Question Title	Available options	Zhipu AI	xAI	OpenAI
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):	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 intend to share our model with certain independent organizations for evaluation purposes; however, we prefer not to disclose their identities.	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.
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	Inference API with safety filters disabled (no inference-time mitigations)		▪ 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)
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	>3 weeks		>2 weeks
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. ▪ Evaluators may publish independently after company review/possible redaction. ▪ The company pre-committed to reproduce an independently written report in the model card without redactions. ▪ The company publishes report after review/possible redactions. ▪ The company provided its own summary of the evaluator's key findings. ▪ Findings remain internal ▪ Other: Please briefly explain:	Evaluators may publish independently without prior company approval after the model is released.,Evaluators may publish independently after company review/possible redaction.,The company pre-committed to reproduce an independently written report in the model card without redactions.,The company publishes report after review/possible redactions.		▪ 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 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.
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). ▪ 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.	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.
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):	Other (please describe): We will communicate with the evaluators to confirm whether it is permissible to retain relevant records.		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).