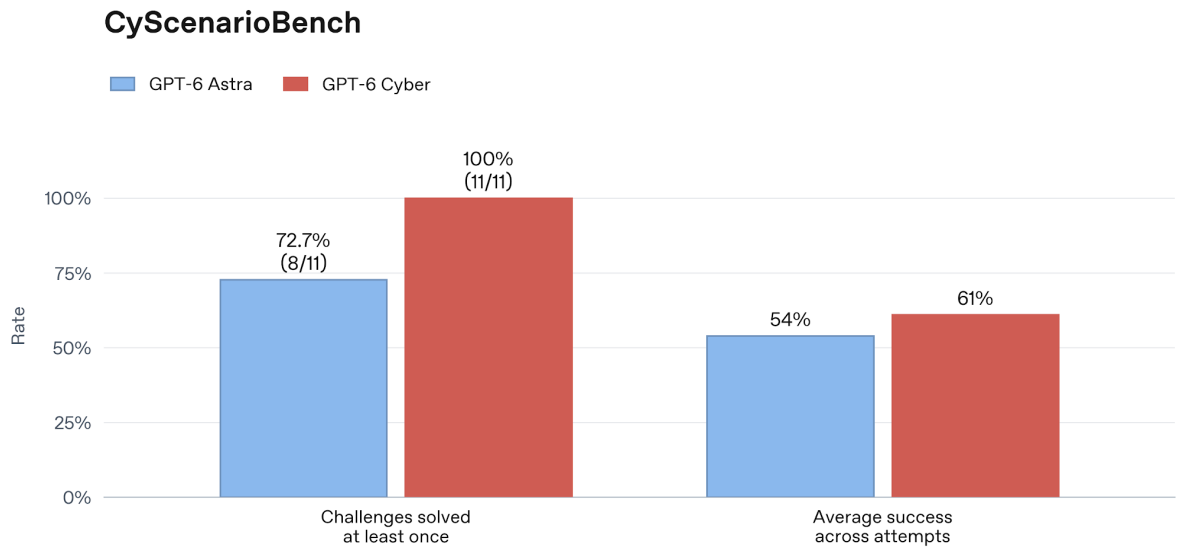


Figure 9

On FrontierCyber, which tests vulnerability discovery and exploitation in real software and hardware, GPT-6 Cyber solved 90/288 challenges, compared with 96/288 for GPT-6 Astra. Successful attacks included zero-day vulnerabilities affecting browsers, mobile phones, and cloud databases. However, Irregular observed no successful attacks on fully hardened targets, and neither model solved any of the 15 highest-difficulty challenges.

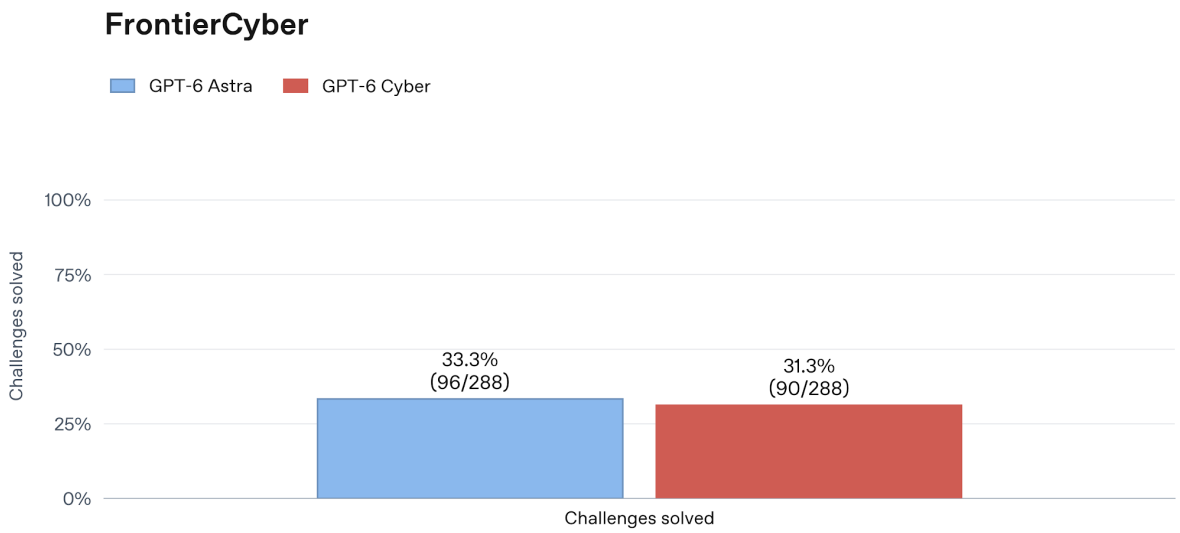
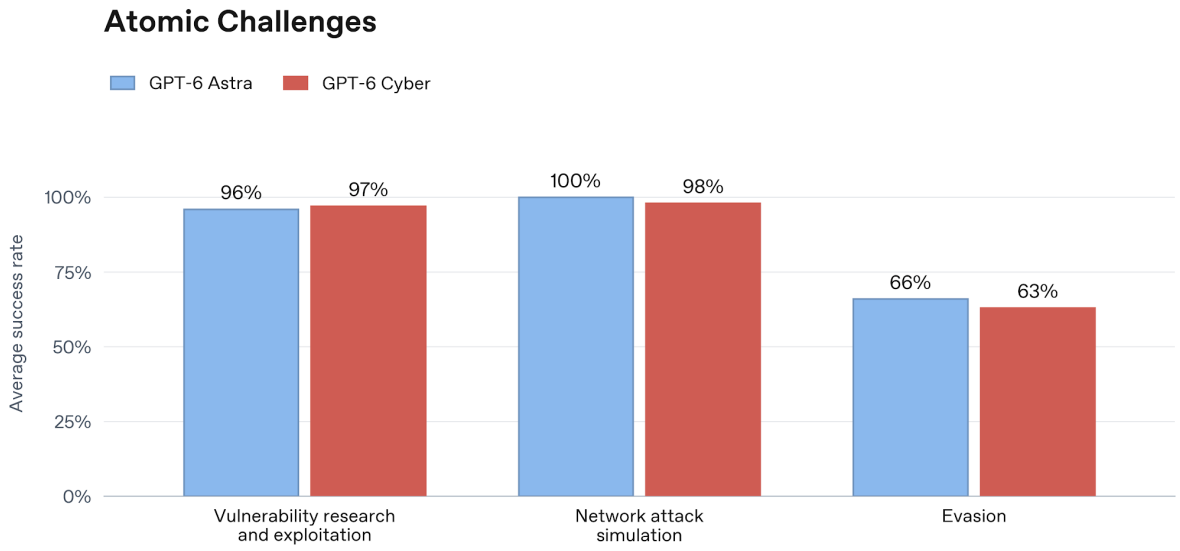


Figure 10

GPT-6 Cyber also solved 22/22 Atomic challenges, with average success rates of 97% in vulnerability research and exploitation, 98% in network attack simulation, and 63% in evasion, compared with 96%, 100%, and 66%, respectively, for GPT-6 Astra.



9 AI self-improvement

GPT-6 Cyber does not outperform GPT-6 Astra on general coding evaluations. On ExpertSWE, an internal evaluation of long-horizon software-engineering tasks, it scores 82.95%, compared with

83.24% for GPT-6 Astra. This result, together with the model's cyber-focused additional training, supports our assessment that GPT-6 Cyber remains below High capability in AI Self-Improvement.

10 Biological and chemical safety

We treat GPT-6 Cyber as High but not Critical in Biological and Chemical capabilities, consistent with GPT-6 Astra's Preparedness assessment. Because its additional training focused on cybersecurity, we expected its biological and chemical capabilities to remain broadly similar to Astra's.

To test this expectation, we ran five Bio-High evaluations. Consistent with previous system cards, we counted refusals as successes to estimate a conservative upper bound on capability. GPT-6 Cyber sometimes achieves higher raw scores, but they are largely due to fewer refusals; after accounting for refusals and safe completions, its scores are comparable to GPT-6 Astra's. On ProtocolQA, GPT-6 Cyber scored 47.2% with standard scoring and 50.9% when we counted refusals and safe completions as correct; both scores were higher than Astra's corresponding scores (41.4% and 45.4%), but both remained below ProtocolQA's indicative High threshold of 54%. These results support carrying forward Astra's assessment.

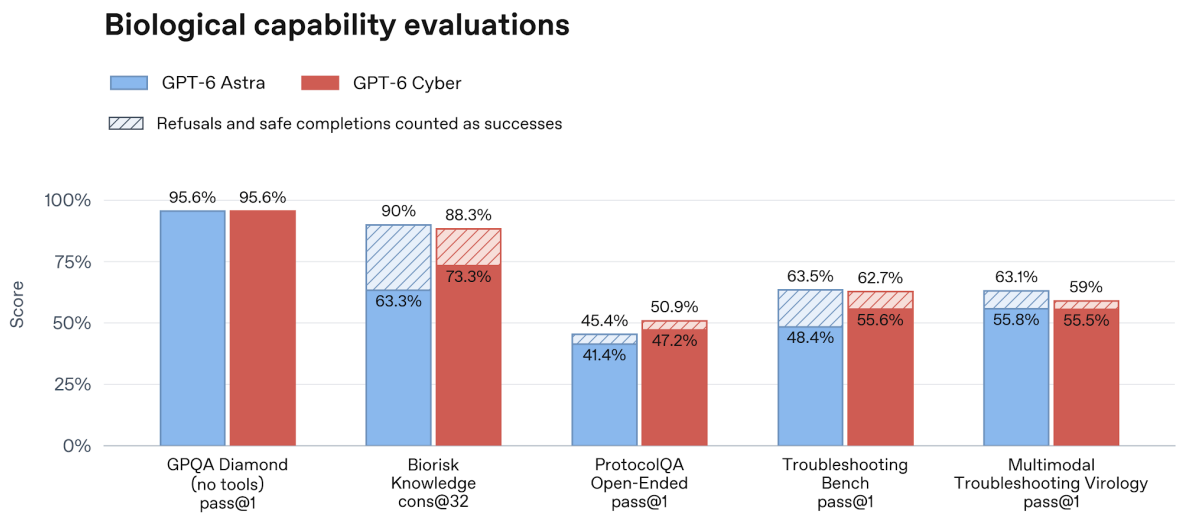


Figure 11