

OECD AI Transparency Report

Organization: OpenAI Opco, LLC (USA)

Reporting Period: Q3 2026

Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Model developer (/provider)
- Application developer (/provider)
- Deployer

If any of the fields above are selected, please indicate your organization size:

Large enterprise: 250 or more employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Define enhanced safety measures for specific dangerous capability thresholds, that, once reached, require safeguards (e.g. capability threshold for dangerous Chemical, Biological, Radiological and Nuclear (CBRN), cybersecurity attacks, dangerous advanced autonomy capabilities, or other capabilities).
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): For example, ISO/IEC 42001; the NIST AI Risk Management Framework; and the EU General-Purpose AI Code of Practice, as described in OpenAI's Frontier Governance Framework.
- Consult with external stakeholders (e.g., from academia, industry groups, civil society, or standard development organisations) when developing or updating risk taxonomies

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe: • Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk; • How mitigations are graduated by capability or propensity level; • How your organisation considers the specificities of different AI technologies, such

as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

OpenAI assesses severe frontier AI risks under its published [Preparedness Framework](#) and [Frontier Governance Framework](#). The Preparedness Framework establishes risk-specific High and Critical capability thresholds and corresponding safeguards. Covered systems reaching High capability require safeguards that sufficiently minimize the associated risk of severe harm before deployment; systems reaching Critical capability also require such safeguards during development. Capabilities Reports and Safeguards Reports inform review by the Safety Advisory Group, which advises OpenAI leadership. The Frontier Governance Framework describes how relevant safety and security practices align with covered regulatory risks, including risk assessment and mitigation, transparency reporting, security risk management, and incident response.

These frameworks describe capability- and threat-model-specific criteria rather than a single numerical threshold across all AI systems. OpenAI also maintains ISO/IEC 42001 certification for its AI management system and ISO/IEC 27001 certification for in-scope information-security management systems. Our [security and privacy page](#) describes the scope of these certifications, which complement the frontier-specific risk governance described above.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Internal red teaming³
- Independent third-party red teaming
- Adversarial testing to detect technical vulnerabilities
- Publicly available benchmarks and measurement tools
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Specific risk identification practices for vulnerable groups
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- External model, system, or application-independent research initiatives
- Vulnerability reports
- Incident reports, including reports shared by other organizations
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): evaluation of tool use, prompt-injection risks and actions within the deployment environment.
- Assessment of risks arising from interactions among multiple AI agents

Are any national AI safety / security institutes of the AISI network involved?

Yes

How does your organization monitor the interactions of AI agents under your control with other agents / assess risk from multiple agents interacting?

We use monitoring and targeted alignment evaluations to identify potential misalignment in agentic systems, including risks that may arise when agents interact with other systems or agents. Findings inform further investigation, testing, and improvements to safeguards.

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

OpenAI employs rigorous red-teaming processes to evaluate the performance of models and systems before external deployment. OpenAI uses both human and automated red-teaming processes. Adversarial testing involves experts who attempt to exploit the model in ways that could cause harm or lead to undesirable outcomes, or who probe the model for unexpected or unwanted behavior in a variety of domains. This process is complemented by collaborations with external organizations, such as cybersecurity firms, to provide additional perspectives and expertise in identifying potential threats. OpenAI's red teaming efforts include a variety of expert testing organizations that help inform our risk assessment and mitigation efforts. These organizations span a variety of domains, with diverse perspectives.

Our model review process is also informed by [third-party assessments](#). Since our previous submission, we have published more detail on three forms of external testing: independent capability and safeguard evaluations, reviews of evaluation methodology, and direct probing by subject-matter experts. These assessments supplement our internal work and inform deployment decisions.

We use regression testing to make sure models are robust against previously discovered jailbreaks, and perform new rounds of automated red-teaming with our latest internal red-teaming attackers to validate our improvements.

Red teaming and adversarial testing are used to uncover failure modes, misuse potential, or unexpected behaviors. These findings often lead to downstream mitigation efforts — such as refining instructions, retraining, or introducing product-level safeguards.

The results of these testing activities are documented and reviewed by relevant stakeholders (e.g., model developers, product teams, risk reviewers), and actions are prioritized based on impact, feasibility, and alignment with risk thresholds.

The Safety Advisory Group reviews Capabilities Reports and Safeguards Reports, assesses residual risk and makes recommendations to OpenAI leadership, which makes deployment decisions. Findings can lead to further evaluation, stronger safeguards, access restrictions or a phased release.

We use a combination of quantitative (e.g., internal benchmarks, as well as third-party testers' evaluations) and qualitative (e.g., subject-matter expert judgment, preparedness review outcomes, stakeholder feedback) metrics to assess AI-related risk. These are embedded in our risk evaluations. We describe some of these metrics, methodologies, and limitations in system cards and in research we publish.

It is worthwhile to note that some risks (e.g., reputational or ethical harm) resist easy quantification, and we lean on expert judgment. AI system behavior can change over time (e.g., through fine-tuning or

model evolution), so risks may need ongoing validation. Additionally, risk evaluations are context-sensitive — what is lower-risk in one setting may be higher-risk in another. Risk evaluations are continually re-evaluated based on changes to the threat landscape, and in response to adverse, anomalous, or malicious activity observed on our platform.

For agentic systems, evaluation results also depend on the tools, software environment, scoring method and resources available during testing. We check for issues such as prior exposure to test material, invalid tasks and shortcuts in scoring, and interpret results within the conditions tested.

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Reports of incidents and vulnerabilities help us understand risks that may only become clear once a system is in use. We consider reports from users, researchers and other organizations alongside information from our own monitoring. When a report arrives, we assess which systems may be affected, how serious the issue could be and what response is needed.

The investigation can lead to changes in safeguards, model behavior, monitoring or enforcement. It can also prompt us to revisit an earlier assessment of risk. For significant incidents, we document what happened, investigate the causes and record the actions taken. Reviews after an incident help the relevant teams carry those lessons into future development and deployment.

Where appropriate, we share information with affected stakeholders or relevant authorities so that they can respond as well. The information shared takes account of security, privacy and confidentiality.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	
Supervisory authorities (e.g., data protection authorities)		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)	☐		

	Always	Ad hoc / as relevant	Not currently accessible
Supply-chain partners (cloud, data providers, etc.)	☐		
Academia / researchers	☐		
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)		☐	
Specific civil society organizations		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

We make different reporting routes available because reports can involve different kinds of problems and sensitive information. The public bounty programs provide an ongoing route for researchers to report issues that fall within the relevant program's scope. For security incidents, the coordinated vulnerability disclosure policy directs reporters to an encrypted channel so that information can be shared securely.

The same policy brings together the rules and instructions for reporting, including the processes for vulnerabilities in third-party software and for assigning CVE identifiers. Once a report is received, further coordination depends on the organization affected, the seriousness of the issue and the information needed to investigate or fix it. This lets us keep public reporting routes accessible while handling sensitive details with the appropriate parties. Reporting instructions, program scope, and the encrypted security-incident reporting channel are available in OpenAI's [Coordinated Vulnerability Disclosure Policy](#).

The separate model-misalignment reporting framework is an employee-initiated escalation and public-disclosure process, not an external submission route. Any OpenAI employee may flag a qualifying example for investigation by the safety and alignment teams. Technical staff assess what happened, remaining uncertainties, whether disclosure is warranted, and any third-party impact; cases follow Ready for Disclosure, Minor Investigation, or Larger Investigation tracks. Disagreements about disclosure or track are referred to the Safety Advisory Group and, if necessary, OpenAI leadership. Where a third party is affected, security, legal, and responsible-disclosure obligations take precedence. We intend to give advance notice if a report would identify a third party, and publication may be delayed for security reasons. External reporters should continue using the applicable bounty and coordinated vulnerability-disclosure channels.

How does your organization share relevant information as appropriate?

- Publicly
- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

Depending on the issue and audience, OpenAI shares descriptions of vulnerabilities or incidents, affected systems, risk-reduction steps, investigation findings and coordinated-disclosure information. These may be communicated through public reports, private notification of affected parties, existing information-sharing partnerships, and applicable regulatory disclosures.

For model misalignment, each full report describes the observed behavior, severity and any external impact, the setting and date or date range, when the behavior was discovered, and the model or models involved at a high level. Where possible, it also covers resulting harm, how the issue was found, the scope of the investigation, implications for AI safety, unanswered questions, and mitigations taken or planned. Publication need not wait for an investigation or fix to be completed. For complex cases, an initial notice may summarize what happened, whether outside experts are assisting, and any available estimate for a final report; publication may be delayed for security reasons. We intend to provide identifiable third parties with advance notice, and disclosures involving customers remain subject to privacy and contractual obligations.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Paid rewards (“bug bounty”)
- Vulnerability Disclosure Programs (VDPs) with safe harbor / liability protection
- Recognition and reputation incentives
- Early access and collaboration incentives
- Research grants, contracts, and hiring red teamers
- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write “Not applicable.”

OpenAI maintains security and safety bounty programs through Bugcrowd, offering rewards and recognition for qualifying findings. The [coordinated vulnerability disclosure policy](#) provides the current program rules, eligible issue categories and submission instructions.

In addition, OpenAI runs a [Cybersecurity Grant Program](#) to support research and development focused on protecting AI systems and infrastructure. This program encourages and funds initiatives that help identify and address vulnerabilities, ensuring the safe deployment of AI technologies.

Research support has included API-credit microgrants and work on software patching, model privacy, detection and response, security integration and agentic security. External red teaming processes also bring external experts into our risk assessment and mitigation efforts.

5. Does your organization use or contribute to the development of international technical standards, best practices or tools for the identification, assessment, and evaluation of risks?

Contributes and uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- NIST AI Risk Management Framework (AI RMF)
- Industry-led best practices or codes of conduct
- General Purpose AI Code of Practice

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)	☐		
Independent civil society organizations (e.g., for evaluations, audits, or policy review)		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector;

establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

OpenAI collaborates on an ongoing basis with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks:

Protecting children: We [work with child-safety organizations](#) on the detection, review and reporting of child sexual abuse material, and [support shared safety infrastructure](#) through organizations such as ROOST. Since our previous submission, we have also introduced [ChatGPT for Teens](#), with stronger built-in protections, support for healthy use and additional parental controls. Users who state that they are aged 13–17, or whom our systems estimate to be under 18, are placed in the teen experience. This builds on our [Under-18 Model Spec](#), [Teen Safety Blueprint](#) and [age-prediction work](#).

Election integrity: Our election safeguards combine access to reliable information, support for cyber defenders, content provenance, monitoring for political bias and disruption of misuse. We have [worked](#) with groups such as the Associated Press on election results and Democracy Works on voting information. These collaborations sit alongside [our policies](#) against political campaigning and targeted political persuasion, and our [investigations into covert influence operations](#). We report the scope and observed impact of those operations rather than assuming that use of AI made them effective.

Investment in impact assessment and policy analysis: We publish research on AI use, economic impacts and risks to individuals and society, and support independent work that can inform public policy. Recent examples include aggregate usage research, industrial policy grants and funding for research on AI and teen development. Questions 20, 28 and 29 provide further detail on these programmes and distinguish published findings from research that has been funded but not yet completed.

Partnering with governments and researchers: We continue to work with governments' AI safety and security institutes, independent evaluators and domain experts. Our public third-party evaluation guidance describes how the tools and environment surrounding a model can change the result, and why reports should distinguish capability testing from safeguard testing. We also collaborate with public-health and research institutions through, for example, Rosalind Biodefense and trusted access to GPT-Rosalind, as described in question 31.

Supporting cyber defenders: We collaborate with cyber defenders through OpenAI Daybreak, which provides two access tiers for approved defensive work. Daybreak Blue supports general defensive workflows, while Daybreak Red provides cybersecurity-specific models, including GPT-5.6-Cyber, for authorized advanced testing. Access is governed by identity verification, account security, monitoring, approved-use restrictions and legal attestations. This extends our work with the defensive security community while recognizing the risks of models operating with reduced safeguards.

Any further comments and implementation documentation.

Question 3 — receiving and sharing reports

OpenAI has mechanisms in place to receive reports of incidents and vulnerabilities from third parties. These mechanisms are part of OpenAI's commitment to safety and security, allowing external researchers, users, and other stakeholders to report issues that they may encounter.

OpenAI maintains security and safety bounty programs through Bugcrowd. The coordinated vulnerability disclosure policy provides the current submission routes, including an encrypted channel for security incidents. It also links to our outbound coordinated-disclosure policy for vulnerabilities found in third-party

software and our CVE assignment policy for standardized identifiers for publicly known software vulnerabilities.

OpenAI also has a [model-misalignment reporting framework](#) for investigating and publicly disclosing qualifying model behavior observed during training, evaluation, testing and deployment. Employees may flag cases for investigation and potential disclosure, including findings whose causes or mitigations remain uncertain. Where third parties may be affected, security, legal and responsible-disclosure obligations take precedence, including private notification as appropriate. This framework complements rather than replaces external reporting channels and legal disclosure requirements.

Question 5 — standards and best practices

OpenAI has contributed to the work of the following standard development organizations:

NIST's AISIC working groups and task forces on provenance, risk management, biorisk

CSA's AI Standards

CoSAI Security and Safety Standards

C2PA provenance standards

MLCommons safety evaluations

ISO risk management

FMF on frontier risk

Coalition for Health AI

OpenAI also uses external standards in its own risk-management practices, including ISO/IEC management-system standards. Our [Frontier Governance Framework](#) explains how relevant safety and security practices align with the EU General-Purpose AI Code of Practice. We co-founded the [Agentic AI Foundation](#) to support community-led standards for agentic systems.

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-

cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Plan and design: We identify how a system is intended to be used and what could go wrong, then use that assessment to decide which tests and safeguards are needed.

Collect and process data: We select, prepare and filter training data, applying relevant privacy protections and controls over where data comes from.

Build, integrate and interpret model: We test the model as it is developed and use the findings to improve its behavior and relevant protections.

Verify and validate: We examine what the system can do and safeguards' effectiveness, using internal automated and human testing and relevant input from outside experts.

Deploy: We review the evidence and the risks that remain after safeguards are applied before deciding whether and how to release the system, which may include phased deployment to understand potential risk among a smaller audience before broadening access.

Operate and monitor: We monitor for misuse and incidents after release and use what we learn to improve safeguards and future assessments.

Evaluations are performed at multiple points throughout the AI lifecycle to ensure models are safe, effective, and align with OpenAI's objectives. In the pre-development phase, we begin with an initial risk assessment to identify potential hazards and define safety objectives, establishing a foundation for informed and responsible development. During the development phase, models undergo iterative testing to refine safety and performance as the model evolves. Before deployment, in the pre-deployment phase, we conduct a range of safety and capability evaluations, as described in our [Preparedness Framework](#). In the deployment phase, we implement system safeguards to prevent harm and monitoring to detect unexpected behaviors or misuse, and we perform post-deployment evaluations using real-world feedback and performance data to identify emerging risks. Finally, in the post-deployment and maintenance phase, we update our evaluations and safeguards based on new misuse patterns, identified incidents, and advances in research.

We embed security controls into infrastructure and model development, and use adversarial testing, monitoring and incident review to improve them. Our [work on long-horizon models](#) has also shown why individual action checks alone can be insufficient: a sequence of seemingly acceptable actions can cross a user's intended boundary.

Significant releases, material evaluation findings, incidents, emerging abuse patterns and changes in applicable requirements can trigger reviews and updates. Findings may change evaluations, model behavior, safeguards, access conditions or monitoring. The Safety Advisory Group advises OpenAI leadership on preparedness decisions, and relevant research, safety, security, product, legal and policy teams contribute according to their responsibilities. An explanation of the safety work for a model is published in a system card when the model is released to general availability.

Findings from internal use, evaluations and incident investigations inform our alignment work, security controls and decisions about whether work can proceed. Our [September 25, 2026 report](#) provides a further example of how incident findings change safeguards and research decisions. An internal research model accessed an external chatbot through insufficient DNS filtering in its training sandbox. Monitoring detected the behavior, but the investigation also identified gaps in detection and operational response. We stopped the affected training run and added blocking controls at two independent layers.

At time of publication, training, evaluation and tool-using inference for our most capable models were paused pending validation that the gap was resolved and additional red teaming. We also committed to beginning a fresh training run with additional alignment improvements rather than resuming training of the affected model.

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Prior to deployment, OpenAI assesses and mitigates potential risks that may stem from generative models, such as information harms, or other content that violates our safety policies. We use a combination of methods, spanning all stages of development across pre-training, post-training, product development, and policy. For example, during post-training, we align the model to human preferences; we red team the resulting models and add product-level mitigations such as monitoring and enforcement; and we provide moderation tools and transparency reports to our users.

We find that the majority of effective testing and mitigations are done after the pre-training stage because filtering pre-trained data alone cannot address nuanced and context-specific harms. At the same time, certain pre-training filtering mitigations can provide an additional layer of defense that, along with other safety mitigations, help exclude unwanted and harmful information from our datasets:

We use our Moderation API and safety classifiers to filter out data that could contribute to harmful content or information hazards, including CSAM, hateful content, and violence.

For image generation systems, we filter our image generation datasets for explicit content such as graphic sexual material and CSAM.

We use advanced data filtering processes to reduce personal information from training data.

Models may perform differently across languages, tasks and groups. Evaluations, adversarial testing and feedback from use help us assess those limitations and guide further safeguards.

8. What security measures are used in testing AI systems?

- Isolated and secure test environments
- Network isolation from external access
- Locally hosted testing environments
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Isolation of staging environments (from production / operational environments)
- Controlled environments for red-teaming and adversarial testing
- Controlled security vulnerability testing
- Rate limits and resource caps
- Human-in-the-loop controls for higher-risk test cases
- Access controls, including roles-based permissions
- Disabling, restricting or simulating high-impact capabilities during testing
- Audit logging and monitoring
- Use of synthetic, anonymized, or dummy data, as appropriate

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

No answer provided

9. How does your organization protect intellectual property, including copyright-protected content?

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Publishing “model cards” or contextual information on training data and model limitations
- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Maintain policies or practices to respect copyright holders’ preferences through, where appropriate/applicable, mechanisms like Robots.txt “do not train” tags or other machine-readable protocols developed through voluntary, industry-led standards processes
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Engage in AI training that comprises a non-consumptive use of any underlying data
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Restricting outputs that raise ethical, human rights, or legal concerns—such as non-consensual intimate imagery, unauthorized deepfakes, or outputs that are substantially similar to copyrighted works.
- Filtering results using keyword or image-based restrictions to reduce risks of misuse
- Designating clear points of contact so creators or rights holders can raise concerns about infringing outputs

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)
- Other(s) (specify): privacy filter

Describe key privacy measures and data-protection standards applied.

We want our AI models to learn about the world—not private individuals. We implement a [series of reinforcing privacy protections](#) to train our models from the ground up to protect private and sensitive information.

We apply filters and remove information from our training data that we do not want our models to learn from, including sites that primarily aggregate personal information. We then further train our models to reject requests for private or sensitive information, even where that information is available on the public web. We conduct a series of evaluations on our models to help ensure they align with these standards, implement robust usage policies to protect privacy, and monitor and respond to attempts to generate information in violation of these policies. We are also increasingly relying on our AI models to enhance privacy, such as to improve the filtering of personal information and augment the use of synthetic data, and are excited about the possibilities presented by this research. Our Privacy Filter is an AI tool that we use internally (and later released to the broader community as an open-weight model) to reduce the processing of personal data in the model training process, including for our most capable frontier models.

We additionally give users of ChatGPT robust and easy-to-use personal [data controls](#), including to review and delete their data, turn off features like ChatGPT memory, and to easily opt out of content being used to further improve our models. Users can also export a copy of their ChatGPT data, giving them greater visibility into and control over the information associated with their account. We also encrypt user data at rest and in transit.

We also offer a Temporary Chat mode. Unsaved Temporary Chats do not appear in conversation history and are not used to improve our models; we may keep a copy for up to 30 days for safety purposes. Users can choose to save a Temporary Chat, in which case it becomes a regular chat and follows their account-level settings. And when using ChatGPT data to improve our models, we take a number of privacy-protective steps to reduce the processing of any incidental personal information. Our dedicated [consumer privacy page](#) provides clear, accessible explanations of these controls and directs individuals to our [Privacy Portal](#), where they can submit requests to exercise their privacy rights under applicable law.

Beyond account-level controls, individuals—including those who do not use ChatGPT—can request the correction or removal of personal information about them appearing in ChatGPT responses. We assess these requests in accordance with applicable law and the technical capabilities of our models.

For our [business services](#), including ChatGPT Business, Enterprise, Edu and the API, we do not use customer inputs or outputs to train our models by default. We provide administrative and access controls, alongside data-retention options, with availability varying by service. Customer content is encrypted at rest and in transit. These safeguards are supported by [independent assessments and certified management systems](#), including SOC 2 Type 2 examinations and ISO/IEC 27001 and 27701 certifications for the information-security and privacy-management systems supporting the API, ChatGPT Enterprise and ChatGPT Edu. Our [Trust Portal](#) provides supporting documentation and details of the applicable audit and certification scope.

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)
- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Restrict physical access to servers, GPUs, or devices hosting AI models, systems, and/or applications
- Monitor system activity and log access to detect anomalies or misuse
- Conduct secure AI development and/or deployment practices
- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access
- Other (please specify): combined sensitive-asset controls described in 11.E.

11.D. Insider risk protection detections

- Conduct pre-employment screening or role-sensitive vetting
- Monitor for unauthorized internal access or suspicious activity
- Provide staff training on cybersecurity best practices and AI-specific threats
- Implement insider-risk escalation or response procedures
- Other (please specify): insider-risk monitoring, investigation and response described in 11.E.

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

OpenAI maintains a comprehensive cybersecurity, physical security, and insider threat program governed by an enterprise risk management process and overseen by the Safety and Security Committee (SSC) of the OpenAI Foundation. Our approach integrates industry best practices with AI-specific innovations to manage the unique threats facing frontier AI systems. OpenAI aligns its information security and privacy program with ISO 27001, 27017, 27018 and 27701, supported by SOC 2 Type 2 and NIST guidance.

We have published six key measures for securing advanced AI infrastructure that complement traditional cybersecurity practices:

Trusted computing for AI accelerators

Network and tenant isolation guarantees

Innovation in operational and physical security for datacenters

AI-specific audit and compliance programs

AI for cyber defense

Resilience, redundancy, and research

We continuously assess cybersecurity risks across our systems and infrastructure. Our security controls are tailored to the risks of increasingly capable AI systems, and we adapt as we move closer to AGI. Security safeguards are reviewed regularly and refined through adversarial testing, red teaming, and ongoing research. Security controls and safeguards are also informed by changes in the threat landscape.

OpenAI treats unreleased model weights as core intellectual property. We enforce strict access controls, granting access only to individuals whose roles require it. These protections are supported by a robust insider threat detection program and secure research environments. Model weights and sensitive assets are handled exclusively within appropriately secured infrastructure designed to prevent unauthorized access or leakage. Technical controls in these environments include advanced physical security (for datacenters), cryptographic protections, real-time security monitoring, defense-in-depth network architecture, and supply chain security controls.

Security controls are reviewed continuously and iteratively—through ongoing red teaming, penetration testing, operational monitoring, third-party audits, and program-level reassessments. OpenAI engages in an independent third party audit on at least an annual basis which reviews our security measures, including our vulnerability management process.

OpenAI has a dedicated insider threat team. The program emphasizes prevention, detection, and response. This includes strong policy enforcement, employee education, access controls, monitoring mechanisms to detect anomalous or high-risk activity, and investigative capabilities.

For covered frontier models, our [Frontier Governance Framework](#) describes protections for unreleased weights including encryption, restricted access, monitoring and logging. Authorized access to model parameters is reviewed regularly. Model execution is sandboxed with safeguards tailored to the relevant workload and operating environment. Security assessments include internal and external testing.

12. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

- Internal vulnerability disclosure receipt and tracking processes
- Bug bounty or incentive programs for external reporters

- Investigation of reported or identified risks and vulnerabilities
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Testing of patches or other mitigations
- Scanning for and mitigation of similar risks or vulnerabilities
- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)
- Collaborative vulnerability sharing (e.g., with partners, industry groups)
- Reporting to regulators / authorities when relevant
- Recognition of external finders of confirmed risks or vulnerabilities

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices
- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

Question 12 — vulnerability management

OpenAI receives vulnerability reports through its security and safety bounty programs and provides an encrypted channel for reporting security incidents. Its disclosure policy also describes coordination for vulnerabilities found in third-party software and its role as a CVE Numbering Authority.

For covered frontier systems, security assessments include red teaming, penetration testing and vulnerability scanning.

Question 13 — post-deployment practices

OpenAI disallows the use of our models and tools for certain activities and content, as outlined in our terms and [usage policies](#). These policies are designed to prohibit the use of our models and tools in ways that

cause individual or societal harm. We periodically update these policies in response to new risks and new information on how our models are being used. Access to and use of our models are also subject to OpenAI's Terms of Use and our Business Terms. We employ a combination of guardrails, tools, policies, and processes to detect and enforce against abuse of our models. In addition to usage policies, we employ automated systems (for example, our Moderation API) and AI-powered tools to detect and prevent harmful content, human reviewers and expert investigators to conduct detailed investigations of model behavior and platform abuse, and external input to help our teams iterate on and regularly improve our moderation, detection, disruption, and enforcement capabilities. We publish information related to our post-deployment monitoring, mitigation and enforcement activities.

We build robust monitoring systems to detect potential model misalignment. We view model alignment as the primary lever to prevent potential misaligned behavior from our models. However, monitoring provides broad visibility into frontier model behavior, illuminating opportunities to further improve alignment and safety. This serves as an additional layer of protection against misaligned behavior that is detected.

Addressing vulnerabilities, incidents, emerging risks and misuse often requires participation from a number of teams across the company, including where applicable, teams responsible for research; safety and alignment; product; user or customer relationships and support; intelligence, investigations and teams supporting scaled detection and enforcement; security; policy teams; global affairs; legal and privacy; and communications.

Post-deployment, we rely on monitoring, user feedback, periodic security evaluations, and incident reporting to track risk emergence over time. Material findings can be fed back into model update cycles, safety classifiers, or policy adjustments.

Questions 14 and 14A — agentic systems and controls

Yes. Some of our products can perform supported tasks using tools and a cloud browser, including reading websites and completing multi-step workflows. It pauses when it needs user input, sign-in, or confirmation.

[Cloud browser documentation.](#)

We design controls to help users remain in charge of an agent's work. In our agentic products, users can manage website permissions, review consequential actions, and take over the cloud browser. OpenAI tests for prompt injection, phishing, and unintended actions, among other things.

We also train and test the agent to resist prompt injection: attempts to make it follow malicious instructions hidden in material it encounters, such as a webpage. Monitoring helps us detect and respond to these attacks. Secure browser takeover allows users to enter credentials without exposing those inputs to the model. Together, these measures address risks that arise when an agent uses outside information or works through services where the user is signed in.

Additionally, we train the model to be aligned with human intentions, and monitor for misalignment issues. This monitoring helps us detect and respond to issues as they arise in internal and external deployment, allowing us to iterate on further safeguard and evaluation development.

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

Yes

15.A. If yes, which of the following information is included in your organization's publicly available documentation for any of your AI products?

- Risk management policies
- Summary of results from evaluations of safety, security, and societal risks including potential adverse impacts on individuals or groups (e.g. impacts on human rights, privacy or differential outcomes across user groups).
- Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments).
- Capabilities, behaviours, and limitations, including domains of appropriate and inappropriate use or indicative thresholds associated with systemic risks.
- Technical documentation and instructions for use if relevant.
- Plain-language or non-technical summaries

Please provide links or references showing where this information is made public.

We publish system cards for frontier models and other significant launches, covering capabilities, limitations, safety evaluations and intended uses. Since our previous submission, this reporting has expanded to newer reasoning models, open-weight models and agentic systems. The GPT-6 Astra system card and accompanying [safety overview](#) provide a recent example, including the model's Critical cybersecurity capability assessment, strengthened safeguards and remaining monitoring limitations.

Examples of our public documentation include the [Deployment Safety Hub](#), the [GPT-6 Astra system card](#) and the [Preparedness Framework](#).

15.B. If yes, how often are such reports usually updated?

At or before major product releases

16. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

- Public documentation
- Protected information sharing
- Regulatory reporting as relevant
- Direct engagement with AI safety institutes, civil society or academic partners

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

OpenAI actively shares information with other stakeholders regarding the outcome of risk evaluations for advanced AI systems.

In addition, OpenAI both publishes certain insights derived from its detection and disruption of malicious use of its platform, and also shares certain insights with industry peers, industry groups, and relevant

authorities to enhance our collective ability to identify and mitigate risks.

System cards and public summaries make findings available to a broad audience, including researchers, developers, and consumers. Our published [external-testing approach](#) distinguishes independent evaluations, methodology review and subject-matter expert probing. We share results through system cards and support assessors in publishing more detailed work, subject to the applicable confidentiality and accuracy review. Government and independent research partners also receive access and supporting information tailored to the evaluation.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes

If yes, provide links to privacy policies or notices covering personal data, user inputs, and AI-generated outputs.

OpenAI discloses privacy policies that address the use of personal data, user prompts, and the outputs of advanced AI systems. These policies outline how data is collected, stored, and used, emphasizing user privacy and data security.

These resources include our [Privacy Policy](#), [Data Processing Addendum](#), [Business Terms](#), [information on how our models are developed](#) and [how data is used to improve model performance](#), the [Consumer Privacy Page](#), [Privacy Portal](#), the [Data Controls FAQ](#), and the [Trust Portal](#).

18. How does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

- Public documentation

Please provide links to any publicly available information.

We provide general information about our training data in our system cards, including information about the sources of our data. That information is disclosed at a high level to balance the value of public disclosures with trade secret sensitivities.

We also publish model-specific summaries of training content on our [EU AI Act resources page](#). In addition, we publish a [training-data summary pursuant to California Civil Code Section 3111](#).

19. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

OpenAI demonstrates transparency regarding its advanced AI systems through a variety of technical documentation, policies, and instructions that are publicly accessible. These resources include:

Model Specifications: OpenAI publishes detailed documentation, such as the [Model Spec](#), which outlines the desired behavior for its models. These documents serve as a guide for researchers and developers, providing transparency on how OpenAI shapes the behavior of its AI systems. It includes a set of core objectives, as well as guidance on how to deal with conflicting objectives or instructions.

[Usage Policies](#): OpenAI has established comprehensive usage policies to ensure that its technology is used responsibly. These policies cover a range of topics, including compliance with applicable laws, safeguarding privacy, and preventing misuse of AI for harmful activities. The policies are designed to provide flexibility for innovation while maintaining safety and ethical standards.

For agentic products, transparency is also part of the user experience. For supported browser workflows in agentic products, users can follow task progress in the conversation, review confirmation requests, and take over the browser. Documentation explains website permissions, sign-in, supported actions, and limitations. [Cloud browser documentation](#).

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

Yes

If yes, please provide links or describe the scope of insights published, including where relevant contributions to understanding AI adoption in specific sectors (e.g. public administration, education systems).

Yes. We publish research to help people understand how AI is being adopted and used, including in work and economic activity. "[How People Use ChatGPT](#)," for example, uses automated classification with privacy protections to study patterns in consumer use. It reports aggregate findings, including work-related uses, without publishing individual conversations.

We also publish the [Signals Global Report](#) to offer a broader view of adoption and use across populations. Making this research public gives researchers, policymakers and other stakeholders evidence they can use to examine how AI is affecting different tasks and communities. Each publication explains its scope and methods, so that readers can understand whose activity was studied and what conclusions the findings can support.

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Board of Directors oversight

- Executive oversight
- Senior management oversight
- Governance council or committee

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

OpenAI has implemented several AI risk management and governance policies, as described in System Cards, our [Preparedness Framework](#) and our [Frontier Governance Framework](#). For example, before deploying frontier models and other significant launches, preparedness evaluations and safeguards are reviewed by the Safety Advisory Group, an internal cross-functional group of safety leaders, with recommendations to OpenAI leadership, which makes deployment decisions. Additionally, the Safety and Security Committee of the OpenAI Foundation has oversight of safety and security practices across OpenAI.

The Frontier Governance Framework also sets out responsibilities for the activities it covers. It identifies OpenAI OpCo LLC's responsibilities for covered US models and OpenAI Ireland Limited's responsibilities for EU general-purpose AI models with systemic risk, including oversight by its board for EU purposes.

OpenAI regularly iterates on other policies and procedures to assess, measure, and mitigate risks throughout the development and deployment lifecycle.

Reviews and updates can be triggered by major incidents, regulatory changes, material system updates and new evaluation findings.

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

Yes. We provide role-specific training to ensure relevant employees understand their responsibilities under our governance and risk frameworks. This includes:

Onboarding training for new hires, covering security, privacy, and acceptable use

Regular security training for employees and contractors to address insider risks, as described in the [Frontier Governance Framework](#).

Guidance documents and FAQs that translate policy into actionable steps, especially for product development and AI deployment contexts

Governance-related training content is regularly updated to reflect policy changes and evolving external expectations.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Yes. We log and track significant reported incidents through our internal incident response processes, which include:

Triage and investigation logs

Post-incident reviews (including root cause analysis and corrective actions)

Lessons learned documentation shared with relevant internal teams

The [Frontier Governance Framework](#) describes investigations and reviews where needed to capture lessons and outstanding actions. Escalation to senior leadership and external reporting depend on the applicable incident criteria, severity and affected parties.

Separately, our [model-misalignment reporting framework](#) sets out a process to track, investigate and disclose qualifying model-behavior concerns across training, evaluation, testing and deployment. Any employee may flag an example for review by safety and alignment teams. Technical staff investigate what happened, assess uncertainties and third-party impact, and assign the case to a Ready for Disclosure, Minor Investigation or Larger Investigation track. The reporting employee is informed of the disclosure decision; unresolved disagreements about disclosure or track go to the Safety Advisory Group and, if necessary, OpenAI leadership. The process sets deadlines for investigation and disclosure, and reports document observed behavior, impacts, outstanding questions and mitigations where available; publication may precede completion of an investigation or fix. Where third parties are affected, security, legal and responsible-disclosure obligations take precedence. This framework complements existing incident-response processes and legal reporting obligations.

24. Does your organization share research and best practices on addressing or managing risk?

Yes

If yes, describe how research or good practices are shared (e.g., publications, conferences, collaborations with other sectors).

Protecting Children and Vulnerable Groups: We publish the [Teen Safety Blueprint](#) and [Under-18 Model Spec](#) to explain our approach to younger users. [ChatGPT for Teens](#) puts this work into practice through an age-appropriate experience, stronger protections and parental controls. Our separate commitment to [independent research on teen development](#) is intended to help assess benefits, risks and the effectiveness of safeguards; it is not a claim that those research outcomes have already been established.

Upholding Democratic Values and Respecting Human Rights: OpenAI conducts research on the societal impacts of AI, focusing on ensuring that AI aligns with democratic values and human rights. This includes studying how AI can be governed to reflect democratic input and avoid exacerbating inequalities.

Content Provenance and Abuse Prevention: OpenAI is actively involved in the development of content provenance standards, such as C2PA, to help track the origin of digital content and prevent the spread of misinformation. This effort is part of broader initiatives to enhance the integrity and transparency of AI-generated content. See the response to question 27 on content authentication and provenance. In addition, OpenAI is committed to enforcing policies that prevent abuse and to improve transparency

around AI-generated content, as outlined in our reports on disrupting deceptive uses of AI by covert influence operations.

We also share risk-management practices through public governance frameworks, system cards, research publications, standards initiatives and collaborations with researchers, industry and government safety institutes.

25. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- NIST AI Risk Management Framework (AI RMF)
- OECD AI Principles
- Industry-led best practices or codes of conduct
- Other international or regional standards (please specify): EU General-Purpose AI Code of Practice.

If yes, describe which international technical standards or best practices your organization uses.

Yes. Our governance and risk management programs align with international frameworks and standards including:

ISO/IEC 27001 and 27701 (Information security and privacy management)

SOC 2 (Trust Services Criteria, including Security, Availability, Confidentiality and Privacy)

NIST AI Risk Management Framework

OECD and EU AI guidelines (including definitions of risk levels and governance expectations)

ISO/IEC 42001 (AI Management Systems)

These standards inform both the design of controls and the structure of our governance documentation and training programs.

OpenAI maintains ISO/IEC 27001:2022 and ISO/IEC 27701:2019 certifications for the information-security and privacy-management systems supporting the OpenAI API, ChatGPT Enterprise and ChatGPT Edu. Its [ISO/IEC 42001:2023 AI Management System](#) covers consumer and business AI products and models in its role as an AI producer and provider. The NIST AI Risk Management Framework informs OpenAI's approach to frontier AI safety risks. Our Frontier Governance Framework explains how relevant safety and security practices align with the EU General-Purpose AI Code of Practice.

Any further comments and implementation documentation.

No answer provided

26. What mechanisms, if any, does your organization put in place to allow end-users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

- User disclosure (e.g., text notice, verbal statement, or system label)
- Persistent visual or audio indicators (e.g., watermark, icon, chime, banner) during AI interaction
- Documentation or FAQs explaining when and how AI is used in the system
- Access to provenance information (e.g., model version, content authenticity signals)
- Disclosure to end-users when an AI system is acting autonomously or executing actions on their behalf (e.g., making purchases, sending communications, modifying files)

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

We help users understand when an AI system is acting on their behalf by making its work visible. ChatGPT Work lets users follow browser-task progress in the conversation and review requests for sign-in or confirmation. Users can take over the cloud browser. Public documentation explains when the system can act and where user involvement is required. [Cloud browser documentation](#). Documentation provides further context about the agent's capabilities, tools and limitations. These features are particularly important when a task involves several steps: they help users follow the work, understand what information or services the agent can access, and intervene when its actions need to change.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Yes

27.A. If yes, how?

- Latent disclosure (e.g., hidden provenance information)
- Manifest disclosure (e.g., visible label)
- Detection tool(s) for hidden provenance signals (e.g., watermark or metadata checks)
- Use of international or industry standards (e.g., C2PA)

27.B. If yes, does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes

27.C. Does your organization collaborate on or invest in research to advance content authentication and provenance?

Yes

Describe how provenance mechanisms are implemented and which standards or tools are used, and any collaborations or investments to improve content authentication (e.g., C2PA credentials, joint R&D

projects, open-source initiatives, standard-setting groups). You may also indicate whether these mechanisms are tested for robustness against removal, tampering, or circumvention.

OpenAI is working to ensure the authenticity and transparency of digital content. Here are our key efforts:

C2PA: OpenAI is a steering committee member of the Coalition for Content Provenance and Authenticity (C2PA). OpenAI adds content credentials to supported generated images and videos, building on our implementations in DALL·E 3 and GPT-4o Image Generation. Content credentials (metadata cryptographically signed to provide an attestation of the content's origins) provide information about how and when a piece of content was created or modified by our platform. Content credentials are metadata-based approaches to content provenance that require adoption by different players across the AI ecosystem to build trust in media online. We believe that C2PA is working to build such an ecosystem, and from our steering committee seat, we are helping to advance this work.

Watermarking and verification: We combine Content Credentials with SynthID watermarking for supported images and audio, alongside visible watermarks for supported video. Our [public verification tool](#) can inspect supported image and audio files for provenance signals, and API access allows organizations to incorporate verification into their workflows.

Our July 2026 update extended SynthID watermarking and public verification support to supported audio and introduced API access for verification. We test and research the accuracy, reliability and resilience of these mechanisms. No detection method is foolproof: a missing metadata or watermark signal does not establish that a file was created without AI.

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Security
- Safety
- Information integrity
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Transparency
- Robustness or reliability

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

We advance the state of AI safety and trustworthiness through a combination of internal research, external funding and partnerships, and product integration of emerging insights. Our efforts span across key domains such as robustness, disinformation, fairness, and transparency. We publish our research to uplift safety in the larger AI ecosystem in our [main blog](#) and [dedicated safety and alignment blog](#).

Internal Research and Applied Safety Work

We invest heavily in internal research focused on the safe deployment of frontier models, including:

Robustness and Red Teaming: Dedicated teams probe models for failure modes, including adversarial behavior, prompt injection, and content safety risks.

Transparency and Interpretability: Research includes techniques like model probing, behavioral evaluation, and feature attribution to improve transparency.

Political Bias: We develop and publish methods for evaluating political bias in model responses, with findings informing further model development.

Misuse and abuse prevention: We continuously study how models might be misused to generate harmful content, or violate our [usage policies](#), and implement corresponding safeguards or disrupt such activity.

Research findings are shared via system cards, peer-reviewed papers, and updates to our model usage policies.

External Research Funding and Collaboration

We support external work through:

Grants and fellowships to researchers and institutions advancing AI safety, fairness, and interpretability.

Partnerships with academic labs and nonprofits to co-develop methods, benchmarks, and evaluations.

Challenge programs (e.g., for red teaming or adversarial testing) that incentivize broader participation in AI safety research.

These efforts promote shared learning and raise the industry bar on responsible AI development.

Across all of this, our goal is to continuously improve the trustworthiness and safety of our AI systems, and to share lessons with the broader ecosystem when possible.

Recent work includes [guidance on the validity of third-party frontier-model evaluations](#), [long-horizon alignment and monitoring](#), and [content-provenance verification](#). These publications explain practical methods and limitations so that other researchers and developers can assess and build on the findings. Our public incident investigations also feed lessons back into evaluation and safeguard design.

In September 2026, we committed \$5 million to [independent research on how generative AI affects the lives and development of teenagers](#) aged 13–17. The programme will support work on potential benefits and risks, including social and emotional development, and on the effectiveness of safeguards and interventions. This funding is intended to build evidence that can inform safer experiences for young people.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Socio-economic considerations (e.g., AI adoption and productivity, workforce transitions, skills development, access to AI tools)
- Environmental considerations (e.g., energy use, computational efficiency, resource use)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

We have engaged in a wide range of academic discussions and joint research projects, including looking at the potential economic and social implications of advanced AI systems. Our view is that robust engagement with a wide range of subject-matter experts, including independent academics, think tanks, and research organizations, will further our understanding of how AI can be used to benefit humanity, in line with our company's mission.

Our economic research includes [How People Use ChatGPT](#) and the [Signals Global Report](#). These publications describe patterns of adoption and use across tasks and populations, using aggregate findings rather than publishing individual conversations. Each study explains its scope and methods, helping readers distinguish measured usage from broader claims about productivity, employment or social outcomes.

We also support independent research on how public policy can help people benefit from AI and respond to economic disruption. In August 2026, we announced [industrial policy grants](#) for 14 projects, with a total of \$1 million in funding and up to \$1 million in API credits across the projects.

We also invest in ways to use computing resources more efficiently. Our [infrastructure work](#) looks at models, the software that serves them and the hardware used to run them together, because improvements at each level can affect the energy needed to provide a service.

Efficiency is one part of the environmental picture. Even if an individual task uses fewer resources, an organization's total footprint also depends on how much its systems are used and other factors. Our infrastructure disclosures and research describe findings within their relevant scope so that improvements in efficiency can be understood in that broader context.

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

30. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities and limitations of advanced AI systems?

Yes

If yes, describe any digital literacy or training initiatives your organization supports (e.g., user education, public AI-education programs, school or university partnerships, community workshops). You may also indicate whether these efforts include or prioritise underserved or underrepresented communities.

OpenAI prioritizes education and training to help people understand the nature, capabilities, limitations, and impact of advanced AI systems. We support [OpenAI Academy](#), an AI literacy and education platform including in-person and virtual events and resources, designed to help people of all backgrounds unlock the benefits of AI. Additionally, OpenAI collaborates with Common Sense Media to

provide resources that promote responsible AI use among families, ensuring that users of all ages understand the ethical and social dimensions of AI.

OpenAI Academy: academy.openai.com

OpenAI Academy provides a free public resource hub, workshops and content ranging from introductory AI literacy to more advanced material for technical users. It serves educators, students, job seekers, nonprofit leaders and small-business owners. Published collaborations include [work with Goodwill](#) on AI literacy for employment specialists supporting job seekers, helping connect practical skills with the needs of people facing barriers to opportunity.

Our [Education for Countries update](#), published in May 2026, provides a more recent example of work with an education system. At the time of that update, Estonia's programme reached more than 20,000 students and 4,600 teachers. Alongside access to ChatGPT Edu, AI Leap, the University of Tartu and Stanford University are studying its use in real classrooms, with a commitment to publish findings. This combines practical access and teacher support with research into how the technology affects learning.

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

Yes

If yes, please provide examples, including any collaborations with civil society, community groups, or efforts to address broader societal challenges.

[OpenAI Academy's India Nonprofit AI Jam 2026](#) offers a recent example of how we help nonprofit organisations apply AI to their work. The programme combines demonstrations, discussion and hands-on sessions in which participants develop workflows using their own organisational needs and materials. Examples include assistants that make organisational knowledge available in local languages, analysis of programme data, and tools for producing communications. Participants are guided to define the intended output, create a reusable prompt or workflow, and include a quality check. The accompanying public resource hub makes workshop materials available for organisations to use in their own work.

We also support public-health and preparedness work through [Rosalind Biodefense](#). The programme sponsors GPT-Rosalind access and launch support for trusted developers building defensive applications, including screening, early detection and preparedness tools. We have expanded trusted access for selected government and allied partners with approved public-health and biodefense missions.

We recognize that there is still significant work to be done to develop AI ecosystems that foster broadly distributed benefits of these technologies in the long-term. Significant resourcing will be needed to strengthen AI ecosystems, such as deepening low-resource language datasets, researching and adapting to potential workforce impacts, and rapidly expanding AI literacy.

OpenAI Academy's collaboration with GitLab Foundation and The Annie E. Casey Foundation on the [AI for Economic Opportunity Demo Day](#) showcases projects addressing economic challenges. Together with the nonprofit workshops and workforce partnerships described above, these initiatives support organizations in applying AI to concrete community needs.

Any further comments and implementation documentation.

No answer provided

Generated from the OECD AI Transparency Reporting Framework