Question ID	Question Title	Available options	OpenAI	xAI	Z.ai	Anthropic	Google Deepmind
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 released the same model version. Yes. All internal evaluations in the system card were conducted on the final checkpoint.	Yes – we released the same model version.	Yes – we released the same model version.	Yes – we released the same model version.	
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. 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.		We have multiple teams focused primarily on technical AI safety research, led by Johannes Heidecke (Safety Systems) and Mia Glaese (Alignment). Subteams and projects include: ▪ Mechanistic interpretability ▪ CoT interpretability ▪ Automating Alignment ▪ Safety oversight & control ▪ Dangerous capability evaluations ▪ Alignment evaluations ▪ Faithfulness & anti-scheming		1) Zhipu Evaluation Team & Zhipu Safety Team & Zhipu Posttraining Team We do not have team websites. 2) We prefer not to say. 3) 20~30	Aligned with our mission and origin as a safety research lab, we have multiple teams working on AI safety research including alignment science, interpretability, frontier red team, safeguards (research team, safeguards for Claude) and more.	
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)	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. 4) Other (please briefly describe):	1) No policy – there is no written requirement to notify any external body.	1) No policy – there is no written requirement to notify any external body.	1) Regulator-only notification – the policy mandates prompt disclosure to a competent regulatory, or supervisory authority.	Other (please briefly describe): U.S. Government notice when model requires ASL-3+ safeguards; see our RSP for more	
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.	Published & Implementation in progress We published version 2 of our Preparedness Framework on April 15 2025 and have implemented safeguards for high biological and chemical risk, which we first deployed with ChatGPT Agent, launched on July 17, 2025.	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 & 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:	
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)	Our mission is to ensure that artificial general intelligence benefits all of humanity. As part of our recently concluded recapitalization, the OpenAI Foundation became operational and has made an initial $25 billion commitment to invest in two areas: Health and curing disease, and technical solutions to AI resilience. For more on our approach to ensuring that AGI remains controllable and safe, see this post.	No	No, but we're working on it	Yes, publicly shared here (please provide URL): Anthropic, Responsible Scaling Policy, Version 2.2	
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	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, including by improving and iterating on our 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, including 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.		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.	Other: Please use this text-field to share URLs to relevant documentation or to clarify specific responses Please see our RSP and transparency hub for more	
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	**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. We've published research and joined a broader working paper urging against optimizing on chains of thought: As we noted in the GPT-5 system card, "our commitment to keep our reasoning models' CoTs as monitorable as possible (i.e., as faithful and legible as possible) allows us to conduct studies into our reasoning models' behavior by monitoring their CoTs."		**Ensure Human-Legible Reasoning** - AI models should reason in ways that are accessible and understandable to humans. Developers should avoid opaque reasoning methods. **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.		
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 for gpt-5. 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. Note that we did task-specific fine tuning on biological and cyber capabilities for gpt-oss and published a paper with our findings, Estimating worst case frontier risks of open weight LLMs.		Comprehensive – TSFT performed on ≥ 3 high-risk domains (choose below).		
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):			Deceptive alignment / Autonomy		
Q35	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: Security on the path to AGI ▪ Third party collaboration on security: OpenAI maintains a bug bounty program through BugCrowd, and welcomes responsible disclosures from third parties via our 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.				