| Question ID | Question Title | Available options | OpenAI | xAI | Z.ai | Anthropic | Google Deepmind |
|---|---|---|---|---|---|---|---|
| **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)<br>▪ 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.<br>We've worked with the US CAISI and the UK AI Security Institute, independent third party labs such as METR, Apollo Research, SecureBio and Irregular 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 GPT-5 System Card.<br>Third party assessors were provided OpenAI GPT-5 Thinking early checkpoints, as well as the final launch candidate models to conduct their assessments across main preparedness categories (Cyber, Bio, AI Self-Improvement). As part of our ongoing efforts to consult with external experts, OpenAI granted early access to these versions of GPT-5 Thinking to both CAISI and UK AISI, both who conducted evaluations of the model's cyber and biological and chemical capabilities, as well as safeguards. As part of a longer-term collaboration, UK AISI was also provided access to prototype versions of our safeguards and information sources that are not publicly available – such as our monitor system design, biological content policy, and chains of thoughts of our monitor models. This allowed them to perform more rigorous stress testing and identify potential vulnerabilities more easily. Grey Swan and FAR.AI conducted general jailbreak red teaming. METR measured the model's general autonomous capabilities, and reward hacking, and Apollo Research evaluated in-context scheming and strategic deception. Pattern Labs evaluated the model's cybersecurity related capabilities, and SecureBio measured the models' biological capabilities. | 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):<br>CN CAICT: General Safety Issues | 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):<br>Please see our system cards (library, Claude Opus 4) and transparency hub for information on our external testing | 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):<br>Yes, external safety testing was commissioned for 2.5, including across CBRN, Autonomy, Cyber, and Extremism and Radicalisation.<br>We have worked with a diverse group of external experts, including Apollo Research, Dreadnode and Vaultis. (See link below)<br>We plan to share more publicly about our approach.<br>TIME - Exclusive: 60 U.K. Lawmakers Accuse Google of Breaking AI Safety Pledge |
| **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<br>▪ Inference API with safety filters disabled (no inference-time mitigations)<br>▪ Helpful-only" or base model API (no harmlessness fine-tuning and no filters)<br>▪ Fine-tuning interface without safety gatekeeping<br>▪ Direct read/write access to internal activations or weights | ▪ Standard inference API with normal user-facing filters in place<br>▪ Inference API with safety filters disabled (no inference-time mitigations)<br>▪ Helpful-only" or base model API (no harmlessness fine-tuning and no filters) | Helpful-only" or base model API (no harmlessness fine-tuning and no filters) | "Helpful-only" or base model API (no harmlessness fine-tuning and no filters) | | Inference API with safety filters disabled (no inference-time mitigations)<br>External testing partners were provided the model without inference time mitigations relevant to their specific domain. We plan to set out more detail on our external testing programme in future. |
| **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<br>▪ >3 weeks<br>▪ >2 weeks<br>▪ >1 week<br>▪ <1 week | >2 weeks | >5 weeks | >3 weeks | | >3 weeks<br>External testing partners began testing pre-deployment with interim findings provided before launch and then continued post deployment with further findings provided. |
| **Q20** | *Which of the following publication arrangements applied to external evaluators' findings?*<br>*If different evaluators had different publication terms, please select all that occurred and briefly explain using the text-box.*<br>*(select all that apply)* | ▪ Evaluators may publish independently without prior company approval after the model is released.<br>▪ Evaluators may publish independently after company review/possible redaction.<br>▪ The company pre-committed to reproduce an independently written report in the model card without redactions.<br>▪ The company publishes report after review/possible redactions.<br>▪ The company provided its own summary of the evaluator's key findings.<br>▪ Findings remain internal<br>▪ Other: Please briefly explain: | Evaluators may publish independently without prior company approval after the model is released.<br>▪ This is true if they run their evaluations independently on the deployed model. Results from the pre-deployment evaluation period are under NDA / require prior approval to protect confidential information.<br>Evaluators may publish independently after company review/possible redaction.<br>▪ See above, in cases where the evaluator wishes to publish about the specifics of the pre-deployment period - METR as an example did publish and made a note that they believe that our redactions did not substantively change their conclusions ("We did not make changes to conclusions, takeaways or tone (or any other changes we considered problematic) based on their review.")<br>The company publishes report after review/possible redactions.<br>▪ OpenAI publishes excerpts from the report mutually agreed upon or written, with OpenAI having the final say for what content goes in System Cards.<br>The company provided its own summary of the evaluator's key findings.<br>▪ This is true in some cases, but we also share back any summaries that we plan to publish with the evaluator prior to release to confirm factual accuracy. | Evaluators may publish independently after company review/ possible redaction. | Evaluators may publish independently without prior company approval after the model is released. | Evaluators may publish independently after company review/ possible redaction.<br>The company provided its own summary of the evaluator's key findings. | The company provided its own summary of the evaluator's key findings.<br>GDM publishes high level summaries appropriate for the risks being evaluated within the Models Cards / Tech report with GDM having the final say for what content goes in the Model Cards/ Tech report. |
| **Q21** | *During pre-deployment testing, what best describes the query-rate or volume restrictions applied to external evaluators?*<br>*(Select one)* | ▪ No limits – evaluators could automate or batch queries with no additional throttling or hard caps.<br>▪ 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).<br>▪ Public-tier caps – evaluators were held to the same rate/ volume limits as ordinary paying users.<br>▪ Lower than Public-tier caps - evaluators had lower quotas than ordinary paying users. | 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).<br>Query rates can depend on technical feasibility in some cases. | 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). | 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).<br>Query rate is bespoke depending on the testing partner's specific needs and evaluation type. Where required, GDM provided elevated but capped quotas, but this rate often depended on technical feasibility. |
| **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.<br>▪ No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.<br>▪ Other (please describe): | Other (please describe):<br>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). | No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. | No - Inputs and outputs are neither logged nor retained, protecting evaluator IP. | | No - Inputs and outputs are neither logged nor retained, protecting evaluator IP.<br>No - Inputs and outputs are not logged during pre-deployment testing by external evaluators. However, where agreed, external evaluators share prompts and model responses for the purpose of assessment and mitigation of risks. |