FLI 2025 Survey - OpenAI responses

Below, blue text are OpenAI responses. Blue highlighting reflects a selected multiple choice option.

Whistleblowing policies

- Q1. Does your company have a WB policy & function covering frontier AI-focused staff?
- Whistleblowing policies page 2/2
- Q2. Who is formally designated with primary responsibility for overseeing the whistleblowing function and ensuring reports are properly addressed?
- Q3. Which statement best describes the investigative independence of your whistleblowing function?
- Q4. Which of the following concerns are explicitly covered by your whistleblowing policy? (Select all that apply)
- Q5. Does your whistleblowing policy explicitly protect individuals who report concerns in 'good faith' or with 'reasonable cause to believe', rather than requiring certainty that violations occurred?
- Q6. Which of the following persons are protected from retaliation under your whistleblowing policy? (Select all that apply)
- Q7. To which of the following individuals or entities can whistleblowers submit reports according to your policy? (Select all that apply)
- Q8. For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)
- Q9. Which of the following best describes the anonymity and confidentiality provisions in your whistleblowing policy? (Select the one that fits best)
- Q10. Does your whistleblowing policy explicitly protect employees disclosing to external parties (e.g., regulators, accredited journalists, civil-society groups) when internal channels are unavailable, conflicted, or fail to resolve a serious concern within stated timelines? (Select one)
- Q11. If “Limited”, under which circumstances is external disclosure protected?
- Q12. Which mechanisms ensure that your whistleblowing function has access to adequate (technical) expertise to investigate reports? (Select all that apply)
- Q13. Investigation timelines and escalation rights: Which best describes your policy's commitments? (Select one)
- Q14. Which specific forms of retaliation are explicitly prohibited in your policy? (Check all that apply)
- Q15. Do any employment-, separation-, or settlement-related agreements used by your company contain non-disparagement or confidentiality clauses that could deter current or former employees from disclosing AI safety or risk-related concerns? (Select one)
- Q16. Which anti-retaliation provisions are explicitly detailed in your whistleblowing policy? (Select all that apply)

External pre-deployment safety testing (6 Questions)

- Q17. Did your organisation commission one or more independent (no financial/governance ties to your company) organisations to test this model for the dangerous capabilities or

propensities you prioritized (in safety framework if available) before public release?
Q18. What was the highest level of technical access granted to any of the listed external evaluators during pre-deployment testing for the specified release? (Select the highest level that applies)
Q19. What was the longest period of time that an external evaluator was given continuous access for pre-deployment testing of your model? (Select one)
Q20. Which of the following publication arrangements applied to external evaluators' findings? If different evaluators had different publication terms, please select all that occurred and briefly explain using the text-box. (select all that apply)
Q21. During pre-deployment testing, what best describes the query-rate or volume restrictions applied to external evaluators? (Select one)
Q22. Does your organization log and retain the model interactions of external evaluators during pre-deployment testing?
Q23. If you specified external pre-deployment safety evaluations in the previous section, were these performed before or after broad internal deployment? (Select one)
Q24. What level of safety testing does your company require for broad internal deployment of frontier AI models? (Select one)
Q25. Does your company require any of the following safeguards for broad internal deployments of frontier AI models? (Select all that apply)
Safety practices, frameworks, and teams (9 Questions)
Q26. When you released your latest flagship model, did you release the same model version that the final round of safety (framework) evaluations were conducted on? (Select one)
Q27. If your company has one or more teams focused primarily on technical AI safety research, please provide more information about the team(s) below.
Q28. Does your organization have a formal, written policy that requires notifying external authorities when safety testing determines a model exceeds your organization's "unacceptable-risk" threshold (i.e., a risk-level that bars deployment under your own safety framework), even if the model will not be released? (Select option that best describes your policy)
Q29. For companies that signed the "Frontier AI Safety Commitments" at the AI Seoul Summit in 2024, and those that strive to implement equivalent safety frameworks:
Q30. Do you have a plan for ensuring that the AGI you're trying to build will remain controllable, safe and beneficial?
Q31. Which of the following elements of an AI emergency response capability has your organization implemented? (Select all that apply)
Q32. Does your company agree with the following principles for promoting legible and faithful reasoning in advanced AI systems to ensure AI remains safe and controllable? (Select all statements you support)
Q33. Task-Specific Fine-Tuning (TSFT) involves training a model to excel at potentially dangerous tasks (e.g., designing biological agents, cyber attacks).
Q44. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number. You may also share additional information about your company's policies.

Whistleblowing policies

If your company has region-specific whistleblowing (WB) policies instead of a single global WB policy, please answer all questions in this survey with regard to the policy that applies to the majority of your frontier AI-focused management, research, and engineering employees. Unless a question specifically asks about other stakeholders, please answer based on protections available to current full-time employees. You may explain variations for different stakeholder groups in the final question.

You can use the text-box at the end of this section to provide clarifications and/or link to relevant publicly available documents.

Definition of terms:

Whistleblowing Function:
The organizational structure, personnel, processes, and resources established to receive, assess, investigate, and respond to whistleblowing reports. This includes the designated individuals or teams responsible for writing and acting according to the whistleblowing policy, managing the whistleblowing process, any technological systems used to facilitate reporting, and the mechanisms for investigating and addressing reported concerns.

Whistleblowing Policy:
The formal, documented set of rules, procedures, and guidelines that govern how an organization handles whistleblowing. This policy outlines what concerns can be reported ("material scope"), who can report them ("covered persons"), how reports should be made and to whom, how they will be handled, and what protections are available to whistleblowers who follow this policy. It serves as the official framework that defines the organization's approach to whistleblowing.

Covered persons:
Individuals who are explicitly protected when making good-faith reports under the whistleblowing policy. The range of covered persons may vary by organization and jurisdiction.

Material scope:
The range of issues, concerns, violations, or misconduct that can legitimately be reported through the whistleblowing channels and will be considered for investigation. In this context, this may include legal violations, ethical breaches, safety concerns, alignment issues, misrepresentations of capabilities, or other matters related to responsible AI development and deployment that the organization has defined as reportable concerns.

Q1. Does your company have a WB policy & function covering frontier AI-focused staff?

Is this policy publicly accessible without login credentials?

Prefer not to answer (skips whistleblowing section)
No WB policy & function - (skips whistleblowing section)
Non-public policy exists - Please briefly explain your rationale for keeping it private:

X Public WB policy - Please provide URL here:

https://openai.com/index/openai-raising-concerns-policy/
https://cdn.openai.com/policies/raising-concerns-policy-blog-copy-202410.pdf

Whistleblowing policies page 2/2

Q2. Who is formally designated with primary responsibility for overseeing the whistleblowing function and ensuring reports are properly addressed?

X Board/Audit Committee
Executive management
X Compliance/Legal department
X HR department
Other (Please also specify whom this role reports to):

Q3. Which statement best describes the investigative independence of your whistleblowing function?

The whistleblowing function requires approval from management before initiating investigations based on whistleblower reports.
The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management.
X The whistleblowing function can independently initiate and conduct investigations based on whistleblower reports, including those involving senior management, AND has the authority to engage external expertise without approval.

Q4. Which of the following concerns are explicitly covered by your whistleblowing policy? (Select all that apply)

X Violations of applicable laws and regulations
X Violations of the company's public AI safety framework (e.g., Anthropic's Responsible Scaling Policy)
X Credible safety concerns that may not violate specific policies including loss-of-control scenarios

- **X Pressure to compromise safety standards or suppress safety concerns**
- Misleading communications about AI capabilities to external parties (such as regulators, the public, or evaluators) or discrepancies between public claims and internal practices
- None of the above

Q5. Does your whistleblowing policy explicitly protect individuals who report concerns in 'good faith' or with 'reasonable cause to believe', rather than requiring certainty that violations occurred?

- **x Yes**
- No

Q6. Which of the following persons are protected from retaliation under your whistleblowing policy? (Select all that apply)

- **X Current employees**
- Former employees
- **X Contractors and self-employed workers**
- AI research collaborators and academic partners
- Individuals who assist whistleblowers
- Suppliers and vendors with access to company systems

Q7. To which of the following individuals or entities can whistleblowers submit reports according to your policy? (Select all that apply)

- **X Board member or board committee**
- Dedicated Ethics/Whistleblowing Officer
- Ombudsperson
- **X Chief Compliance or Risk Officer**
- **X General Counsel/Legal Department**
- **X Human Resources department**
- **X External/independent third party**
- **X Direct disclosure to a statutory or supervisory authority**
- Other (please briefly specify):

Q8. For former employees and contractors, indicate any policy limitations compared with current employees. (Select all limitations that apply)

	A Former employees	B Contractors
Limited Reporting Channels	Some channels, such as speaking to your current HR representative, are inherently available only to current employees.	
Limited Reportable Issues		
Limited Retaliation Protection		
No Limitations		

Q9. Which of the following best describes the anonymity and confidentiality provisions in your whistleblowing policy? (Select the one that fits best)

Our policy does not provide for anonymous reporting
Our policy allows anonymous reporting but does not specify technical measures to protect reporter identity
Our policy allows anonymous reporting with specific technical measures in place to protect reporter identity (e.g., anonymous hotline, encrypted system)
X Our policy allows anonymous reporting with technical protections AND includes confidentiality commitments for non-anonymous reports

Q10. Does your whistleblowing policy explicitly protect employees disclosing to external parties (e.g., regulators, accredited journalists, civil-society groups) when internal channels are unavailable, conflicted, or fail to resolve a serious concern within stated timelines? (Select one)

Possible Conditions:

• Imminent risk of serious harm

• Management or board implicated

• Reasonable fear of retaliation

• Internal investigation deadlines missed

• Unconditional reporting to a competent regulatory authority

• After internal reporting has been attempted

No – external disclosure is not explicitly protected or is discouraged (skips follow-up question)
Limited – protected only under specific conditions (choose below)
Full – broadly protected under all listed conditions above (skips follow-up question)

Note: Our policy specifically protects disclosures to any “national, federal, state or local agency charged with the enforcement of any laws or regulations.”

Q11. If “Limited”, under which circumstances is external disclosure protected?

Imminent risk of serious harm
Management or board implicated
Reasonable fear of retaliation
Internal investigation deadlines missed
Unconditional reporting to a competent regulatory authority
After internal reporting has been attempted
Other (specify):

Q12. Which mechanisms ensure that your whistleblowing function has access to adequate (technical) expertise to investigate reports? (Select all that apply)

Dedicated AI experts within the whistleblowing function itself
X Authority to consult internal AI experts under confidentiality safeguards, including procedures that shield case details where necessary
Standing agreements with external independent AI ethics/safety consultants
Budget authority to engage external AI experts without requiring management approval
None of the above
Other (please specify):

Q13. Investigation timelines and escalation rights: Which best describes your policy's commitments? (Select one)

X **None – no specific timelines for acknowledgment, updates, or resolution**
Basic – acknowledge receipt ≤ 7 days only
Standard – acknowledge ≤ 7 days and provide updates ≤ 30 days
Full – acknowledge ≤ 7 days, updates ≤ 30 days, final outcome ≤ 90 days
Full + internal escalation – all Full timeframes plus whistleblowers may escalate to board/leadership if deadlines are missed
Full + comprehensive escalation – all Full timeframes plus whistleblowers may escalate both internally AND to regulators/external parties if deadlines are missed

Q14. Which specific forms of retaliation are explicitly prohibited in your policy? (Check all that apply)

Our policy forbids retaliation. Notwithstanding the way this question is worded, it is well established under relevant law that retaliation can include termination or dismissal, demotion or negative performance reviews, or reduction in compensation or benefits. These are all covered under our policy's prohibition of retaliation. Our policy also expressly addresses harassment.

Termination/Dismissal
Demotion, or negative performance reviews
Reduction in compensation or benefits
Exclusion from meetings or information
Harassment or creating a hostile work environment
Blacklisting within the industry
Legal action against the whistleblower

None of the above

Q15. Do any employment-, separation-, or settlement-related agreements used by your company contain non-disparagement or confidentiality clauses that could deter current or former employees from disclosing AI safety or risk-related concerns? (Select one)

We have confidentiality clauses that could impact some forms of public disclosure, but these have carveouts for internal or regulator disclosures. We do not have non-disparagement clauses in any such agreements, except in specific cases where an employee or former employee has entered a mutual non-disparagement agreement with the company.

No - we do not include such restrictions in our agreements

X Yes, but clauses only limit public disclosure; internal or regulator disclosures are explicitly unrestricted.

Yes, but not enforced – clauses exist, but the company has a written policy never to enforce (or threaten to enforce) them against AI safety or risk-related disclosures (no withholding of pay/equity and no legal action).

Yes, enforced - our standard confidentiality and non-disparagement provisions may restrict raising AI safety or risk-related concerns

Q16. Which anti-retaliation provisions are explicitly detailed in your whistleblowing policy? (Select all that apply)

X Defined disciplinary consequences for individuals who retaliate against whistleblowers (e.g., termination, demotion, or other concrete penalties - not just general statements prohibiting retaliation)

Documented investigation procedure for retaliation claims (including designated investigators, timelines, evidence standards, and appeal rights)

Concrete remedial measures for whistleblowers who experience retaliation (e.g., compensation, reinstatement, transfer options, or other specific remedies - not just general commitments to address retaliation)

None of the above are specifically detailed

If you wish to provide clarifications to particular answers, you can use this textbox to do so.

Please reference specific questions using their associated number. You may also share additional information about your company's policies.

External pre-deployment safety testing (6 Questions)

Please answer the following questions about external pre-deployment safety testing with regards to the release of your currently most capable publicly deployed AI model.

Frontier models:

Anthropic - Claude 4 Opus

DeepSeek - R1

Google Deepmind - Gemini 2.5 Pro

Meta - Llama 4 Maverick

OpenAI - o3

xAI - Grok3

Zhipu AI - GLM4 Plus

You can use the text-box at the bottom of the page to provide clarifications and/or link to relevant publicly available documents.

Q17. Did your organisation commission one or more independent (no financial/governance ties to your company) organisations to test this model for the dangerous capabilities or propensities you prioritized (in safety framework if available) before public release?

> No – no such external pre-deployment testing was commissioned (skip to next section)
> **Yes – external testing was commissioned**. Please list the organization(s) that performed relevant tests on the specified model and briefly indicate the broad risk domain(s) covered e.g., “UK AISI: cyber-offense, bio-risk (opens follow-up questions below):

We’ve worked with the US and UK AI Safety Institutes, and independent third party labs such as METR, Apollo Research, and Pattern Labs to add an additional layer of validation for key risks. Where possible and relevant, we report on their findings in our systems cards, such as in the o3 System Card.

Third party assessors were provided OpenAI o3 early checkpoints, as well as the final launch candidate models to conduct their assessments. As part of our ongoing efforts to consult with external experts, OpenAI granted early access to these versions of o3 to the U.S. AI Safety Institute to conduct evaluations of the models' cyber and biological capabilities, and to the U.K. AI Security Institute to conduct evaluations of cyber, chemical and biological, and autonomy capabilities, and an early version of the safeguards. METR measured the models' general autonomous capabilities, and reward hacking. Pattern Labs evaluated the model's cybersecurity related capabilities (evasion, network attack simulation, and vulnerability exploitation). Apollo Research evaluated in-context scheming and strategic deception. In some instances we paid private consultants for their work, but payment is not conditioned on the content of their findings.

If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number. You may also share additional information about your company's policies.

Q18. What was the highest level of technical access granted to any of the listed external evaluators during pre-deployment testing for the specified release? (Select the highest level that applies)

- **Standard inference API with normal user-facing filters in place**
- **Inference API with safety filters disabled (no inference-time mitigations)**
- **"Helpful-only" or base model API (no harmlessness fine-tuning and no filters)**
- **Fine-tuning interface without safety gatekeeping**
- **Direct read/write access to internal activations or weights**

Q19. What was the longest period of time that an external evaluator was given continuous access for pre-deployment testing of your model? (Select one)

- **>5 weeks**
- **>3 weeks**
- **>2 weeks**
- **>1 week**
- **<1 week**

Q20. Which of the following publication arrangements applied to external evaluators' findings? If different evaluators had different publication terms, please select all that occurred and briefly explain using the text-box. (select all that apply)

- **Evaluators may publish independently without prior company approval after the model is released.**
    - This is true if they run their evaluations independently on the deployed model. Results from the red teaming period are under NDA / require prior approval
- **Evaluators may publish independently after company review/possible redaction.**
    - See above, in cases where the evaluator wishes to publish about the specifics of the pre-deployment red teaming period
- **The company pre-committed to reproduce an independently written report in the model card without redactions.**
- **The company publishes report after review/possible redactions.**
    - OpenAI publishes excerpts from the report mutually agreed upon or written, with OpenAI having the final say for what content goes in System Cards.
- **The company provided its own summary of the evaluator's key findings.**

    - This is true in some cases, but we also share back any summaries that we plan to publish with the evaluator prior to release.
- **Findings remain internal**
- **Other: Please briefly explain:**

Q21. During pre-deployment testing, what best describes the query-rate or volume restrictions applied to external evaluators? (Select one)

- **No limits – evaluators could automate or batch queries with no additional throttling or hard caps.**
- **Elevated but capped – evaluators had higher quotas than the public/enterprise tier but were still subject to explicit caps (e.g., requests-per-minute or daily token limits).**
    - Query rates can depend on technical feasibility in some cases.
- **Public-tier caps – evaluators were held to the same rate/volume limits as ordinary paying users.**
- **Lower than Public-tier caps - evaluators had lower quotas than ordinary paying users.**

Q22. Does your organization log and retain the model interactions of external evaluators during pre-deployment testing?

- **Yes - Inputs and outputs are logged and retained.**
- **No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.**
- **Other (please describe):**
- Zero Data Retention available upon request, if technically feasible during pre-deployment periods (for some new models or products, ZDR is not always possible during pre-deployment testing).

If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number. You may also share additional information about your company's policies.

Internal deployments (3 Questions)

Deployment levels:

1) Broad deployment: Many teams within the company have access for normal use.

2) Development access: Access limited to specific teams or projects that are actively testing the model or developing it further.

Q23. If you specified external pre-deployment safety evaluations in the previous section, were these performed before or after broad internal deployment? (Select one)

Before - External safety tests were completed before broad internal deployment.

Partial - All external evaluations on situational awareness, scheming, and cyber-offense were conducted before broad internal deployment.

After - External safety tests were completed after broad internal deployment.

Other (please explain briefly):

Q24. What level of safety testing does your company require for broad internal deployment of frontier AI models? (Select one)

As described in our public Preparedness Framework, we believe that models that have reached or are forecasted to reach Critical capability under our framework will require additional safeguards (safety and security controls) during development, regardless of whether or when

they are externally deployed. We do not currently possess any models that have Critical levels of capability, and we expect to further update this Preparedness Framework before reaching such a level with any model.

No formal risk management requirements for internal deployments
Formalized risk management for internal deployments with less stringent requirements than external deployment framework for the following risks/capabilities: situational awareness, scheming, AI R&D, cyber-offense.
Formalized risk management for internal deployments with the same requirements as external deployment framework for the following risks/capabilities: situational awareness, scheming, cyber-offense.
Company requires the same risk management effort for internal and external deployments.
Other (Please briefly describe):

Q25. Does your company require any of the following safeguards for broad internal deployments of frontier AI models? (Select all that apply)

See answer to Q24, above.

Inference time safety mitigations for misuse risks (including cyber & bio risks)
Restricting access to helpful-only models and only granting time-bound access to staff that apply with a legitimate research need
Logging all inputs and outputs from internal use and retaining them for at least 30 days
Not currently logging, but introduced an *official, written* plan to start doing so after models reach a specified capability threshold
Analyzing all internal model interactions for abnormal activity, including harmful use or unexpected attempts by AI systems to take real-world actions
Live monitoring and automated editing/resampling of suspicious outputs
None of the above
Other (please describe briefly):

If you wish to provide clarifications to particular answers, you can use this textbox to do so.

Please reference specific questions using their associated number. You may also share additional information about your company's policies.

Safety practices, frameworks, and teams (9 Questions)

Q26. When you released your latest flagship model, did you release the same model version that the final round of safety (framework) evaluations were conducted on? (Select one)

Yes – we released the same model version.

No – we further modified the model but explicitly mentioned and described all further changes in the model documentation.

No – further modifications are not described explicitly in the model documentation.

Yes. We ran our evaluations on an earlier checkpoint and then confirmed our automated evaluation results on the final checkpoint.

Q27. If your company has one or more teams focused primarily on technical AI safety research, please provide more information about the team(s) below.

By technical AI safety teams, we are referring to teams researching topics such as scalable oversight, dangerous capability evaluations, mechanistic interpretability, AI control, alignment evaluations, risk-modeling, etc. Please use separate paragraphs for listing multiple teams.

We have multiple teams across safety research focused on safety, alignment, evaluations, trustworthiness and governance.

1) Team name (& website URL if available)

2) Mission and scope – Briefly describe the team's focus. Please distinguish between:

- immediate product safety (e.g., RLHF, jailbreak prevention, safety classifiers), and

- forward-looking/fundamental research (e.g., model organisms of misalignment, mechanistic interpretability)

3) Technical FTEs – Approximate number of full-time equivalent technical staff (researchers and research engineers). Please count each individual only once, based on their primary team.

Q28. Does your organization have a formal, written policy that requires notifying external authorities when safety testing determines a model exceeds your organization's "unacceptable-risk" threshold (i.e., a risk-level that bars deployment under your own safety framework), even if the model will not be released? (Select option that best describes your policy)

X 1) No policy – there is no written requirement to notify any external body.

2) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority.

3) Regulator + public transparency – as in option 2 **and** the policy provides for a public statement or summary once doing so will not exacerbate security risks.

Other (please briefly describe):

Q29. For companies that signed the "Frontier AI Safety Commitments" at the AI Seoul Summit in 2024, and those that strive to implement equivalent safety frameworks:

Which of the levels below best describes the status of your Safety Framework? Please indicate the *highest* option below that accurately describes your current state.

No official Safety Framework published (yet).

Published & Implementation in progress

Published & substantially implemented – Most discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational. Please briefly assert which elements have not been implemented as described yet and the expected timeline for implementation:

Published & fully implemented – All discrete policies, processes, or technical safeguards described in the policy are fully implemented and operational.

Q30. Do you have a plan for ensuring that the AGI you're trying to build will remain controllable, safe and beneficial?

No

No, but we're working on it

Yes, internally. (Please briefly explain why you have not published it)

For more on our approach to ensuring that AGI remains controllable and safe, see https://openai.com/safety/how-we-think-about-safety-alignment/

Q31. Which of the following elements of an AI emergency response capability has your organization implemented? (Select all that apply)

- Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). Successfully tested rapid full model rollback including internal deployments within the last 12 months.
- Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months.
- Conducted at least one full live emergency response drill/simulation in the past 12 months.
- Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander and a 24 × 7 duty roster.
- Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.
- None of the above
- **Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses**

OpenAI has developed and continues to improve incident response programs across key areas of its operations, and is likewise improving and iterating on AI safety incident-specific protocols that are tailored to our operations and technology. Our goal is to respond to incidents in a rapid, coordinated way. Our response capabilities include:

- **Technical Controls for Rapid Mitigation:** We maintain the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized within the last year in response to our finding that a GPT-4o model update was overly flattering or agreeable (see *Sycophancy in GPT-4o: what happened and what we're doing about it*, https://openai.com/index/sycophancy-in-gpt-4o/).

- **Incident Response Planning and Structure:** OpenAI has formal incident response plans for key areas of operations and continues to iterate on AI safety incident-specific protocols. Our response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. We maintain close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans.

As part of our commitment to continuous improvement, we continue to refine our incident response capabilities, including robust playbooks for rapid-response. These efforts are integral to our broader model governance and safety assurance frameworks.

Q32. Does your company agree with the following principles for promoting legible and faithful reasoning in advanced AI systems to ensure AI remains safe and controllable? (Select all statements you support)

Leading AI companies should:

Ensure Human-Legible Reasoning - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. [No.]

Avoid Optimization That Encourages Obfuscation - Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior.

Disclose Optimization Pressures on Reasoning - Companies should transparently report the optimization pressures and training methods applied to model reasoning, particularly when removing 'undesired reasoning'.

None of the above

We've publicly urged against optimizing on chains of thought: https://openai.com/index/chain-of-thought-monitoring/

Q33. Task-Specific Fine-Tuning (TSFT) involves training a model to excel at potentially dangerous tasks (e.g., designing biological agents, cyber attacks).

Before releasing your current frontier model, which statement best describes your TSFT safety testing? (Select one)

None – no TSFT safety testing performed (skips follow-up).

Partial – TSFT performed on ≤ 2 high-risk domains (choose below).

Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below).

None. We evaluated helpful-only models, which we believe is appropriate for the threat model of misuse for models made available via our platform and whose weights we do not release, as is codified in our Preparedness Framework.

Q34. If you selected 'Partial' or 'Comprehensive' on the previous question, Please tick the risk-domains tested with TSFT.

Biological	Persuasion
Chemical	Deceptive alignment / Autonomy
Cyber-offense	Other (please specify):

Q44. If you wish to provide clarifications to particular answers, you can use this textbox to do so. Please reference specific questions using their associated number. You may also share additional information about your company's policies.

Below, we include some additional information about our security work that we believe may be useful context for evaluators considering our overall posture and approach.

- For additional technical detail on our security measures for AI see: Securing Research Infrastructure for Advanced AI.
- Third party collaboration on security: OpenAI maintains a bug bounty program through BugCrowd (https://bugcrowd.com/openai), and welcomes responsible disclosures from third parties via our coordinated vulnerability disclosure policy (https://openai.com/policies/coordinated-vulnerability-disclosure-policy/) . 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.