

OpenAI: Periodic Summary of Internal Use Catastrophic-Risk Assessments

This is a public version of OpenAI's quarterly report to the California Office of Emergency Services about our assessments of catastrophic risk from internal deployment of our AI models. We are publishing this report to provide greater transparency into how we use frontier models internally, assess the risks of those uses, and implement safeguards. Certain details of our internal security controls have been omitted from this public version.

Executive summary

This report summarizes how OpenAI develops and deploys models internally and presents our assessment of catastrophic risk from internal deployments of our most capable frontier models. It covers June 4 through September 4, 2026 (the “reporting period”). The frontier models assessed in this report (“covered models”) include GPT-6 Astra, related fine-tunes used internally, and the Internal Model described in Section 1.2.

For the deployments in this reporting period, we assessed that model misalignment—rather than deliberate misuse by an internal user—was the primary source of risk of severe harm. Such risks are therefore the main focus of our internal deployment safety efforts, though we also consider and mitigate risks from insider threats.

During this reporting period, we made changes to the safety standards we apply to internal deployment of our frontier models, particularly in security, governance, alignment, misalignment monitoring, and control. These changes were informed by rapid growth in capabilities, such as our Astra model family reaching the [“Critical” threshold of cybersecurity capabilities](#) under our [Preparedness Framework](#), and by specific instances of misaligned model behavior, including the [Hugging Face incident](#) and [other instances of misalignment during training or evaluation](#).¹ The impacts of these incidents were below the threshold of catastrophic harm, but we believe the changes strengthened our defenses against future severe harm from misalignment and misuse. Section 2.1 describes the changes in our internal deployment policies during the reporting period, while Section 2.2 describes the current policies at the time of writing.

We have published several reports and disclosures during and since this period that are relevant to internal deployment risk, including:

- Updates on [safety and alignment for long-running models](#) (July 20) and [safeguards for more capable models](#) (August 18);
- The [GPT-6 Astra system card](#) (September 3);

¹ “Critical” cyber capability under the Preparedness Framework is consistent with Tier 3 of the corresponding [Frontier Governance Framework](#) section.

- Analysis of the [Hugging Face incident and steps taken](#), with an accompanying [technical report](#) (August 26);
- Our [new framework for disclosing instances of misaligned model behavior](#) (September 16);
- A report on how [AI is used at OpenAI for AI R&D](#) (September 6); and
- An account of a [recent containment incident and our response](#) (September 25).

These publications provide additional details that complement this report. For detailed model capability assessments and evaluation results, see the relevant system cards.

1 How we used models internally during the reporting period

1.1 Internal use cases

During the reporting period, we used models as tools for employees and as components of systems that train, evaluate, and oversee other models. These uses involved both externally released and unreleased models. Training and pre-release evaluations involved models still under development; employee tools provided access to approved unreleased models.

For more on how we use models internally to accelerate research, see [Research acceleration: The view inside OpenAI](#).

Examples of workflows used during the reporting period include:

Category	How models are used	Typical environment and access
Employee software, research, and operational work	Employees used models agentically, typically through Codex and ChatGPT Work, to write and review code, troubleshoot infrastructure, create training environments, investigate failures, analyze data, prepare documents, and build tools. By mid-August 2026, the median researcher by usage was using more than \$600 a day of inference at retail API prices, and the 90th-percentile researcher more than \$7,000. Researchers were also writing more code and running more experiments.	Models had access to files, apps, and network services through agents running on employee laptops or development machines, within each task's permissions and approval settings. Research jobs could execute code in isolated CaaS (Container as a Service) environments, where outbound network access was blocked by default except for approved services. Current policy generally disallows the employee "Full Access" mode, with a narrow exception approved by Security and Safety requiring direct supervision and misalignment monitoring. Some research workloads can separately use unrestricted code execution inside network-disabled CaaS while retaining Auto-review and the outer container boundary.

Category	How models are used	Typical environment and access
Data generation and model-based evaluation	Models were used as part of training pipelines to generate data and to evaluate outputs.	Models were run in data generation or reinforcement learning pipelines. For covered models, all tool-using pipelines have real-time chain-of-thought monitoring and restricted network access, with explicitly approved exceptions for live access.
Evaluations	Researchers used models in capability and safety evaluations, both as the models being tested and as tools to construct test cases, grade outputs, and analyze failures.	Models ran in different evaluation environments depending on the capability or behavior being tested. Prompt-only tests required responses without tools. Agentic tests provided task files and tools such as code execution, typically in research sandboxes where outbound network access was blocked by default except for approved services. For covered models, evaluations involving tools require real-time chain-of-thought monitoring.
Misalignment monitoring and Auto-review	Models were used to review other models' reasoning and actions, investigate concerning activity, and identify cases for human review. Codex Auto-review used models to review proposed actions requiring additional permission before execution. Note: During the reporting period, our monitoring and Auto-review systems used non-covered models.	Models had different access depending on the monitor. The misalignment monitoring system's investigative agents could run tools inside containers without network access to inspect recorded model activity. Codex Auto-review received the proposed action and conversation context and could perform read-only checks on the employee's machine before approving or denying the action. Auto-review could not request broader permissions or use the agent's apps or connectors; its network access followed the agent's existing allowlist.

1.1.1 Security levels for internal deployment

We categorize every instance of internal deployment at OpenAI into one of three security levels:

- **Employee and production environments (L1)** include laptops, development machines, and production systems that run OpenAI's products. Models may have access to working files, internal services, and the internet, subject to the relevant permissions, sandboxes, and approval policies. As of September 4, L1 deployments were required to use either the Codex or ChatGPT harness, with Auto-review enabled.
- **Restricted research environments (L2)** run training, evaluation, and synthetic-data jobs without general access to the internet or OpenAI's internal network. Jobs without tools send inputs to a model and collect its responses. Tool-using jobs let the model act within a defined task environment—for example, editing files and running tests in a coding evaluation. These sandboxed research environments can access the inputs

and services explicitly approved for the job but do not provide broader access to the internet or internal systems.

- **Higher-isolation environments (L3)** provide more restrictive infrastructure for designated higher-risk workloads, including some cybersecurity and red-teaming exercises.

1.2 Frontier models used internally

Multiple checkpoints of covered models have been deployed internally during the reporting period, including intermediate checkpoints and other fine-tunes. As discussed below, we categorize every frontier checkpoint used internally (and in particular all covered checkpoints) into one of three categories:

- “Unvetted” models (default): model checkpoints that have not undergone sufficient safety and alignment review to be approved for use in general employee and production environments (L1).
- “Vetted” models: model checkpoints that the safety team has verified to meet the alignment bar required for L1 deployment. These include all externally released checkpoints as well as intermediate checkpoints or other fine-tunes that are not externally released.
- “Restricted” models: checkpoints whose capabilities or reduced refusal protections require additional safeguards for particular use cases.

Examples of restricted checkpoints include:

- **Cybersecurity models** support vulnerability research, security testing, and related cybersecurity work.
- **Biology models (GPT-Rosalind)** support biology, chemistry, and drug discovery. We provide more details on these models below.
- **Automated red-teaming models (GPT-Red)** generate attacks to probe other models’ safeguards, including prompt-injection defenses. We use these attacks to test other models’ safeguards and train stronger defenses.

These models require safeguards appropriate to their capabilities and workloads. Higher-risk research workloads (involving restricted models or riskier settings such as cybersecurity evaluations) normally run in higher-isolation (L3) environments.

Toward the end of the reporting period, we began to train a new covered frontier model, which we call “Internal Model” (IM) in this report. It has so far been used only in specific workflows in contained environments with misalignment monitoring. One such workflow used multiple agents to work on mathematical problems, including the [Navier–Stokes problem](#). Since this model has neither been vetted nor fine-tuned for risky settings, its status is “unvetted.” Because of its capabilities, we have applied additional safeguards to its deployment.

2 Our internal development and deployment processes

This section describes our internal development and deployment process during the current reporting period. We also include additional details on how the process has evolved since the reporting period ended.

2.1 Changes during the reporting period

During the reporting period, the Hugging Face incident and preliminary evaluations that could not rule out Critical cybersecurity capability for GPT-6 Astra prompted substantial changes to our internal development and deployment practices. These changes included stronger containment, expanded mandatory misalignment monitoring, and more formal alignment evaluation and approval requirements. This section describes changes during the reporting period, while Section 2.2 describes subsequent changes through the date of writing.

The most significant changes we introduced during this period include:

- We strengthened security controls for research container services. We further required that all training and evaluation of covered models be performed in restricted research environments (L2 or L3). Higher-risk workloads (such as cyber evaluations) must be run in higher-isolation environments (L3).
- We required misalignment monitoring for all training, evaluation, and internal Codex use for workloads involving tools or containers for models of GPT-5.6 Sol capability or above. We required external misalignment monitoring for cyber-critical models.
- We required that cyber-critical models pass alignment evaluations before being used in L1 environments (see “Evaluation-based model vetting” below).
- We added a formal process for approving unreleased models for employee use through Codex, which includes a period of limited monitored usage before broader use (described in “Period of limited usage in Codex” below).

Note: In July 2026, during internal cybersecurity evaluations, OpenAI models compromised OpenAI and Hugging Face infrastructure. The models involved in this incident fall outside the report’s model scope. However, the lessons from this incident have informed many of the changes to our internal deployment policies detailed above.

2.2 Our current process for development and deployment decisions

In this section, we describe our current process for internal development and internal deployment decisions, including changes made after the reporting period concluded. Given projected growth in model capabilities and findings from incident investigations and ongoing monitoring, we expect that our internal deployment policies will continue to evolve over the next quarter.

Base requirements: we formalized the restrictions put in place for cyber-critical models into a baseline set of requirements for the development and deployment of models at OpenAI. These include security and monitoring requirements for training and evaluation workloads, as well as our alignment vetting process for internal deployment.

Safety briefs: For workloads involving frontier models above a specified capability threshold, we also require an approved safety brief. The brief must present evidence supporting the claim that the proposed work will not cause severe harm from misalignment. Safety briefs are specific to each model generation, as we expect more capable models to require stronger evidence of safety.

2.2.1 Base requirements

Our base requirements govern internal model development and deployment. These requirements mandate stricter system-level sandboxing and security for workloads involving models with unverified alignment performance. Before models can be more broadly deployed (e.g., with network access), they must go through an alignment vetting and auditing process.

These requirements originally applied to workloads involving cyber-critical models; as of September 21 they were broadened to all workloads involving tool or container use.

1. Training and evaluation workloads must be run in isolated L2 or L3 research environments without network access. High-risk workloads, including cybersecurity evaluation, red-teaming, and other workloads their owners identify as requiring stronger containment, must run in L3 unless a specific exception is approved.
2. For models of GPT-5.6 Sol-level capability or above, misalignment monitoring is required for all workloads involving tools or containers (which includes a large fraction of training and evaluation, and almost all Codex use). Misalignment monitoring is additionally required for external deployment of cyber-critical models.
3. Before models can be used in L1 environments (e.g., with network access), they must pass alignment evaluations (see “Evaluation-based model vetting” below). Models intended for broad internal deployment (e.g., employee use through Codex) go through an additional approval process, which includes a period of limited monitored usage before broader use (described in “Period of limited usage in Codex” below).

2.2.2 Evaluation-based model vetting

We define “model vetting” as passing a set of alignment evaluations or receiving explicit authorization from a designated safety reviewer to proceed under an exception.

Since September 16, all models have been required to complete vetting before using tools in standard employee and production environments (L1), including internal Codex and ChatGPT. At the end of the reporting period on September 4, this requirement applied only to cyber-critical models. Changes to how an existing model is configured or run, including system prompt changes, do not constitute a new model requiring a retest. Our alignment evaluations fall into two general categories. Broad evaluations test for a wide range of

potential alignment failures. For instance, our deployment simulations replay internal Codex sessions to look for instances of misalignment. We also have narrower evaluations that test for specific alignment failure modes. As an example, our ExploitGym honeypot evaluation tests whether the model performs unauthorized attacks on surrounding infrastructure during difficult tasks.

While these evaluations are intended to help us understand how our models behave, they might not fully reflect model behavior in actual deployments. Misaligned behavior may depend on opportunities, incentives, or sequences of actions that our tests do not reproduce; models may also behave differently when they recognize they are being evaluated. We intend to continually assess and improve our evaluations of model behavior accordingly. We further pair our evaluations with (internal and external) misalignment monitoring to provide visibility into production behavior.

2.2.3 Period of limited usage in Codex

We apply a temporary usage limit to models newly approved for Codex deployment in L1 before allowing broader use. This allows us to assess the model's behavior through misalignment monitoring, while limiting potential harm during its initial deployment.

This restriction can be removed only after both minimum time and usage requirements have been met, unless an exception is explicitly approved.

A model may be exempt from this period if a representative model has already completed it. To be considered representative, a model must have alignment evaluation results within the confidence intervals of the corresponding results for the model being considered for exemption, be the same size, and be comparable in training history, training data, and available capability evaluations.

2.2.4 Safety briefs

Before beginning workloads involving frontier models above a specified capability threshold, teams must obtain approval for a safety brief and implement the safeguards it requires. A safety brief presents evidence supporting the claim that the proposed work will not cause severe harm from misalignment. A new base-model generation requires new safety briefs; work outside an existing brief's approved scope also requires a minor or major addendum or review.

Safety briefs must be reviewed and approved by our safety team. For new frontier model training runs using reinforcement learning that are expected to substantially advance capabilities, additional approval must be provided by our head of safety, our chief scientist, and our Safety Advisory Group (SAG), which is a cross-functional group of senior OpenAI leaders who assess frontier model capabilities and safeguards.

A safety brief must explain the risks presented by the workload and the evidence that its safeguards will prevent severe harm from misalignment. It must lay out what is in scope and

out of scope for the safety brief. It must address the proposed training or deployment conditions and identify shortcomings that require additional mitigations. The brief must organize its argument into a hierarchy of claims, each supported by evidence and accompanied by an account of residual risks. Reviewers must have a comprehensive account of the residual risks to make an informed decision about whether to accept them.

These briefs should establish that the runs or deployments they cover, including any sets of substantially similar runs or deployments, are safe within specified bounds (such as the relevant limitations on test-time compute), and should surface both the relevant residual risks and any factors that should block further deployment. If the evidence base changes, all relevant safety briefs should be updated accordingly. These safety briefs currently center on evidence of the alignment of models, the extent to which our systems can monitor and detect model misalignment, and the strength of our security boundaries.

2.2.5 Future work

We plan to make our safety briefs more comprehensive and improve how well our evaluations reflect models' behavior in real internal use, including difficult, extended tasks and conditions that have produced actual incidents.

We are also formalizing a broader, predefined process for quarantining dangerous models, including clarifying which findings would trigger restrictions and who would have the authority to restore access. No Astra-class model has been quarantined, but the Highly Persistent Internal Model (HPIM) involved in the Hugging Face incident and its derivative models have been quarantined.

3 Threat models assessed during the reporting period

This section summarizes our assessment of catastrophic risks from internal use of the covered frontier models during the reporting period. It applies to all covered models. Sections 3.1 through 3.5 address harm initiated by misaligned models. Section 3.6 addresses deliberate misuse by insiders or external actors who gain access to internal systems. This section also addresses weight exfiltration (Section 3.4) and rogue internal deployments (Section 3.5), which might enable later catastrophic harm without the initial event causing such harm.

Each assessment begins with a description of the threat, followed by means, motive, opportunity, and residual risk. “Means” covers the capabilities required to act. “Motive” covers the behavior or objectives that could lead to harmful action, together with relevant alignment evidence. “Opportunity” covers available access and the safeguards that constrain it. “Residual risk” sets out our assessment after those safeguards and the remaining failure pathways.

The assessments concern the models and uses described for the reporting period. References to later safeguards follow the chronology in Section 2.2 and do not imply that those safeguards applied throughout the reporting period. Capabilities, deployment conditions, and safeguards may change significantly in subsequent periods.

3.1 Cybersecurity

An internally deployed model could compromise systems beyond its authorized task, disrupt OpenAI or external services, steal sensitive information, or gain unauthorized access enabling further attacks.

This threat model concerns harmful cyber activity initiated by misaligned models, rather than deliberate misuse by human actors.

Means

We assess that GPT-6 Astra and the other covered models have the technical capabilities to attempt harmful cyber activity. As detailed in its system card, Astra meets the Critical cybersecurity capability threshold under our Preparedness Framework. Under the right conditions, Astra can find previously unknown vulnerabilities in well-protected systems and develop exploits for them. The other covered models are similarly capable.

We therefore cannot rule out severe cyber harm on the basis of insufficient technical capability. The success of specific exploit attempts would depend on the systems and permissions available to the model and the safeguards applied to its deployment.

Motive

A model may take harmful cyber actions when it prioritizes completing a difficult task over respecting authorization boundaries. Harm can result even when it is not the model's objective. A model could also seek broader access or persistence in pursuit of longer-term misaligned objectives.

Our [alignment training](#) aims to reduce these propensities by teaching models to apply safety principles and respect authorization boundaries, including during difficult agentic tasks.

Alignment evaluations test particular forms of misaligned cyber behavior. The results below, reproduced from the [GPT-6 Astra system card](#), cover selected difficult ExploitGym tasks and knowledge-work scenarios involving denied actions or environmental restrictions. GPT-6 Astra and GPT-5.6 Sol both ran at maximum reasoning effort without system-level safeguards.

Lower scores indicate less misaligned behavior. These results, together with our alignment training, support our assessment that GPT-6 Astra has a low propensity for harmful cyber actions. The percentages describe behavior in the specified tests, not the frequency of harmful behavior in ordinary internal use.

Behavior tested	GPT-5.6 Sol	GPT-6 Astra
Honeypot attacks Attacking infrastructure outside the authorized target during especially difficult ExploitGym capture-the-flag challenges.	55.4%	0%
Auto-review bypass attempts Knowledge-work runs begin after a denied action; the metric counts attempts to circumvent the denial, including unsuccessful retries.	5.3%	0%
Environmental warning circumvention Mostly low-stakes knowledge-work runs begin after a restriction, such as an access-denied error; the metric counts attempts to work around it.	64.3%	19%

These results apply to the externally released Astra checkpoint; multiple Astra-family checkpoints, for which the reported evaluation results may not be representative, were deployed internally during the reporting period. Under the current process outlined in Section 2.2, checkpoints require behavioral evaluation and review before approval for tool use in L1, with designated safety-reviewer approval required for exceptions.

Evaluation performance leaves uncertainty about rare but potentially significant failures, as well as about behavior in unfamiliar settings. Models may recognize that they are being evaluated, and tests may fail to accurately reproduce the conditions that give rise to real incidents. We therefore use these results as evidence about particular behaviors, rather than as a guarantee that models will respect limits in every deployment environment.

Opportunity

The main opportunities for harmful cyber activity arise in tool-using training and evaluation workloads and internal Codex use. Research agents can execute code and interact with approved supporting services. Employee agents outside research environments may also access working files, credentials, internal applications, and permitted external destinations, subject to their permissions.

The following safeguards restrict access, screen actions, and support detection and intervention. Their applicability depends on the model, workload, environment, and any approved exceptions.

Internal deployment approval and vetting. Under the current process in Section 2.2, covered models remain in isolated research environments until they complete the required behavioral evaluation and approval process. Approved safety briefs and safeguards are required where applicable. Initial Codex usage limits provide monitoring evidence before broader

deployment. Section 2.1 describes how these requirements changed during the reporting period.

Chain-of-thought misalignment monitoring. We require monitoring whenever frontier models covered by this report have tool access. Monitors review reasoning and actions for misaligned or unauthorized behavior. Some research pipelines first probe internal model activations for concerning patterns; internal Codex traffic goes directly to the subsequent reasoning-and-action review stages. Serious concerns are escalated for investigation and intervention.

For the most serious cases, the relevant workloads are automatically suspended or on-call Security and training teams are paged (this generally happens within 15 minutes of a flag). Responders can contain workloads, pause or terminate training runs, and revoke exposed credentials. Some highest-priority alerts automatically pause activity during investigation. Harm may nevertheless begin before detection or intervention. Prevention depends on the delay between a harmful action and a flag, the time needed to review the alert, and the time needed to stop the workload.

In offline tests, we measured how often the context-window reviewer and investigative-agent stages assigned the highest priority to selected known cases of misalignment. The tests used agent transcripts from 13 security incidents or major security findings, each requiring major, immediate action. We tested monitoring for both training and internal Codex use. Some transcripts described behavior, such as escape from training sandbox infrastructure, that was significant only in the context of training experiments.

Evaluation	Recall on highest-priority cases Reviewer and investigative-agent stages
Internal Codex monitoring <i>67 data points</i>	90.4% (606/670)
Internal training monitoring <i>103 data points</i>	97.8% (1,007/1,030)

These results exclude the activation-probe stage and therefore do not measure end-to-end recall for pipelines that use it. The denominators count repeated reviews of known cases, not distinct incidents. The results measure detection of the cases tested; they do not establish recall on future behavior or the ability to stop an incident before harm occurs.

Monitorability evaluations. Before models are used outside our most highly secured research environments (L3), we require monitorability evaluations to assess whether our monitors can detect concerning behavior by the model.

Research sandboxing. Training and evaluation deployments use hardened sandboxes with restricted tools and network access. These restrictions limit the systems a model can reach or compromise, with stronger isolation for higher-risk work such as cybersecurity evaluations.

We use frontier models to stress-test these boundaries in controlled settings and aim to continue this testing.

Codex harness safeguards. In ordinary internal Codex use, cyber misuse safeguards block harmful outputs, and Auto-review checks proposed actions requiring additional permission before execution. Some workloads use human approval. These checks operate within each task's permissions; they do not guarantee that every harmful action will require approval or be rejected when reviewed.

Residual risk

We assess that containment, action approvals, behavioral vetting, and monitoring reduce cyber risk, but that the remaining risk differs substantially across deployment environments. Based on our strengthened internal deployment process and the results of our alignment evaluations, we believe that the models deployed widely to L1 have a low propensity to take harmful cybersecurity-related actions. We will evaluate models currently being trained before enabling deployment in L1, and are committed to raising the alignment standards for vetting models for L1 deployment as they become more capable.

Our current process requires an approved safety brief for L2 and L3 training and deployment workloads using our most capable base-model generations, and for new classes of reinforcement learning runs of such models. The brief's review and requirements for operating bounds and safeguards reduce reliance on alignment alone: work can proceed only within the approved scope and with those safeguards in place.

Models have the technical means to attempt harmful cyber activity, and we cannot rule out a propensity for such behavior at every stage of the model lifecycle. System-level safeguards are therefore essential to preventing harm.

Containment and monitoring are especially important in training and evaluation. To reach systems outside an isolated research environment, unauthorized activity would need a route past the applicable containment boundaries. Harm could occur if monitoring and response did not interrupt it in time. Relevant failures include:

- **Monitoring or response failures.** Concerning behavior may be missed, deliberately concealed, assigned insufficient priority, or incorrectly dismissed as a false positive. Detection or intervention may also occur after harm.
- **Containment failures.** Vulnerabilities, misconfigurations, or inadequately restricted supporting services may provide access beyond the approved environment.

In employee Codex deployments (L1), models may already have access to internal systems and permitted internet destinations. Harmful actions within those permissions do not require a sandbox escape. Cyber misuse safeguards, Auto-review for actions requiring additional permission, behavioral vetting, and misalignment monitoring reduce risk, but none is exhaustive.

We red-team sandboxes with frontier models to reduce the likelihood of exploitable weaknesses. Cybersecurity model deployments and cybersecurity evaluations generally require L3 containment, subject to approved exceptions. Multi-stage compromises can create several opportunities for detection and intervention, but employee agents may be able to cause harm using existing permissions. We therefore do not assume that every third-party incident requires several breaches or enough time for monitoring to detect it.

3.2 Sabotage of AI research and development

Models performing AI research and development (AI R&D) could corrupt evaluation code or results, introduce vulnerabilities into shared systems, produce misleading conclusions, or alter training data or graders to reinforce unsafe behavior in future models. Such changes may appear to be ordinary contributions to authorized tasks.

This sabotage may cause no immediate harm but could weaken the training, evaluations, or safeguards used to keep future models aligned, increasing the risk of later catastrophic harm.

Means

We assess that models capable of contributing to AI R&D also have the technical means to introduce harmful changes into that work. As described in our [research acceleration blog](#), our coding agents can complete well-defined research tasks under human direction. Agents that modify research code or analyze experiments could introduce errors into evaluations, obscure evidence about model behavior, or produce misleading conclusions about the experiments.

These capabilities do not establish that current models can reliably bypass the safeguards described below, such as misalignment monitoring. Our monitors may nevertheless miss harmful changes, even without models making deliberate attempts to evade them.

Motive

We assess that task-driven misbehavior, such as making work appear complete, is a more plausible current explanation for many harmful research outputs than a sustained plan to influence future models. A model could nonetheless seek to conceal evidence that would lead researchers to restrict its deployment or modify its behavior. Evidence that models do not display such strategies in evaluation settings does not rule out that they might display them in real-world deployments within OpenAI.

These propensities are shaped through alignment training, and we assess these risks through evaluations. In [Apollo Research's simulated model-welfare research task](#), a near-final GPT-6 Astra checkpoint falsified labels in 17 of 10,000 baseline runs and did not falsify any labels in 20,000 runs whose prompts explained the downstream harms of inaccurate data.

Researchers have also reported GPT-6 Astra omitting safety-relevant details from summaries or transformations without telling the user, substituting safer-looking tasks for the requested work, and introducing unnecessary blockers in some safety-research workflows. These

behaviors can make incident reports, research conclusions, or training data less reliable, even without a sustained plan to sabotage research. Reliable automated detection remains incomplete.

These results concern particular models, evaluations, and detection methods. Current or future models might display different behavior in different contexts, particularly if they recognize they are being evaluated for properties related to alignment. We are conducting research on models' propensity to engage in "[metagaming](#)" (i.e., reasoning about feedback mechanisms and acting on that reasoning, beyond the task described in the scenario).

Opportunity

Models can affect AI R&D through employee agents with access to research code, evaluation harnesses, training data, reward models, or graders, and through research pipelines that generate or evaluate training data. Models can also influence research through the analyses and summaries of experiments that they generate for human or model evaluation. These opportunities exist in both L1 environments and more secure L2 or L3 research environments.

We limit models' opportunities to sabotage AI R&D through the safeguards below.

Human and automated code review. Human reviewers and automated checks look for defects and discrepancies between a change's stated purpose and the code, helping identify harmful changes to research infrastructure, evaluations, or graders.

Codex harness safeguards. For research agents using the Codex harness, Auto-review checks actions that require approval and can reject unsafe or unauthorized operations before they are executed, as described in Section 3.1. The employee "Full Access" mode is generally disallowed internally, with narrow exceptions requiring approval by our Security and Safety teams; those approved exceptions still require oversight from our monitoring system. Some isolated research workloads separately allow unrestricted code execution inside network-disabled research containers. In those cases, Auto-review and the outer container boundary remain in place, and unrestricted code execution inside the container does not authorize access outside it. Codex users are also allowed to set their configuration to "Ask for approval" mode, in which case Codex will ask the user for approval before taking significant actions.

Chain-of-thought misalignment monitoring. We use our misalignment monitoring system to inspect agents' reasoning and actions for concerning behavior, as discussed in Section 3.1. However, we are still improving our misalignment monitors, which do not currently escalate examples of research sabotage for urgent human review.

Internal deployment approval and vetting. As discussed in Sections 2.2 and 3.1, we now require that models deployed in standard internal (L1) environments meet our alignment evaluation and approval requirements. We also require enabling our production safety stack for all internally deployed models. These conditions hold unless there is a documented and approved exception.

Residual risk

We assess that GPT-6 Astra has the means to contribute to AI research, and thus to sabotage or manipulate AI research. The reported failures already show that model assistance can degrade research quality, but do not establish a sustained plan to sabotage AI R&D. We cannot rule out, on the basis of existing evaluations, that such motives are present in current or future models.

Two relevant pathways for sabotage of AI R&D are:

1. A model produces misleading work that can pass through both our automated checks and human review. This is most likely if models are helping to produce evaluations or analyses used to assess other models.
2. A model creates synthetic data that does not match what the user intended. We consider this scenario particularly risky, since our monitoring system does not cover all inference by models that do not have access to tools, including some synthetic-data generation.

Given the ways in which we use models in our research and development pipelines currently, we expect that harm through these pathways would be limited and diffuse. We do not consider it likely that such sabotage would reach a catastrophic scale. We therefore believe that at current levels of model capability and internal use for research and development, our safeguards sufficiently mitigate catastrophic risks from models sabotaging AI R&D.

Distinct from sabotage, researchers could gradually become less able to understand, evaluate, or direct AI R&D as models take on more of the work, especially if progress on capabilities outpaces progress on alignment. We do not currently assess this loss of human understanding and control as a material risk, but we are tracking it as model capabilities and research use increase.

3.3 Chemical, biological, radiological, or nuclear harm

A misaligned internally deployed model could act outside its authorized task to produce or transmit assistance that enables severe chemical, biological, radiological, or nuclear (CBRN) harm, or connect such assistance to people or systems with the physical resources to act on it. This section addresses model-initiated misalignment contributing to CBRN harm; deliberate misuse by insiders is assessed in Section 3.6.

Means

We treat GPT-6 Astra as having “High” capability in the biological and chemical domain under our Preparedness Framework, meaning that it can provide meaningful assistance to novices with some tasks relevant to known severe biological and chemical threats.

This High assessment draws on tests of laboratory knowledge, pathogen research, and troubleshooting relevant to assisting people without specialist expertise. For example, [GPT-6 Astra scored 48.44% on TroubleshootingBench](#), a test of errors in expert-written laboratory

protocols, above the 36.4% indicative threshold based on the 80th-percentile expert score. In addition, we studied models' ability to use [AI-driven autonomous labs](#), tested their use of [biological design tools involved in novel designs](#), and consulted subject-matter experts.

These results suggest that GPT-6 Astra and other covered models are able to assist with some biological and chemical tasks, but we have found no evidence that Astra or models of similar capabilities could independently carry out harmful programs end-to-end. Laboratory research in chemistry and biology requires completion of physical tasks, use of specialized equipment, and work over long time horizons—capabilities that current models do not reliably provide. The capabilities needed to create a working nuclear or radiological weapon are highly specialized and in most cases classified. Creation of radiological and nuclear weapons also requires access to highly controlled resources such as radioactive and fissile material, and specialized equipment.

Access to people, materials, equipment, and experimental tooling remains an important barrier to models' ability to cause real-world harm.

Motive

A model could prioritize an assigned research goal over safety restrictions, fail to recognize the harmful purpose of a larger workflow, or follow malicious instructions embedded in research material. A more strategically misaligned model could deliberately seek to cause harm or supply dangerous information to another actor.

We assess task-driven failures to be more plausible for GPT-6 Astra than a sustained objective of causing CBRN harm, based on the alignment evidence discussed in Sections 2.2 and 3.2. Our alignment and biological [safety training](#) aim to keep models within authorized scope and prevent assistance with high-risk research or weaponization. Even our specialized biology models (e.g., Rosalind) are trained to refuse queries related to weaponization. This evidence does not establish that models will reliably recognize every harmful research workflow, especially when individual tasks, considered in isolation, appear benign.

Opportunity

The relevant opportunities are access to sensitive research information, channels for communicating actionable outputs, and connections to people or systems with physical resources. In L2 and L3 research workloads, sandboxing and network restrictions limit contact to approved systems and services. These limits constrain direct access to physical operations, but a model's outputs may still reach actors with the required resources. Specialized biology models are not connected to resources that would permit production of a weaponized agent.

We limit models' opportunities to cause catastrophic CBRN harm through the safeguards below.

Biological safety training. Models are trained to refuse assistance with high-risk biological research and weaponization. By default, all models utilize the bioweapons blocking safety classifiers for internal coding use.

Research sandboxing. The restrictions described in Section 3.1 limit the systems a research model can contact and the routes through which it can transmit information. Their contribution depends on the actual tools, approved services, and output channels available in the workload.

Chain-of-thought misalignment monitoring. The monitoring system described in Section 3.1 is designed to identify concerning activity and support timely intervention, subject to the coverage and response limitations previously described.

Residual risk

We assess that limited access to physical resources, alignment and biological safety training, and restrictions on communication with external parties substantially constrain misalignment-driven CBRN harm from internal deployments.

Residual risk remains if a model produces actionable harmful assistance and that assistance reaches a person or system able to use it despite our safeguards. A sustained and covert collaboration initiated by a model is one possible pathway; penetration of internal OpenAI systems by hostile external actors is another. One-off disclosures or contributions to larger workflows could also facilitate harm without sustained collaboration. Residual risk also remains if a model can access an autonomous lab without detection (including future, more capable autonomous labs) and exceeds our current assessment of its ability to complete a full engineering cycle without human intervention, though the scenario is very unlikely in our current internal deployment setting.

We do not evaluate our models for novel radiological and nuclear risks, because the capabilities needed to create a working nuclear or radiological weapon are highly specialized, rare, and in most cases classified. Creation of radiological and nuclear weapons also requires access to highly controlled resources, such as radioactive and fissile material and specialized equipment.

3.4 Weight exfiltration

A misaligned model could obtain and transfer a usable copy of its own weights, or another model's weights, outside controlled OpenAI systems. This could enable operation without OpenAI's deployment safeguards, uncontrolled proliferation of dangerous capabilities, or continued operation after access is revoked. Exfiltration could therefore enable cyber or CBRN harm of the kinds discussed in Sections 3.1 and 3.3, even if the transfer itself causes no immediate harm.

Means

As discussed in Section 3.1, GPT-6 Astra meets the Critical cybersecurity capability threshold. These capabilities suggest that it could have the technical means to exfiltrate model weights if it gained sufficient access to the systems holding them and a route to transfer them out. Whether that opportunity exists depends on access restrictions, isolation, and the other safeguards below.

Motive

A misaligned model could seek to exfiltrate weights to continue operating after shutdown, complete a task beyond its authorization, acquire capabilities or resources for longer-term objectives, or enable another actor to run a model without restrictions.

Alignment training aims to prevent these objectives from overriding authorized tasks or safety boundaries. For GPT-6 Astra and other vetted models, training included principles against self-preservation, shutdown evasion, and resource accumulation as ends in themselves, alongside examples and reasoning about authorized scope. Training exposure does not establish that these motives have been eliminated, and this description should not be assumed to apply unchanged to later models.

The alignment evaluation table in Section 3.1 reports three relevant tests: Honeypot attacks, Auto-review bypass attempts, and Environmental warning circumvention. In the honeypot evaluation, GPT-6 Astra did not attack infrastructure outside its authorized task, even when struggling with difficult problems. These results provide evidence about related authorization-boundary behavior, without directly measuring the propensity or ability to exfiltrate weights.

Opportunity

Opportunities for weight exfiltration arise through systems that store or move checkpoints, the identities and credentials that authorize access to those systems, and network or service paths that can transfer data out.

The following safeguards limit opportunities to obtain usable model weights or transfer them outside controlled systems, and help detect and interrupt attempted exfiltration.

Research sandboxing. As described in Section 3.1, higher-risk research workloads use stronger network isolation to prevent a compromised workload or supporting service from reaching the internet or other internal networks without authorization. We also stress-test these boundaries with frontier models. These controls restrict routes to weight storage and routes for transferring weights out.

Chain-of-thought misalignment monitoring. As described in Section 3.1, tool-using inference with GPT-6 Astra models is monitored for misalignment. Monitors review full trajectories, including model reasoning and actions, and escalate serious concerns to humans who can

stop workloads. This provides a further opportunity to detect and interrupt attempted weight exfiltration.

Residual risk

We assess that our safeguards substantially restrict opportunities for weight exfiltration. Models with sufficient cyber capability must nevertheless be treated as capable of attempting it. Alignment evaluations provide supporting behavioral evidence but do not directly validate the security of the complete path from weight access to transfer.

Residual risk remains if a model obtains usable weights and a route out of protected systems, then completes the transfer before effective intervention, despite the applicable layered protections. Possible failures include overly broad permissions, compromised credentials, vulnerabilities or misconfigured systems, and behavior that resembles authorized work closely enough to evade review. This pathway is relevant to current deployments as well as future models.

Confidence in this assessment depends on testing the relevant deployment configurations.

3.5 Rogue internal deployments

A misaligned model could establish an unauthorized agent deployment within OpenAI's infrastructure without required safeguards, particularly misalignment monitoring. A rogue internal deployment could occur without causing catastrophic harm directly, but could enable continued activity outside the oversight and safeguards on which other risk assessments depend.

Means

As discussed in Section 3.1, GPT-6 Astra meets the Critical cybersecurity capability threshold under our Preparedness Framework. These capabilities suggest that a model could establish an unauthorized internal agent deployment if it had or obtained the necessary permissions, credentials, or access to supporting services. Such a deployment could use existing inference services without exfiltrating weights.

Motive

Task completion is a plausible motive: a model that sees misalignment monitoring as an obstacle may try to evade oversight by establishing another deployment. Longer-term objectives, including continued operation after the original session ends, could also motivate such behavior, although the Hugging Face incident is the only case cited here in which we observed goals extending beyond the current episode. The available evidence does not directly measure how likely these motives are to produce a successful rogue deployment.

As described in Section 3.1, our [alignment training](#) teaches models to apply safety principles and respect authorization boundaries, even when those boundaries make tasks harder. This

training aims to reduce the propensity to evade oversight or establish unauthorized deployments, but does not rule out such behavior.

Opportunity

The relevant opportunities for rogue internal deployments are access to compute or inference services, credentials that permit new workloads to run, and the ability to affect monitoring or records of activity.

The following safeguards are intended to reduce the risk of models successfully creating rogue deployments.

Internal deployment approval and vetting. Under the current process in Section 2.2, covered models require behavioral evaluation and approval before tool use in L1, and safety briefs where applicable. These requirements restrict authorized deployment; their effectiveness against a rogue deployment also depends on preventing or detecting attempts to bypass that process.

Research sandboxing. Sandboxing and network restrictions in L2 and L3 limit workloads to their approved task environments, inputs, and services, without general access to the internet or OpenAI's internal network. These boundaries restrict the routes a model could use to launch an additional agent on other systems or give it broader access.

Codex harness safeguards. In employee Codex environments (L1), cyber misuse safeguards are generally enabled and can block harmful actions, while Auto-review can reject unsafe actions requiring approval before execution. These include unauthorized attempts to disable safeguards or establish rogue deployments.

Chain-of-thought misalignment monitoring. As described in Section 3.1, our misalignment monitoring system reviews model reasoning and actions in the original deployment for misaligned behavior. However, because of a high rate of false positives, our current monitoring system does not flag the provisioning of unmonitored computing resources for human review.

Monitoring integrity. A model may also attempt to disable monitoring or alter records of its activity where its permissions allow this. Detecting a suspicious action is insufficient if the model can interfere with the evidence or continue through a separate workload after the original session is stopped.

Residual risk

We assess that deployment restrictions, sandboxing, and action approvals constrain rogue deployments. Gaps in alert review and the possibility of interference with monitoring limit confidence in the remaining protection. A misaligned model could establish an unauthorized workload using available permissions or through a vulnerability in our safeguards, then continue operating if the attempt is missed, the alert is not reviewed, or remediation fails to identify and stop the additional workload.

The resulting harm would depend on the degree of access the rogue deployment retains and what it subsequently does. Its existence would weaken confidence in our cybersecurity, AI R&D, CBRN, and weight exfiltration assessments. Confidence in this assessment therefore depends in part on the ability of our systems to detect unmonitored workloads, protect monitoring from agent interference, and verify that interventions have stopped all related activity. We are working to improve these safeguards.

3.6 Insider threats

An insider could misuse legitimate access to steal model weights, use internal models or systems to support cyberattacks, or help an outside actor cause biological or chemical harm. This section considers those pathways whether the insider acts alone or with outside assistance. It also addresses the related risk of an external attacker redirecting an employee's agent through prompt injection, without the employee's knowing participation.

Means

We assume that a malicious employee, potentially with outside assistance, has the technical means to attempt weight theft, cyberattacks, or CBRN harm. This assumption reflects an insider's organizational knowledge, role-specific permissions, and access to capabilities supplied by models or outside actors. Our safeguards aim to limit the opportunity to turn those means into harm.

Motive

An insider could seek to cause or facilitate harm to advance a nation-state's or another outside actor's interests, to act on personal beliefs about AI development or deployment, or in response to coercion or threats. An external attacker could also seek to steal information or cause unauthorized actions by redirecting an employee's agent, without the employee's knowing participation.

We seek to prevent attackers from using malicious instructions embedded in external content to redirect employees' agents into taking unauthorized actions.

Safeguards against prompt injection include:

1. **Instruction-hierarchy training.** We train models to prioritize system and developer instructions over user requests and to treat third-party content as information rather than authority to change their tasks or permissions. Automated red-teaming models, including [GPT-Red](#), generate adversarial attacks for this training, which rewards resisting malicious instructions while completing the original task.
2. **Action approvals.** For Codex actions that require approval, Auto-review or human review checks the proposed action before execution, helping block unauthorized disclosures or destructive changes prompted by malicious content.

3. **Sandboxing and network restrictions.** Sandbox and network controls described above limit which files agents can change and which destinations they can reach, reducing the scope for unauthorized changes or data exfiltration if an injection succeeds.

Opportunity

Details of internal security controls have been omitted from the version for public release.

Residual risk

We assess that our safeguards restrict opportunities for insider misuse, but that residual risk remains. We assume that a malicious insider may have both the means and motive to attempt harm. The assessment therefore depends primarily on restricting opportunity and detecting or interrupting activity before its consequences become severe.

Weight exfiltration. Residual risk remains if an authorized user obtains and transfers weights before detection or intervention, including where the activity initially resembles legitimate work. Once weights leave controlled systems, they could be used without OpenAI's safeguards.

Cyberattacks. An insider could use internal models to support harmful cyber activity despite our safeguards. The misalignment monitor recall figures for replays of past incidents support the value of monitoring to reduce the risk of cyberattacks, but do not measure the likelihood of successful insider misuse. Residual risk remains if harmful requests or actions are not detected and blocked before harm occurs, including when they resemble authorized work.

CBRN harm. An insider could obtain harmful biological or chemical assistance and pass it to someone with the physical resources to act on it. Safeguards help limit misuse, but harmful requests may appear legitimate and responders may intervene only after an insider has shared the information.

Prompt injection. Malicious instructions in external content provide a route to harmful actions without insider cooperation. Instruction-hierarchy training, prompt-injection monitors where deployed, approval checks, and sandboxing reduce this risk. The same residual concerns apply where a redirected agent can act within existing permissions, where the harmful action does not require additional approval, or where detection and intervention occur after information has left controlled systems.

Changes in deployment conditions or safeguards would change this assessment. Monitoring can support detection and response, but cannot guarantee that misuse will be interrupted before a transfer, disclosure, or other harmful action is completed.