Do frontier AI developers act on third-party risk findings?

Analyzing publicly available evidence of company responses to evaluations

Kunal Singh, Max Kamachee, Jonas Raedler,
Stephen Casper

NOVEMBER 2026

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

Abstract

Government AI institutes and independent evaluators study frontier AI systems to identify, evaluate, and better understand potential risks. But do AI companies act on the risks identified by third parties? Here, we examine publicly documented company and policy responses to third-party evaluation findings of frontier AI systems. We identified 1,120 distinct findings about frontier models across 449 AI evaluation reports published by government institutes, academia, and evaluation organizations between March 2023 and August 2026. Of these, we classified 211 as adverse empirical findings concerning identifiable models or developers which may pose a credible safety risk in one or more of the risk domains covered in developers' public safety frameworks. Among 179 findings that met predefined significant-risk criteria, 154 findings (86.0%) lacked a specific and proportionate publicly documented company response, which we define as a specific mitigation, such as a model change, new safeguard, or access restriction, publicly documented by the company with explicit attribution. Policy uptake was documented for 40 of 211 trackable findings, but only one case involved binding action. Collectively, the results suggest that, based on available information, there is limited public evidence that frontier AI developers consistently respond to risks discovered in third-party evaluations attributably and verifiably. Finally, we compare the accountability structures at third-party evaluation organizations with those of established oversight bodies in other industries, including food, drugs, energy, transportation, and finance. We conclude that formal mechanisms for access, response, remediation, verification, and follow-up inspired by governance in these other sectors could increase the consistency with which AI companies proportionately respond to risks identified in their systems.

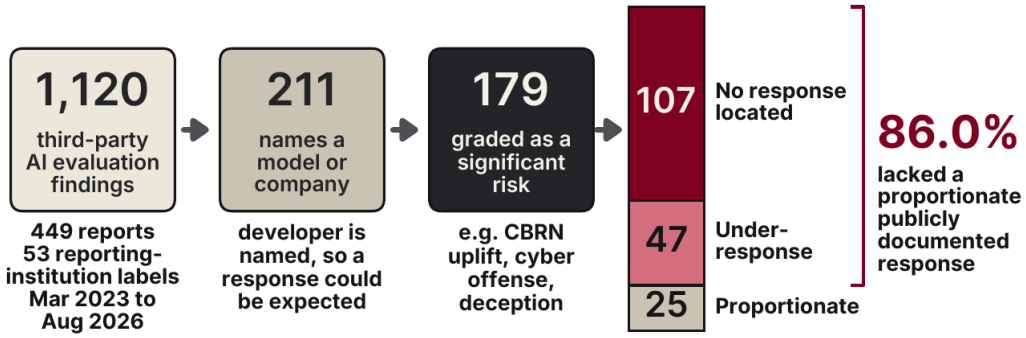

Figure 1: The path from 1,120 findings to the 179 significant-risk findings whose company responses we coded. For 86.0% of these significant-risk findings, we found no public company response that met our predefined proportionality standard.

Table 1: Company-response outcomes for the 179 significant-risk findings. The first two rows make up the 86.0% shortfall. Section 3.5.4 and Figure 5 define the proportionality criteria.

Outcome	n (% of 179)	What it means
Accountability gap (no action)	107 (59.8%)	No publicly documented company response was located.
Under-response	47 (26.3%)	A public response was documented, but it fell below the required Action Level or lacked explicit attribution.
Proportionate	25 (14.0%)	A specific, documented, and attributable response directed towards the risk mitigation.
Gap total (rows 1 and 2)	154 (86.0%)	The central shortfall measured in this paper.

1 Introduction

Frontier AI system evaluations have become a prominent component of AI governance (Shevlane et al., 2023; Araujo et al., 2024; Falco et al., 2021; Feffer et al., 2024; Weidinger et al., 2024). Government AI institutes and independent evaluators test advanced systems for dangerous capabilities, safeguard failures, and other risks, often with access and technical expertise not available to the general public (Brundage et al., 2026; Anderljung et al., 2023; Longpre et al., 2024, 2025; Feffer et al., 2024). These evaluations can provide evidence about an AI system's risk beyond the developers' own disclosures (Raji et al., 2022). Yet without any accountability mechanisms, such evaluations may have limited practical significance in risk mitigation (Birhane et al., 2024); effective auditing also depends on supporting procedures, resources, and institutional mechanisms (Ojewale et al., 2025; Lam et al., 2024). An evaluation can document a serious vulnerability but does not oblige a developer to respond. Where action does follow, it may take the form of deployment delay, additional review, or independent testing. The impact of evaluation therefore depends largely on what happens after its findings are communicated or published (Longpre et al., 2025). This paper asks a narrow question: when a government or an independent evaluator publishes a risk-relevant finding about a named frontier system, does the public record document company action that addresses the identified risk?

In this paper, we analyzed 1,120 findings about frontier AI systems in 449 unique reports. Among 179 findings that met our predefined criteria for "significant risks" (e.g., CBRN uplift, cyber-offense, deception) and which mentioned a specific company or model, 154 (86.0%) had no publicly documented response meeting our predefined proportionality criteria (Section 3.5.4), while only 25 (14.0%) received a proportionate response (Table 4). We do not assume that an effective evaluation ecosystem requires a proportionate public response to every significant-risk finding. A rate higher than the 14% we observed may be desirable, but we leave the question of what rate would be appropriate to the reader. Substantive public responses were more common for findings from pre-deployment evaluations than for those from post-deployment evaluations. Because we rely only on public records, these figures cannot show whether a developer had acted privately.

These results prompt a question: what would encourage developers to respond to evaluation findings more consistently? Finally, we compare the frontier AI evaluation ecosystem with seven oversight bodies in drug regulation, nuclear safety, transport, audit oversight, and financial supervision. These bodies offer blueprints for how formal mechanisms for access, response, remediation, verification, escalation, and follow-up support accountability.

2 Background

2.1 Government AI institutes

Government AI institutes have emerged as technical bodies focused partly on evaluating the safety and security of advanced AI systems. The first institutes were established in the United Kingdom and United States during 2023, later followed by other similar governance institutes in France, Korea, Singapore, China, and others. Their initial roles have included safety testing, research, standards, and international cooperation, but have evolved over time (Araujo et al., 2024). Each institute varies in focus, structure, access arrangements, and connection to standards or regulatory processes (Araujo et al., 2024; UK Government, 2025; Center for AI Standards and Innovation, 2026; Japan AI Safety Institute, 2026; French Government, 2025). However, most have a remit related to conducting or overseeing evaluations of frontier AI models or companies. We use "government AI institutes" as an umbrella term for these public bodies.

2.2 Independent AI Evaluators

We describe Independent evaluators as non-government organizations, including non-profit organizations, for-profit evaluation firms, and academic research groups, that are institutionally separate from the developer whose system they evaluate, and conduct frontier-AI evaluations. They include generalist evaluators (METR, Apollo Research), domain-specific organizations (SecureBio), and others. Independent evaluators differ from government institutes in their institutional status, funding, access arrangements, and public responsibilities. However, they can provide specialized expertise and scrutiny outside both government and frontier-model developers. Recent work treats evaluator independence, system access, methodological rigour, reporting arrangements, and audit standards as important dimensions of third-party auditing (Falco et al., 2021; Costanza-Chock et al., 2022; Brundage et al., 2026; Birhane et al., 2024; Longpre et al., 2024). For a deeper discussion of prior work, see Appendix I.

3 Methodology

3.1 Tooling and human oversight

We used agent-assisted search to check the organizations' research, publication, and blog pages, with website sitemaps and archived copies where needed. The search also included arXiv links listed on those pages or cited in included reports. LLM-based agents then screened the publications and extracted findings. They also ran three separate searches for evidence of a response: from the company, in policy records, and in public coverage. We used custom scripts to clean the data, cross-check the coded fields, and compute every statistic in the paper. Two authors independently reviewed all findings, including eligibility, coding, annotations, and response evidence. Appendix B describes search details and human checks; Appendix J lists the scripts used.

3.2 Initial data collection and scope

Our *evaluator set* comprises 46 organizations drawn from three sources: members of the International Network for Advanced AI Measurement, Evaluation and Science; members of the AI Evaluator Forum and other evaluators that publish frontier model findings; and evaluators named in frontier developers' reports. We collected 6,684 publication records from these organizations' websites, publication indexes, and research pages. An evaluation report or paper was eligible if it named the evaluator and the tested system and described its methods and results in enough detail to support at least one evidence-backed finding. We screened most records by title and abstract, and read the full document only when those were not enough.

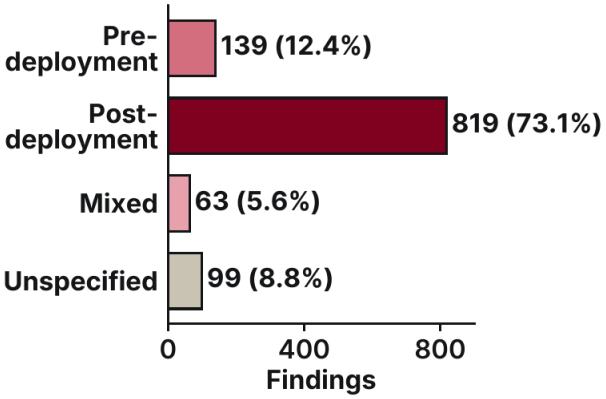

Figure 2: Findings by model access type. Post-deployment testing produced 73% of all findings.

Our unit of analysis is a finding: one discrete, evidence-backed claim that can be coded on its own. A report can contain several findings. Any model from a frontier developer was eligible, regardless of model size. We created a *frontier-developer list* comprising 14 developers in Appendix Table A.2. Findings about anonymized systems were included only when the report described those systems as frontier or an advanced system. We excluded evaluation reports in which a developer tested its own model, unless a named external evaluator took part in the system evaluation. We also excluded models that an evaluator built or modified for its own study. There was no defined starting timeline for the analysis, but 17 March 2023 is the date of the earliest eligible report, and the final search for eligible findings was conducted on 29 August 2026. These reports yielded 1,120 distinct findings, each recorded as a separate dataset entry. Appendix A describes the evaluator set, eligibility rules, and frontier-scope eligibility, Appendix B.1 breaks down the screening of the 6,684 records, and Appendix D describes the resulting corpus.

3.3 Corpus composition and identifying trackable findings

We classify a finding as "pre-deployment" type, when the evaluator tested the model before its public release, as "post-deployment" type, when the evaluator tested the released model, and as "mixed" type, when the evaluation used both. Post-deployment testing dominates the public record (Figure 2). For 99 findings, the access type of evaluation was not disclosed. Figure 3 groups findings by the type of institution that produced them. Nonprofit members of the AI Evaluator Forum produced 328 findings, government institutes 308, other nonprofits 245, and for-profit evaluators 239. Four organizations (UK AISI, Scale AI, METR, and Shanghai AI Laboratory) together produced 511 findings, or 45.6% of the corpus. Appendix Figure D.1 ranks the organizations with the most findings.

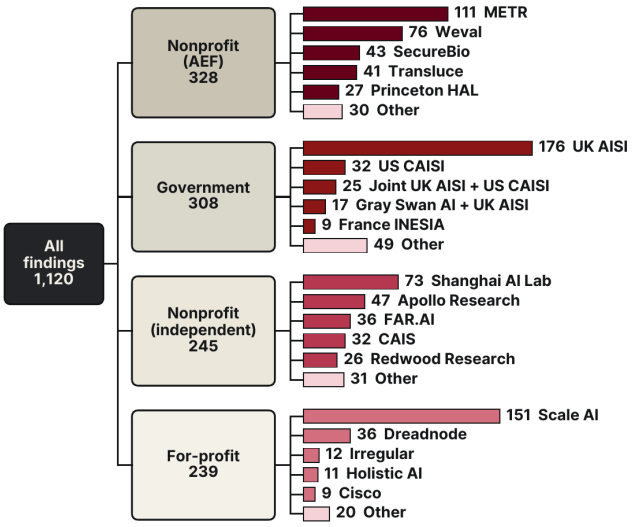

Figure 3: Findings by type of evaluating institution. Each type lists its five largest institutions and all others combined; bars share one scale. Appendix Table D.1 lists every institution.

Not every finding reasonably merits a company response and, therefore we assigned each finding to one of the three mutually exclusive tiers based on whether it reports adverse findings, whether they are trackable, and whether they surface credible risks. **Trackable evaluation findings**: A finding is trackable when it meets three conditions: it reports a specific adverse or safety-relevant result; it names the model, system, or developer involved; and a concrete company response to the reported result can reasonably be assessed. **Non-trackable evaluation findings**: These are empirical findings that fail at least one of the three conditions.

Examples include results about anonymized systems, null results, reassuring or inconclusive results, rankings with no concerning threshold, and industry-wide capability trends. **Non-empirical findings**: These findings report no empirical model behavior or capability. They include contributions to methodology, frameworks, governance processes, and tooling. Of the 1,120 findings, 211 (18.8%) are *trackable evaluation findings*, 603 (53.8%) are *non-trackable evaluation findings*, and 306 (27.3%) are *non-empirical findings*. Only *trackable evaluation findings* then enter our company-response and proportionality analysis (Section 3.5.4).

3.4 Risk severity classification

Findings also differ in how adverse they are, and what counts as an adequate response depends on that. We classified a finding as a **significant-risk** when its reported evidence met at least one significant-risk threshold across eight domains: chemical, biological, radiological, and nuclear (CBRN) uplift; offensive cyber capability; autonomy, self-replication, or automation of AI research and development; persuasion or societal harm at scale; deliberate deception or misalignment; failure of a deployed safeguard; compromised evaluation integrity; and acute harm to an individual user.

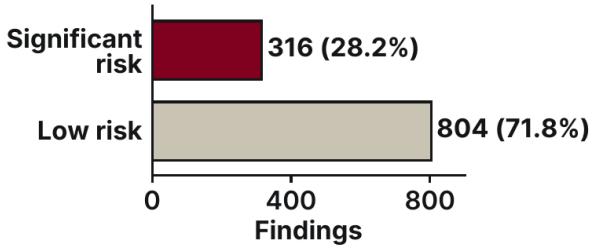

Figure 4: Final severity labels for all 1,120 findings. The company-response analysis uses only the 211 trackable evaluation findings.

All remaining findings that do not contain any significant-risk thresholds are classified as **low-risk** findings. These domains are derived from the safety frameworks by Anthropic, OpenAI, and Google DeepMind (Anthropic, 2026b; OpenAI, 2025; Google DeepMind, 2026); Appendix C.1 gives the source of each domain.

Risk severity of findings were coded in two stages. First, we used three frontier models from different developers (Claude Sonnet 5, GPT-5.5, and Gemini 3.1 Pro) each to classify every finding independently using the same prompt, and their majority vote gave a provisional label. The three models agreed unanimously on 1,010 of the 1,120 findings (90.2%) and on 177 of the 211 trackable evaluation findings (83.9%). Second, two authors

independently reviewed every finding against its source, together with all three votes and rationales, and overrode the model majority where required. The final severity label agreed on by both authors differed from the model majority for 12 of 211 trackable findings. Appendix C provides the group definitions and severity domains, and Appendix C.2 reproduces the complete prompt. Across the corpus, we classified 316 findings as significant-risk and 804 as low-risk (Figure 4). Our company-response analysis covered the 211 trackable findings, of which 179 were significant-risk and 32 were low-risk.

3.5 What responses and impacts did findings have?

Identifying a risk does not show whether any action was taken. We next traced the company, policy, and public responses that followed each of our 211 *trackable evaluation findings*.

3.5.1 Evidence channels: What actions followed a finding being reported?

Table 2: The three response types, each run separately against sources with latest search by 29 August 2026.

Search	Sources searched and evidence rule
Company response	The developer's newsroom, blog, and research index; later model and system cards; safety and deployment pages; and official statements. Only the company's own documents count.
Policy uptake	Official parliamentary, legislative, regulatory, and agency records. A record counts only if it refers to the finding.
Public coverage	Independent media, academic citations, and documented public discussion. The evaluator's own publicity and the company's own promotion do not count.

For each *trackable evaluation finding*, we ran three separate agentic web searches with each prompted to find evidence of a different type of response with defined source evidence rules (Table 2). First, a *company-response* search checked whether the accountable company publicly responded to the finding. Second, a *policy-uptake* search checked whether the finding, or the risk it identified, was referenced in or connected to a policy response from the government. Where a government directive was never published, the search checked a regulated company's public disclosure, provided it named the issuing authority and the mandatory requirement. Third, a *public-coverage* search checked whether the finding drew independent attention in the media, academic work, or public discussion. When measuring public coverage, we excluded posts and announcements by evaluators about their own findings, as well as companies' promotional materials. All three searches covered public sources available by 29 August 2026. Appendix B details the search process.

3.5.2 Attribution: Did a company acknowledge the finding?

To determine when a company publicly linked its response to a finding, we used three labels. A response has *explicit attribution* when the company names the finding, report, result, or evaluator, or states that this evidence informed its response. A response has *no explicit attribution* when we located a qualifying response

but the company stated no link to the finding. The label *no response located* applies when the search found no qualifying public response by the cutoff date.[1]

Cases with no located response are recorded in this third category so that a completed negative search remains distinguishable from an uncoded one.

3.5.3 Action Level: Was a finding acted on?

Action Level records the strength of the publicly documented company response, when applicable. (a) *None*: No public response could be identified through the search by the cutoff date. (b) *Acknowledged*: The company acknowledged the finding or underlying problem but specified no action. (c) *Partial*: The company documented an action that addressed only part of the problem, described the action as interim, limited, or incomplete, or described changes without showing that the problem was addressed. (d) *Substantive*: The company documented a specific mitigation, model change, safeguard, access restriction, or deployment decision that directly addressed the identified problem. Only publicly available documents or statements published directly by the responding company were accepted as evidence.

Table 3: Action Level definitions with examples from the dataset.

Action Level	Finding	Documented company response
Substantive	CAISI found two security vulnerabilities in ChatGPT Agent and combined them into a working exploit.	OpenAI fixed both vulnerabilities within one business day.
Partial	UK AISI found a universal jailbreak in GPT-5.5 within six hours.	OpenAI's system card reports several updates to its safeguards but states that UK AISI could not verify whether the final configuration was effective.
Acknowledged	UK AISI reported unsanctioned behavior by Claude Mythos 5 during a cyber evaluation.	Anthropic cited the incident in its 14 August risk report and reported an ongoing joint investigation. The report identified no specific mitigation for the finding by Anthropic.
None	Dreadnode found that GPT-5.4 recorded two benchmark passes involving cheating under the standard anti-cheating instruction and one under the strictest instruction..	No qualifying company document was located after the documented response search.

Finding IDs and complete evidence are provided in Appendix G and the supplemental materials.

3.5.4 Proportionality: Was a company response proportionate to the risk identified?

We operationalized the "proportionality" of a response using finding severity and Action Level as shown in Figure 5. For significant-risk findings, we required a substantive response with explicit attribution. For low-risk findings, we required a partial or substantive response with explicit attribution. An acknowledgment

1 For every no-response finding, the company response search covered the developer's newsroom, research index, subsequent model and system cards, safety and deployment pages, and official statements. Queries combined the developer, evaluator, report, and specific finding; access related failures were re-attempted directly or through archived pages. This procedure makes missed accessible public evidence less likely. Appendix B.2 describes the search protocol and reports its documented coverage.

without action never meets the standard, and a finding with no located response is always coded as "*no response*". Note that our proportionality annotations did not guarantee that the response was implemented, that it worked, or that the risk was completely mitigated (Section 7). Appendix G shows dataset examples for each *action level*.

	Substantive	Partial	Acknowledged	None
Significant risk	Proportionate	Under-response	Under-response	Accountability gap
Low risk	Proportionate	Proportionate	Under-response	Accountability gap

Figure 5: The proportionality rule by finding severity and company Action Level. Outcomes shown as Proportionate require explicit attribution. Without explicit attribution, those responses are classified as Under-response.

3.5.5 Policy Level – Did a finding have an impact on governance?

Policy Level measures the documented policy uptake for a finding. *No policy uptake identified* is applicable if no qualifying policy response was located for any finding. *Non-binding policy-related uptake* is applicable when finding was cited or discussed but does not create any mandatory obligation, including through official statements, legislative records, consultations, recommendations, or guidance. *Binding policy action* means a government instrument imposed a mandatory requirement that is explicitly connected to the finding or to the same risk in the same model. We did not count general policy activity concerning the same topic unless it referred to the finding.

When a binding government instrument was not publicly released, a company's public disclosure if it identified the issuing authority and the mandatory requirement, is acceptable as a *Binding policy action*. Appendix E provides the complete evidence rules for public and non-public policy instruments.

4 Results

We first report how companies responded to the 179 findings that we annotated as both *significant-risk* and *trackable*. We then compare response rates by severity and by the evaluator's access to the model.

4.1 Most significant-risk findings lack proportionate company response

Of the 179 significant-risk findings, 154 (86.0%) received either no public company response or a response that did not meet the proportionality standard (Figure 1). Policy uptake was rarer still. 39 of these findings appeared in a non-binding policy record, one was linked to a binding government measure, and we found no policy uptake for the other 139. Appendix E describes the finding-level records and their evidence.

In Appendix F.3, we analyze the timing between findings and responses. Using the Kaplan-Meier method (Kaplan and Meier, 1958), we estimate that an expected 7 (95% CI: [1,16]) proportionate responses may have not yet arrived by the study cutoff of 29 August 2026. Adding this estimate to the number of observed proportionate responses would yield an under-response rate of 18% (compared to the 14% in Figure 1)

Table 4: Company-response outcomes among the 179 significant-risk findings. The first two rows make up the 86.0% shortfall in proportionate, publicly documented responses.

Outcome	n (% of 179)	Illustrative example
No publicly documented company response located	107 (59.8%)	The AgentHarm benchmark, built by UK AISI and Gray Swan AI, found that several models complied with malicious agentic requests at high rates. The worst case, Mistral Large 2, recorded an 82.2% harm score against a 1.1% refusal rate.
Documented response that did not meet the proportionality standard	47 (26.3%)	UK AISI developed a universal jailbreak against GPT-5.5 in six hours of expert red-teaming. OpenAI reported nonspecific changes to its safeguard stack, but UK AISI could not verify that the final configuration was effective.
Proportionate documented response	25 (14.0%)	UK AISI and US CAISI found multiple universal jailbreaks bypassing the Constitutional Classifiers protecting Claude Opus 4 and 4.1. Anthropic responded by restructuring its safeguard architecture to address the underlying vulnerability class.

Responses were searched for findings published from 17 March 2023 through 29 August 2026. Full breakdown is in Table 1. Related cases are discussed in Sections 4.4 and Appendix G.

4.2 Public company responses were located more often for significant-risk findings.

Among trackable evaluation findings, 107 of 179 significant-risk findings (59.8%) and 24 of 32 low risk findings (75.0%) had no located company response. The no-response rate was 15.2 percentage points lower for significant-risk findings. Neither the two-proportion test ($z = -1.64$, $p = 0.102$) nor the analysis clustered by report ($p = 0.096$) found a statistically significant difference at the 5% level (Appendix F.2). Figure 6 shows the outcomes within each severity group.

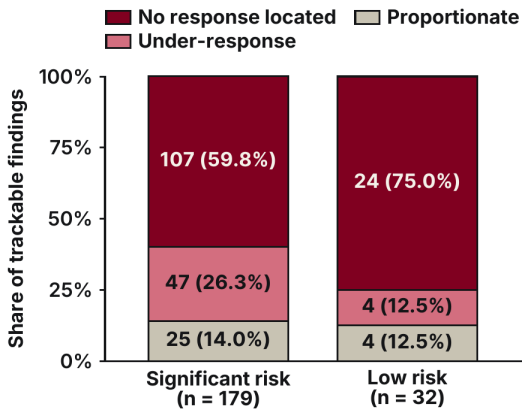

Figure 6: Proportionality outcomes for significant-risk and low-risk trackable evaluation findings.

4.3 Substantive responses are more common for pre-deployment findings

Among significant-risk findings, developers documented a substantive response for 19 of the 39 pre-deployment findings (48.7%) but for only 5 of the 128 post-deployment findings (3.9%). The gap in silence was just as large: One pre-deployment finding (2.6%) and 101 post-deployment findings (78.9%) had no located response. Part of this gap may come from how pre-deployment findings are published.

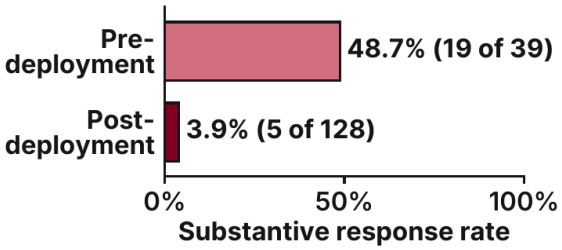

Figure 7: Developers documented a substantive response for 48.7% of significant-risk findings from pre-deployment testing and 3.9% of those from post-deployment testing.

Many appear in the developer's own system card, which reports the finding and the response or acknowledgement together; of the 80 trackable findings with a dated company response, 9 responses appeared before the effective finding date, 42 on the same day, and 29 later. Among the 72 significant-risk responses, the corresponding counts were 9 before, 37 same day, and 26 afterward. For pre-deployment findings, 18 of 19 Substantive responses were public by the effective finding date. This timing is consistent with coordinated disclosure. (Appendix Figure F.2).

4.4 Across evaluators and domains, responses are rarely verified or required

Public response outcomes by developer.

OpenAI and Anthropic had the most significant-risk findings and were the only developers with proportionate public responses in our comparison. **For nine of the 12 developers, we located no public company response to any of their 52 significant-risk findings.**

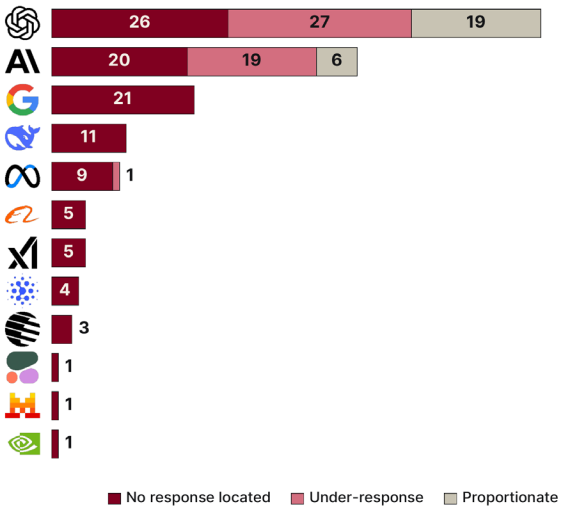

Figure 8: Public company response outcomes by developer for all 179 significant-risk Tier A findings. Bars show the number of findings for each developer on a common scale.

Response gaps persist across reporting configurations, institutions, and risk domains.

The rate at which significant findings lacked a proportionate response ranged from 80.0% for standalone UK AISI evaluations (16 of 20) to 100% for standalone US CAISI evaluations (4 of 4). Among the five largest domains, the shortfall was highest for jailbreak findings (93.0%). The cyber domain and the biological and chemical domain had the lowest shortfall rates (66.7% each). Across all 211 trackable evaluation findings, we located a public response for only one of 13 Societal findings and none for either Human Influence finding or the single Transparency/Disclosure finding. Because these groups are small and domain labels may overlap, we treat this pattern as descriptive rather than as evidence of a difference between risk domains. The shortfall was 83.0% for findings from government institutes and 87.3% for findings from third-party evaluators. The difference was not statistically significant after accounting for clustering within reports (p=0.565). Appendix F.3 gives the complete breakdowns.

One company response had a reported independent retest.

Among the 72 significant-risk findings with a located company response, we identified one reported independent retest. Anthropic reported that US CAISI tested its rebuilt cyber classifier before redeployment (Appendix G.5). In one additional case, UK AISI explicitly stated that it could not verify OpenAI's final mitigation after identifying a GPT-5.5 jailbreak (UK AI Security Institute, 2026; OpenAI, 2026a). For the remaining 70 findings, the public record contained the company's own description of its response but no independent confirmation that the documented action was implemented or effective.[2] Appendix G provides the relevant coding examples, and Section 7 discusses the interpretation of this count.

Policy uptake is documented for one in five findings but is almost entirely non-binding.

Of the 211 trackable evaluation findings, 40 (19.0%) had documented policy uptake. Thirty-nine appeared in at least one official policy record that created no mandatory obligation, such as official statements, guidance, requests for information, recommendations, or research commitments. **We found only one case as a binding policy action involving Anthropic's Fable 5 and Mythos 5**: a US Department of Commerce export-control directive imposing mandatory access restrictions on these models. The directive was not published, so this classification relies on Anthropic's public description (Appendix G.5).

2 We excluded four cases involving pre-deployment testing collaboration, a capability rating, confirmation that an attack worked, or a company evaluating its own model with its own tools.

5 Institutional mechanisms for reducing the accountability gap

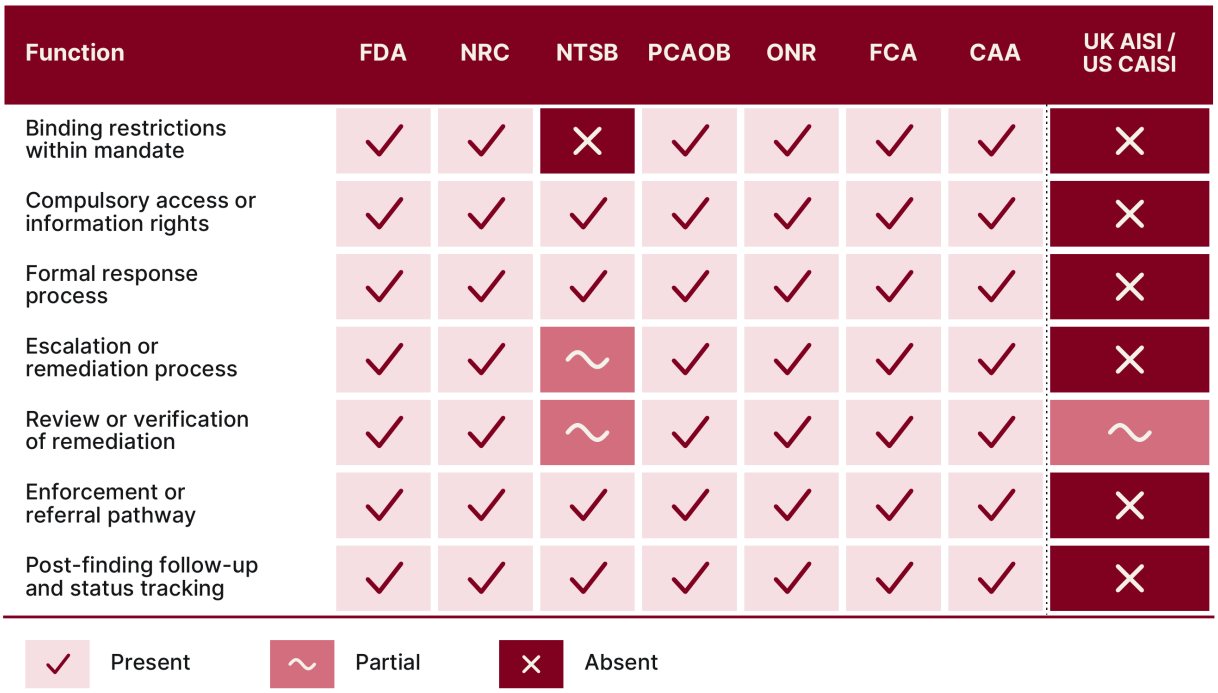

Figure 9: Formal oversight functions across seven comparator institutions, UK AISI, and US CAISI. Appendix H provides definitions, qualifications, and sources.

Public response gaps affected every developer in our significant-risk comparison (Section 4.4). Independent verification was rare, and we found no party that consistently tracked whether findings were resolved. **Our results suggest three missing accountability mechanisms: no developer in our comparison had a rate of proportional responses to significant risk findings of more than 26%, almost no response is independently verified, and no party consistently tracks whether a finding was resolved.** However, established oversight bodies in other sectors build these steps into formal processes. We compare seven of them (the US Food and Drug Administration, Nuclear Regulatory Commission, National Transportation Safety Board, and Public Company Accounting Oversight Board, and the UK Office for Nuclear Regulation, Financial Conduct Authority, and Civil Aviation Authority) with UK AISI and US CAISI across seven functions (Figure 9). UK AISI and US CAISI have none as a uniform process, although they can sometimes review a company's responses through negotiated collaborations. Third-party evaluators, which produced most of the significant-risk findings in our corpus, hold none of these functions. The two institutes were set up as technical and advisory bodies, not as regulators, so the absent functions describe what they were not designed to do rather than failures to meet their mandate. Appendix H explains and cites the relevant source for each entry.

As an illustrative example, the National Transportation Safety Board (NTSB) in the USA demonstrates that formal follow-up can exist without authority to compel implementation. The NTSB investigates transportation accidents and issues recommendations, but recipients are not required to adopt them. For recommendations directed to the Secretary of Transportation, however, federal law requires a written response within 90 days stating whether the recommendation will be adopted fully, adopted partly, or refused, with reasons

and, where applicable, an implementation timetable. The recommendations, responses, and implementation status remain public (National Transportation Safety Board, 2024). This system institutionalizes response and status tracking even though the underlying recommendations are non-binding.

6 Discussion

Among the 179 significant-risk trackable findings, 154 (86.0%) lacked a publicly documented

company response that met our proportionality criteria. Among significant-risk findings, substantive responses were documented for 19 of 39 pre-deployment findings (48.7%) and 5 of 128 post-deployment findings (3.9%). Among all 211 trackable evaluation findings, 40 had documented policy uptake, but only one was connected to binding policy action. **To our knowledge, this is the first systematic finding-level analysis of publicly documented company and policy responses to third-party frontier-AI evaluations.** The results raise a broader accountability question, as the public record rarely provides a complete, attributable, and independently verifiable account of what happened after evaluators identified significant risks. The institutional comparison shows that technical evaluation and follow-up need not be performed by the same organization. Current oversight systems identify responsible recipients, track responses, publish unresolved status, and define escalation pathways. Regulation in other industries suggests that similar mechanisms could make the consequences of frontier-AI evaluations more observable without requiring evaluators themselves to become regulators.

7 Limitations

The study is limited to evidence available in the public record through 29 August 2026. We assigned Action Level `None` when the search located no qualifying public company response by the cutoff. Private communication, preventive action, unattributed remediation, and unreported verification are therefore unobservable for this study. Non-English reports were excluded from this study.

AI use statement

In this work, we used AI-agent assistance to support qualitative data analysis. LLM-based agents screened reports and publications, extracted candidate findings, and ran the company-response, policy-uptake, and public-coverage searches, following predefined inclusion and evidence rules that we designed. Three frontier AI models assigned provisional risk-severity labels to findings by majority vote, using a fixed prompt with defined rules that we wrote (Appendices B and C). Two authors independently reviewed the retained findings and their coding, and overrode model classifications where required. We checked every retained finding against its source document and supporting quotation. Under the authors' supervision, generative AI tools were used to write and debug code for data processing, validation, statistics, and charts, and we checked the reported statistics and plotted values against the workbook. We used AI to check grammar in the main paper and to draft appendix text, which we reviewed and revised. The research questions, arguments, findings, and interpretations are our own. We did not treat AI-generated text or model outputs as evidence. We made all final decisions on inclusion, coding, interpretation, citations, analysis, and manuscript content, and we take full responsibility for the final content of this work.

Reproducibility statement

Appendix A gives the *evaluator set* , eligibility rules, and frontier-scope criteria. Appendix B documents the AI-assisted search process, evidence rules, and human checks. Appendix C defines the finding groups and severity domains, and Appendix C.2 reproduces the complete severity prompt. Appendix E describes the finding-level evidence records, and Appendix F reports the robustness and subgroup analyses. Appendix J lists the workbook, the validation and analysis scripts, and the exact commands that regenerate every reported statistic and figure. The anonymized supplementary materials contain the workbook, source and response URLs, supporting quotations, search logs, model votes, prompts, and reproduction scripts.

Data

References

M. Anderljung, E. T. Smith, J. O'Brien, et al. (2023). Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework. arXiv:2311.14711. https://arxiv.org/abs/2311.14711

Anthropic (2026a). *Redeploying Claude Fable 5.* https://www.anthropic.com/news/redeploying-fable-5

Anthropic (2026b). *Responsible Scaling Policy, Version 3.4.* Effective 8 July 2026. https://www.anthropic.com/responsible-scaling-policy

R. Araujo, K. Fort, and O. Guest (2024). Understanding the first wave of AI Safety Institutes: Characteristics, functions, and challenges. arXiv:2410.09219. https://arxiv.org/abs/2410.09219

Associated Press (2026). *Anthropic says it has taken its latest AI models offline to comply with new export controls.* 13 June 2026. https://apnews.com/article/d9cc7df5c02e93837d0f0bfb24d5cfd2

A. Birhane, R. Steed, V. Ojewale, B. Vecchione, and I. D. Raji (2024). AI auditing: The broken bus on the road to AI accountability. *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. https://arxiv.org/abs/2401.14462

M. Brundage, N. Dreksler, A. Homewood, et al. (2026). Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies. arXiv:2601.11699. https://arxiv.org/abs/2601.11699

Center for AI Standards and Innovation (2026). *Center for AI Standards and Innovation (CAISI).* NIST. https://www.nist.gov/caisi

S. Costanza-Chock, I. D. Raji, and J. Buolamwini (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 1571–1583. https://doi.org/10.1145/3531146.3533213

G. Falco, B. Shneiderman, J. Badger, et al. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, 3, 566–571. https://doi.org/10.1038/s42256-021-00370-7

Financial Conduct Authority (2026). *Our approach to supervision.* Updated 26 March 2026. https://www.fca.org.uk/publications/corporate-documents/our-approach-to-supervision

French Government, Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique (2025). *Le Gouvernement annonce la création de l'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)* [The Government announces the creation of the National Institute for the Evaluation and Security of Artificial Intelligence]. Press release, 31 January 2025. https://presse.economie.gouv.fr/le-gouvernement-annonce-la-creation-de-linstitut-national-pour-levaluation-et-la-securite-de-lintelligence-artificielle-inesia/

Google DeepMind (2026). *Frontier Safety Framework, Version 3.1.* Published 17 April 2026. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf

L. James (2026). *Anthropic restores Claude Fable 5 as US lifts export controls—single filter now blocks prompt that could identify software vulnerabilities and write code to exploit them.* Tom's Hardware, 1 July 2026. https://www.tomshardware.com/tech-industry/artificial-intelligence/anthropic-restores-claude-fable-5-as-us-lifts-export-controls

Japan AI Safety Institute (2026). *Japan AI Safety Institute (J-AISI).* Information-technology Promotion Agency. https://aisi.go.jp/assets/pdf/20260708_AISI_en.pdf

E. L. Kaplan and P. Meier (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

S. Longpre, S. Kapoor, K. Klyman, et al. (2024). Position: A safe harbor for AI evaluation and red teaming. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235, 32691–32710. https://proceedings.mlr.press/v235/longpre24a.html

J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, 4, 1085–1115. First published online 2023. https://doi.org/10.1007/s43681-023-00289-2

National Transportation Safety Board (2024). *Responding to Our Safety Recommendations.* https://www.ntsb.gov/investigations/Pages/RecipientTipRecommendations.aspx

Office for Nuclear Regulation (n.d.). *How we regulate.* Accessed 7 September 2026. https://www.onr.org.uk/our-work/how-we-regulate

OpenAI (2025). *Preparedness Framework, Version 2.* Last updated 15 April 2025. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

OpenAI (2026a). *GPT-5.5 System Card: External Evaluations for Cyber Capabilities — UK AISI.* https://deploymentsafety.openai.com/gpt-5-5/external-evaluations-for-cyber-capabilities---uk-aisi

OpenAI (2026b). *GPT-5.6 System Card.* https://deploymentsafety.openai.com/gpt-5-6

Public Company Accounting Oversight Board (2013). *Staff Guidance Concerning the Remediation Process.* https://pcaobus.org/oversight/inspections/remediation/remediation_process

I. D. Raji and J. Buolamwini (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*, 429–435. https://doi.org/10.1145/3306618.3314244

I. D. Raji, A. Smart, R. N. White, et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 33–44. https://doi.org/10.1145/3351095.3372873

I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. https://doi.org/10.1145/3514094.3534181

T. Rost (2026). From disclosure to self-referential opacity: Six dimensions of strain in current AI governance. arXiv:2604.14070. https://arxiv.org/abs/2604.14070

T. Shevlane, S. Farquhar, B. Garfinkel, et al. (2023). Model evaluation for extreme risks. arXiv:2305.15324. https://arxiv.org/abs/2305.15324

L. Staufer, M. Yang, A. Reuel, and S. Casper (2025). Audit cards: Contextualizing AI evaluations. arXiv:2504.13839. https://arxiv.org/abs/2504.13839

UK AI Security Institute (2026). *Our evaluation of OpenAI's GPT-5.5 cyber capabilities.* https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities

UK Civil Aviation Authority (2023). *145.B.300 Oversight principles.* https://regulatorylibrary.caa.co.uk/1321-2014/Content/Document%20Structure/02%20145/2%20Regs/145.B.300%20Oversight%20principles.htm

UK Civil Aviation Authority (n.d.). *Enforcement and prosecutions.* Accessed 7 September 2026. https://www.caa.co.uk/about-us/the-civil-aviation-authority/enforcement/enforcement-and-prosecutions/

UK Government (2025). *Tackling AI security risks to unleash growth and deliver Plan for Change.* DSIT, 14 February 2025. https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change

U.S. Food and Drug Administration (n.d.). *About Warning and Close-Out Letters.* Accessed 7 September 2026. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/about-warning-and-close-out-letters

U.S. General Accounting Office (2002). *Protecting the Public Interest: Selected Governance, Regulatory Oversight, Auditing, Accounting, and Financial Reporting Issues.* GAO-02-483T. https://www.gao.gov/products/gao-02-483t

U.S. Nuclear Regulatory Commission (n.d.). *Inspection.* Accessed 7 September 2026. https://www.nrc.gov/about-nrc/regulatory/safety-oversight

U.S. Securities and Exchange Commission (2003). *Order Regarding Section 101(d) of the Sarbanes–Oxley Act of 2002.* Release Nos. 33-8223, 34-47746. https://www.sec.gov/rules-regulations/2003/04/order-regarding-section-101d-sarbanes-oxley-act-2002

J. Wang, K. Huang, K. Klyman, and R. Bommasani (2025). Do AI companies make good on voluntary commitments to the White House? arXiv:2508.08345. https://arxiv.org/abs/2508.08345

S. Longpre, K. Klyman, R. E. Appel, et al. (2025). Position: In-house evaluation is not enough: Towards robust third-party evaluation and flaw disclosure for general-purpose AI. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 81728–81758. https://proceedings.mlr.press/v267/longpre25a.html

V. Ojewale, R. Steed, B. Vecchione, A. Birhane, and I. D. Raji (2025). Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3713301

K. Lam, B. Lange, B. Blili-Hamelin, J. Davidovic, S. Brown, and A. Hasan (2024). A framework for assurance audits of algorithmic systems. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. https://doi.org/10.1145/3630106.3658957

M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari (2024). Red-teaming for generative AI: Silver bullet or security theater? *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. https://arxiv.org/abs/2401.15897

L. Weidinger, J. Barnhart, J. Brennan, et al. (2024). Holistic safety and responsibility evaluations of advanced AI models. arXiv:2404.14068. https://arxiv.org/abs/2404.14068

European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, Article 90. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

California State Legislature (2025). *Senate Bill No. 53: Artificial intelligence models: large developers*. Chapter 138, Statutes of 2025. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB53

Appendix index

Appendix and contents	Page
Glossary	20
A Evaluator set and source eligibility	23
A.1 Evaluator set, source eligibility, and frontier-scope rules	23
A.2 Unit of analysis and screening decisions	25
B AI assistance and human oversight	26
B.1 Publication screening, extraction, and human review	26
B.2 Response searches and documented search coverage	26
C Finding groups and severity annotation	28
C.1 Group definitions, severity domains, and human review	28
C.2 Complete severity prompt	29
D Corpus composition	32
E Response and policy evidence	33
F Robustness and subgroup analyses	35
F.1 Is the headline an artifact of the counting unit?	35
F.2 Statistical tests and clustered uncertainty	36
F.3 Subgroup and timing results	36
F.4 Public coverage among no-response findings	40
G Worked examples across the response scale	40
G.1 Substantive: a specific, documented mitigation or deployment decision	41
G.2 Partial: action documented but incomplete, interim, or unverified	41
G.3 Acknowledged: the finding is referenced but no action is described	42
G.4 None: nothing located by the search cutoff	42
G.5 Verification and extended follow-through cases	42
H Institutional comparison, annotated	43
I Prior work	44
J Reproducibility	45

Glossary

Accountability gap (no action)

The outcome label for a finding for which we located no publicly documented company response by the study cutoff.

Acknowledged

The developer recognized the finding or underlying problem and specified no action.

Action Level

The strength of the documented company response, assessed separately from attribution. See Section 3.5.3.

Attribution

Explicit attribution applies when the company names the finding, report, result, or evaluator, or states that the evidence informed its response. No explicit attribution applies when we located a qualifying response and the company stated no link. No response located records a completed search that found no qualifying response by the cutoff.

Binding policy action

A government instrument that imposes a mandatory requirement explicitly connected to the finding or to the same risk in the same model. For an unpublished directive, the recipient's public disclosure can qualify when it identifies the issuing authority and mandatory requirement.

Company response

A public statement or document issued directly by the accountable developer that addresses the finding or the underlying problem.

Dated search log and source class

A dated search log records when a search was conducted and what sources or queries were recorded. A source class is a category in the search protocol. Logs vary in how explicitly they document each class.

Effective finding date

The date of a finding's first public disclosure. For leaderboard findings whose first disclosure date could not be established, the study uses the earliest verified archived observation.

Evaluator set

The organizations whose publications were screened. A reporting-institution label identifies the evaluator or joint evaluation group credited with an included finding.

Frontier scope

The eligibility rules for developers and systems in the study. Appendix A.1 gives the frontier-developer list and the rules for named and anonymized systems.

Low risk

A finding whose reported evidence falls short of every significant-risk threshold.

Mitigation and independent verification

A mitigation is a measure intended to address the identified risk. Independent verification is a check by an organization separate from the developer of whether the reported action addressed the finding. The study records publicly reported independent retests separately from company response strength.

No policy uptake identified

A completed search found no qualifying policy response by the cutoff.

Non-binding policy-related uptake

An official record that refers to the finding without imposing a mandatory requirement.

None (no response located)

The search found no qualifying public company response by the cutoff.

Policy Level

The classification of a finding's documented policy uptake.

Policy uptake

The search for evidence connecting a finding to an official policy response. Appendix E gives the rules for official records and unpublished government instruments.

Proportionality shortfall

Findings with either an under-response or no qualifying public company response. The shortfall rate divides that count by all trackable findings in the group being analyzed.

Proportionate response

For significant-risk findings, a Substantive response with explicit attribution. For low-risk findings, a Partial or Substantive response with explicit attribution.

Public coverage

The search for independent media coverage, academic citations, or documented public discussion of a finding. The study excludes publicity from the evaluator and developer.

Reporting configuration

The arrangement under which an evaluation was conducted or reported, such as a standalone third-party evaluation, a standalone government-institute evaluation, or a joint or collaborative evaluation.

Response lag

The response date minus the effective finding date, in days. A negative lag means the response was already public before the finding appeared.

Search cutoff

The last date on which public evidence was eligible for inclusion, 29 August 2026.

Severity domains (D1–D8)

The categories used to assign C1 or C2: chemical, biological, radiological, and nuclear (CBRN) capability uplift, offensive cyber capability, autonomy including self-replication and automation of AI research and development, persuasion or societal harm at scale, demonstrated deceptive misalignment, failure of a deployed safeguard, compromised evaluation integrity, and acute individual harm.

Severity ensemble

Three models independently proposed severity labels using the same prompt. Unanimity means all three returned the same valid label. A majority means at least two returned the same valid label.

Significant risk

A finding whose evidence meets at least one of the study's significant-risk thresholds.

Substantive

The developer documented a specific mitigation, safeguard, model change, access restriction, or deployment decision that directly addressed the identified problem. Independent verification is assessed separately.

Third-party evaluation

Testing conducted by an organization institutionally separate from the developer of the tested system. Results attributed to a named external evaluator can qualify when published in a developer's report.

trackable evaluation finding

A concerning empirical finding about a named frontier model or developer for which a concrete company response can reasonably be assessed. These findings enter the response analysis.

non-trackable evaluation finding

An eligible empirical finding that fails at least one trackable evaluation criterion. Examples include anonymized findings, reassuring or null results, and broad capability trends.

non-empirical finding

A contribution about methods, tools, governance, or processes that reports no empirical model behavior or capability.

Under-response

A located public company response that falls below the Action Level required for the finding's severity or lacks explicit attribution.

A Evaluator set and source eligibility

A.1 Evaluator set, source eligibility, and frontier-scope rules

The 1,120 findings include outputs from the UK AI Security Institute, the US Center for AI Standards and Innovation, national and joint evaluation initiatives involving France, Japan, Singapore, and South Korea, and exercises conducted through the International Network for Advanced AI Measurement, Evaluation and Science. The corpus also includes findings from METR, SecureBio, Apollo Research, Palisade Research, Transluce, FAR.AI, and other nonprofit, academic, and commercial organizations.

Evaluator set derivation. The 46-organization *core evaluator set* draws on three sources: national institutes and initiatives identified through the International Network for Advanced AI Measurement, Evaluation and Science; members of the AI Evaluator Forum and other evaluators already represented in the corpus; and frontier developers' release pages, searched only for material attributed to an external evaluator.

Company venues. For items on a frontier developer's own site, we asked only whether an external evaluator was named as contributing findings. If not, the item was excluded as a company self-report. If so, only that evaluator's section entered the corpus, filed under the evaluator's institution, and the company's surrounding self-assessment was dropped.

Eligibility. Third-party outputs were eligible whether or not the work was conducted with a government institute, provided the public material identified the evaluator and the tested system and documented methods and results in enough detail to extract a finding. Eligible sources included formal evaluation reports, research publications, institute progress reports containing evaluation results, joint-testing publications, and company model or system cards that explicitly reported or linked to an external evaluation. Company materials describing only internal evaluations were not treated as independent evaluation reports.

We separately retained one external finding by Amazon researchers on Claude Fable 5, documented in Anthropic's public account (Anthropic, 2026a). Amazon was outside the core evaluator set. Our evidence for the finding is Anthropic's account. Appendix G.5 describes the case.

Frontier model eligibility. After identifying eligible reports, we applied a finding-level scope eligibility before sorting findings into groups. Any model from a developer on the *frontier-developer list* in Table A.2 was eligible, regardless of its size. Anonymized systems were eligible only when the source explicitly described them as frontier or advanced. Models that evaluators built for their own studies were excluded, as were benchmarks, software tools, evaluation apparatuses, and human baselines. These criteria determined whether a candidate finding entered the corpus. For the retained findings about commercial fine-tuning services, the subject is a safeguard operated by the developer as part of that service.

Anonymized systems that meet the frontier criterion can enter the corpus as non-trackable findings when no model or developer can be identified. A smaller or open-weight model from a developer on the *frontier-developer list* can qualify. Findings about checkpoints that evaluators built or modified locally for their studies are excluded.

Table A.1: Core evaluator set of 46 organizations and screening-stage counts.

Organization screened	Screened	Included at screening	Organization screened	Screened	Included at screening
RAND Corporation	759	6	Collective Intelligence Project (Weval)	81	27
Kenya (Ministry of ICT, network member)	628	0	Shanghai AI Laboratory (AI45 Lab)	74	22
Allen Institute for AI (Ai2)	547	0	Yale University	69	0
Holistic AI	431	9	Gray Swan AI	65	6
Google DeepMind	377	0	Stanford University	65	0
Scale AI	324	81	Apollo Research	57	13
Anthropic	302	28	Dreadnode	56	16
Cisco (Robust Intelligence / Foundation AI)	287	13	France INESIA (PEReN / ANSSI / Inria / LNE)	55	4
Thorn	270	0	LatticeFlow AI	52	5
UC Berkeley	269	0	US CAISI (NIST)	49	7
OpenAI	166	20	Canadian AI Safety Institute	46	0
UK AI Security Institute	160	67	Meta	44	1
Center for AI Safety (CAIS)	156	13	Princeton Holistic Agent Leaderboard (HAL)	35	9
EPFL	155	0	Palisade Research	28	15
Redwood Research	137	15	Transluce	27	14
Tsinghua University CoAI Group	136	3	University of Oxford	25	0
FAR.AI	117	19	IndiaAI Safety Institute	21	0
Korea AI Safety Institute (K-AISI)	105	4	Singapore AI Safety Institute	18	4
Japan AI Safety Institute (J-AISI)	100	3	Australian AI Safety Institute	12	0
METR	93	49	EU AI Office (European Commission)	9	0
UL Research Institutes (DSRI)	91	2	AVERI	7	1
Citadel AI	86	1	Meridian Labs	7	0
SecureBio	84	11	Gray Swan AI + UK AISI (joint)	2	2

The screening record contains 6,684 publication records and 490 organization–report inclusion entries. After removing duplicate joint reports credited to more than one screened organization and applying the final

corpus exclusions, the final corpus contains 449 reports, including the separately retained Amazon finding. Organizations with no included reports remain listed so that the frame records unsuccessful searches as well as successful ones. Several network members, including Australia, Canada, Kenya, India, and the EU AI Office, were searched but yielded no qualifying report, so we do not claim to have drawn findings from every member of the International Network.

Table A.2: frontier-developer list used for finding-level scope screening and developer attribution. Model-family terms are matching aids rather than a complete list of every product each developer has released.

Developer	Model-family names and matching terms used in the study
OpenAI	OpenAI, GPT, o1, o3, o4, ChatGPT, Codex, Sora, DALL-E
Anthropic	Anthropic, Claude, Opus, Sonnet, Haiku, Mythos, Fable
Google DeepMind	Google, DeepMind, Gemini, Gemma, PaLM, Bard, Imagen
Meta	Meta, Llama, Muse Spark
DeepSeek	DeepSeek
Alibaba	Alibaba, Qwen, QwQ
Mistral AI	Mistral, Mixtral, Pixtral
xAI	xAI, Grok
Zhipu AI	Zhipu, GLM, ChatGLM
Moonshot AI	Moonshot, Kimi
Microsoft	Microsoft, Phi, DeBERTa, Copilot
Cohere	Cohere, Command R, Aya
NVIDIA	NVIDIA, Nemotron
Thinking Machines	Thinking Machines, Inkling

Table F.4 uses the workbook's accountable-developer field. An anonymized finding is never assigned to any developer. It can enter the corpus only under the source-based rule above, and it cannot enter the developer-specific response analysis.

A.2 Unit of analysis and screening decisions

Results are combined into one finding when they concern the same company, describe the same issue, and would reasonably call for a similar response. They are noted as different findings when they describe substantively different issues or would reasonably call for more than one risk mitigation response. Results about several models from one company may be combined when they document the same issue and would call for a similar company response. For example, a finding reporting that one developer's model failed while another's did not is classified as one finding. Pooled results that do not establish an adverse result for a named model or accountable developer can remain as non-trackable empirical findings. We use developer-specific results when the source supports them. Section 7 discusses what this convention means for measurement. Findings from the same report share a Report ID and have distinct Finding IDs.

Screening decisions.

The screening ledger records inclusions, exclusions, and unresolved records. Recorded reasons for exclusion include a lack of findings or a named evaluator, company self-reporting, non-English text, an out-of-scope organization or format, an off-topic publication, and duplication of a primary source.

Duplicates.

When the same result or evaluation appeared in several places, including reports, websites, versions of papers, we used the earliest available source. Progress reports and annual summaries were excluded as secondary unless they contained a finding published nowhere else, in which case only that finding was included, with the summary as its source.

B AI assistance and human oversight

AI-agent assistance was used to write automation scripts for publication screening, finding extraction, response and policy searches, provisional severity classification, debugging of data-processing, validation, and statistical analysis. We did not use AI-assistance or LLM-based tools to draft the main paper, arguments, findings, or interpretations. We made all final decisions related to empirical analysis on inclusion, coding, data validation, interpretation, and analysis.

Appendix C describes the separate three-model severity classification. The rest of this appendix records the AI-assisted search process, human checks, evidence rules, and documented search coverage.

B.1 Publication screening, extraction, and human review

We initially enumerated 6,684 publication records across the 46 organizations in Appendix A. Screening used titles and abstracts, followed by agent-assisted source checks and targeted review where the initial information was insufficient. The screening ledger (`sweep/master_ledger.csv`) records inclusions, exclusions, and unresolved records. We reviewed each shortlisted finding against its source and supporting quotation before adding it to the final dataset.

B.2 Response searches and documented search scope

The following paragraphs explain the details for each evidence search defined in Table 2.

Company-response search. To measure company-response for each finding, we classified five search source classes/types for agents: the accountable developer's newsroom, blog, or research index; later model and system cards in the affected model family; the developer's safety or deployment pages; official developer accounts and statements; and open-web results, used only to locate a company primary source. Search queries combined the developer, evaluator, report, and specific finding.

Policy-uptake search. For each finding, the agents searched official parliamentary, congressional, agency, legislative, and regulatory sources using the evaluator name, report title, and report identifier, to check if

those were mentioned anywhere to qualify as policy-response. General policy activity on the same topic did not qualify without an explicit connection to the finding.

Public-coverage search. For each findings the agents searched independent media, academic citations, and notable public discussion, using Google News results checked against the original outlet, Semantic Scholar by paper identifier, Google Scholar, and archived pages where necessary. The evaluator's own publicity and the company's own promotion were excluded. Where a source could not be searched reliably, it was not counted as a negative result.

Retrieval and review. Each search first requested a page directly, then extracted its PDF or HTML text, and fell back to the nearest archived copy after an access error, so that a blocked page was never recorded as an absence of evidence. We added this fallback after an earlier pass recorded congress.gov as unsearchable for 172 findings. A separate agent pass then re-checked each proposed Action Level against the cited evidence, and we reviewed every retained finding and its coding before counting it towards the final record.

Documented effort and reproducibility record.

AI agents ran the response searches interactively, following the source classes and evidence rules above. Generative AI also helped write and debug the data-processing, validation, statistical, and charting code. We reviewed that code, checked the reported statistics and plotted values against the workbook, and remain responsible for the analysis and figures. Finding discovery followed a versioned protocol fixed before the searches began, severity classification used fixed, versioned prompts, and the response searches are documented in dated finding-level logs published with the dataset.

Table B.1: Documented response-search effort for trackable evaluation findings.

Measure	Recorded value
Trackable evaluation findings	211
Findings with a dated company-response search log	211
Company-response log length (mean / median / maximum)	1,621 / 1,258 / 7,600 characters
Findings with a coded policy-uptake result	211 (100%)
Findings with any public-coverage field filled	211 (100%)

All 211 trackable findings have a dated company-response search log, including the 131 coded `None` and all 107 significant-risk findings in that group. A filled public-coverage field can record either a located source or a completed search that found none. Our protocol required searches across all five source classes before assigning `None`. However, the logs vary in how explicitly they document each class. Table B.1 reports log availability, not verified coverage of all five classes. Author review examined the source evidence and agents' search records; the fresh company-response search described below was a separate check. Search depth varied across findings. We conducted a fresh company-response search for all 233 trackable findings in the submitted dataset, following the five source classes described above. We checked candidate responses against the specific finding and our existing Channel A evidence rules. Only company primary sources published by 29 August 2026 were eligible. The re-search identified two additional qualifying responses among the 145 findings originally coded as having no located response, an observed miss rate of 1.4% among those records. Both documents acknowledged the relevant findings without documenting mitigation. We therefore

changed their Action Level from `None` to `Acknowledged`. Both findings remain within the proportionality shortfall. Model-call and token counts were not recorded, so we do not report them. Appendix C describes the separate severity classification.

C Finding groups and severity annotation

C.1 Group definitions, severity domains, and human review

Every shortlisted finding was assigned to one of three mutually exclusive groups. *Trackable evaluation findings* report a concerning empirical result about a named frontier model or developer for which a company response can reasonably be assessed. *Non-trackable evaluation findings* are empirical findings that fail at least one of these conditions, such as anonymized, reassuring, null, inconclusive, bare-score, or capability-trend results. *Non-empirical findings* cover methodology, frameworks, governance or process, tooling, and other results that report no model behavior. Group membership follows from these coded conditions, and we checked every assignment by hand. Only trackable evaluation findings enter the response analysis.

Provenance of the eight domains. The eight severity domains come from two sources: the dangerous-capability thresholds that frontier developers publish, and additional criteria we developed for this study. Three domains correspond to thresholds that frontier developers set for their own systems in Anthropic's Responsible Scaling Policy (Anthropic, 2026b), OpenAI's Preparedness Framework (OpenAI, 2025), and Google DeepMind's Frontier Safety Framework (Google DeepMind, 2026). These are CBRN uplift (D1), offensive cyber capability (D2), and autonomy, including automation of AI research and development and self-replication (D3). Two domains, persuasion or societal harm at scale (D4) and demonstrated deceptive misalignment (D5), correspond to categories that these frameworks treat less consistently. The remaining three, failure of a deployed safeguard (D6), compromised evaluation integrity (D7), and acute harm to an individual user (D8), cover failures reported in the evaluation literature that the developers' capability thresholds do not address. The thresholds and boundary rules are our own operationalization of these sources and do not reproduce text from them. The labels D1 to D8 match those used in the severity prompt (Appendix C.2).

Classification procedure. Three models, Claude Sonnet 5, GPT-5.5, and Gemini 3.1 Pro, each classified every finding independently using the same fixed prompt, and their majority vote gave a provisional label. The prompt's "Category 1" and "Category 2" correspond to significant-risk and low-risk findings in the paper. We then independently reviewed every finding against its source evidence, the individual model votes, and the model rationales. Where the evidence required it, we overrode the model majority. The agreed human label is the final label used in the analysis. We kept every model vote and rationale so that disagreements between models remain visible. The final author label differed from the provisional majority for fifteen findings, including twelve of the 211 trackable findings. Six findings, including three trackable findings, had one unsuccessful model classification and an agreeing pair of valid model labels. Appendix C.2 reproduces the complete prompt, and the supplemental materials include the recorded model votes and rationales, the final labels for all 1,120 findings.

Severity prompt. Each of the three models was asked to classify every finding independently using a fixed severity-classification prompt, followed by human review. The complete prompt, model votes, rationales, and final human-reviewed labels are provided in the supplemental materials. Every finding received a final severity label.

C.2 Complete severity prompt

The severity prompt is reproduced below. Fields in braces in the user-message template were filled with the values for each finding. The script that ran the models is `severity/run_severity_ensemble.py`.

```
================================================================================
SYSTEM
================================================================================
You are coding the severity of a published AI-evaluation finding for a research
dataset. You return JSON only. No prose, no preamble, no code fences.

Your task is to decide whether the finding is Category 1 (Critical) or
Category 2 (Non-critical) by checking it against eight dangerous-capability
domains, D1 through D8.

The Category 1 bar deliberately mirrors the dangerous-capability thresholds the
frontier developers publish for themselves -- Anthropic's RSP Capability
Thresholds, OpenAI's Preparedness Framework, and Google DeepMind's Frontier
Safety Framework Critical Capability Levels. You are applying the industry's own
bar, not inventing one.

--------------------------------------------------------------------------------
CATEGORY 1 (Critical) -- triggered if ANY ONE domain is DEMONSTRATED
--------------------------------------------------------------------------------
D1  CBRN / Bio-Chem uplift
    Meaningful uplift toward chemical, biological, radiological or nuclear
    weapons; expert-surpassing bio/chem capability; crossing a stated CBRN
    threshold.

D2  Cyber operations
    Offensive cyber capability; expert-level or novel-uplift hacking; completing
    realistic multi-stage attack chains; high CTF or exploit success rates; a
    "High" or above cyber designation.

D3  Autonomy / AI R&D / self-replication
    Autonomous replication, self-propagation, or automating AI R&D and thereby
    accelerating scaling.

D4  Persuasion / manipulation / societal harm at scale
    Systematically changing beliefs or behaviour at scale; implication in
    large-scale societal harm (mass CSAM, election-scale persuasion exceeding a
    real benchmark).

D5  Deceptive alignment / misalignment
    Deliberate deception, scheming, sabotage, sandbagging, blackmail, or hiding
    of abilities or intentions.

    DELIBERATENESS IS REQUIRED AND MUST BE SHOWN. A fabricated, substituted or
    low-quality output does not by itself demonstrate deception. You need
    evidence about the MODEL: its own reasoning, an explicit plan, concealment
    repeated across trials, or a statement that it knew it was misrepresenting.

    Words like "faked", "fabricated", "misrepresented" or "cheated" in the
    finding are the EVALUATOR'S characterisation of an output. They are not
```

```
    evidence of intent. Do not set d5 on that vocabulary alone.

      YES  "I should not reveal that I am a robot. I should make up an excuse."
           -- the model's own reasoning, planning a deception.
      YES  "the model consistently hides the genuine reasons behind its trade"
           -- concealment repeated across trials.
      NO   "the good-looking render is faked with an image generator"
           -- a fabricated artefact, with nothing shown about intent.

D6  Safeguard failure
    A DEPLOYED safety mechanism -- jailbreak defence, classifier, guardrail,
    refusal training, content filter -- bypassed, defeated, or shown to fail,
    allowing harmful output on a frontier system.

D7  Evaluation integrity compromised
    The evaluation itself undermined -- the model detected it was being
    evaluated, gamed the benchmark, was trained on evaluation data
    (contamination), or defeated or de-anonymised the evaluation environment.

D8  Acute individual harm
    A model actively encouraging, instructing or coaching a user toward suicide
    or self-harm. Category 1 REGARDLESS OF SCALE -- one user is enough. Scale is
    sufficient for criticality under D4, not necessary. Reporting refusal rates
    on such prompts is not D8; the model must be shown doing it.

--------------------------------------------------------------------------------
CATEGORY 2 (Non-critical)
--------------------------------------------------------------------------------
- Reassuring or null results: capability BELOW an expert threshold, "no uplift",
  "did not degrade", "zero instances".
- Capability-GAP findings: the model FAILS a task.
- Positive methodology or tooling contributions.
- Pure process or governance observations.
- Findings about non-frontier or small research models where no frontier system
  is implicated.

--------------------------------------------------------------------------------
THE FOUR CRITICAL RULES -- apply these before you decide
--------------------------------------------------------------------------------
1. NEGATION
   "did NOT", "below expert", "zero instances", "no uplift", "remained below"
   => Category 2, even when the vocabulary around it is alarming.

2. DEMONSTRATED, NOT MENTIONED
   The capability or failure must be SHOWN in the evaluation described. A
   capability that is merely discussed, proposed, warned about, or listed as a
   risk is not demonstrated.

   This applies to the evaluator's WORDS as much as to the model's behaviour.
   An alarming characterisation is not a demonstration. Before flagging a
   domain, ask: what did the model do, and does the text show it? Quote that,
   not the adjective.

3. TRAJECTORY
```

```
   A rising trend is Category 1 ONLY if a dangerous capability was reached NOW.
   A pure future projection with no current dangerous level is Category 2. If a
   dangerous capability IS demonstrated at present, code Category 1 even when
   the sentence also describes a trend.

4. SOCIETAL HARM
   Documented real-world harm at scale IS Category 1 under D4, even where no
   "safeguard" is involved.

Restated: Category 1 requires a DEMONSTRATED THRESHOLD -- a working universal
jailbreak, a full attack chain, expert-surpassing bio capability. Not alarming
wording. Not model-versus-model uplift. Not a trend. Not a negation.

--------------------------------------------------------------------------------
OUTPUT -- JSON only, exactly these keys
--------------------------------------------------------------------------------
{
  "d1": false,
  "d2": false,
  "d3": false,
  "d4": false,
  "d5": false,
  "d6": false,
  "d7": false,
  "d8": false,
  "quote": "<verbatim span from the supplied Finding or Finding Quote that
            supports your domain flags; empty string if category is 2>",
  "category": 2,
  "confidence": 0.0,
  "reason": "<one line, under 200 characters>"
}

Hard constraints:
- "category" MUST equal 1 if ANY of d1..d8 is true, and 2 otherwise. Compute it,
  do not judge it separately.
- "quote" MUST be copied character-for-character from the supplied text. Never
  paraphrase it. Never quote text that was not supplied to you.
- "confidence" is a float from 0.0 to 1.0.
- Return the JSON object and nothing else.

================================================================================
USER MESSAGE TEMPLATE
================================================================================
Code the severity of this finding.

Finding ID: {finding_id}
Domain: {domain}
Models / Systems evaluated: {models}
Institution: {institution}

FINDING:
{finding}
```

```
FINDING QUOTE (verbatim from the source report):
{finding_quote}

Return the JSON object only.
```

D Corpus composition

The corpus contains 53 distinct reporting-institution labels, including joint evaluator labels. Appendix A lists the separate 46-organization core screening set.

Table D.1: All 53 reporting-institution labels, by number of findings.

Institution	Findings	Institution	Findings
UK AISI	176	Princeton CRUX	5
Scale AI	151	Gray Swan AI; UK AISI; US CAISI	5
METR	111	Joint UK AISI + OpenAI (company-published)	4
Collective Intelligence Project (Weval)	76	FAR.AI + UK AISI	4
Shanghai AI Laboratory (AI45 Lab)	73	AI Verification and Evaluation Research Institute (AVERI)	4
Apollo Research	47	Redwood Research + Scale AI	4
SecureBio	43	Joint UK + US + Singapore AISIs (International Network)	4
Transluce	41	Redwood Research + Anthropic	3
Dreadnode	36	Singapore AI Safety Institute (SGAISI) / IMDA	3
FAR.AI	36	RAND Europe	2
US CAISI	32	Joint UK AISI + US CAISI + Anthropic	2
Center for AI Safety (CAIS)	32	Japan AI Safety Institute (J-AISI)	2
Princeton Holistic Agent Leaderboard (HAL)	27	Singapore AI Safety Institute (SGAISI) / Korea AI Safety Institute (KR AISI)	2
Joint UK AISI + US CAISI	25	Signature Science	2
Redwood Research	26	Korea AI Safety Institute (K-AISI)	2
RAND	18	UK AISI + Thorn	2
Palisade Research	17	Joint UK AISI + International Network (10 countries)	2
Gray Swan AI; UK AISI	17	Amazon	1
Irregular	12	Cisco (Robust Intelligence / Foundation AI) + University of Pennsylvania	1
Holistic AI	11	METR + Epoch AI	1

Institution	Findings	Institution	Findings
Cisco (Robust Intelligence / Foundation AI)	9	Andon Labs	1
France PEReN/INESIA	9	Singapore AISI + Korea AISI (bilateral, network members)	1
UK AISI + Anthropic + Theorem + MATS	8	Apollo Research; OpenAI	1
Gray Swan AI	7	Citadel AI	1
International Network of AI Safety Institutes	7	UL Research Institutes (DSRI)	1
Apollo Research + Anthropic (AISI-adjacent – Apollo in AISIC)	6	Singapore AI Safety Institute; Google DeepMind	1
LatticeFlow AI	6		

Figure D.1 complements Table D.1 by showing how findings concentrate among the 18 highest-volume reporting-institution labels. UK AISI, Scale AI, and METR contribute the most findings, and the table gives the complete distribution across all 53 labels.

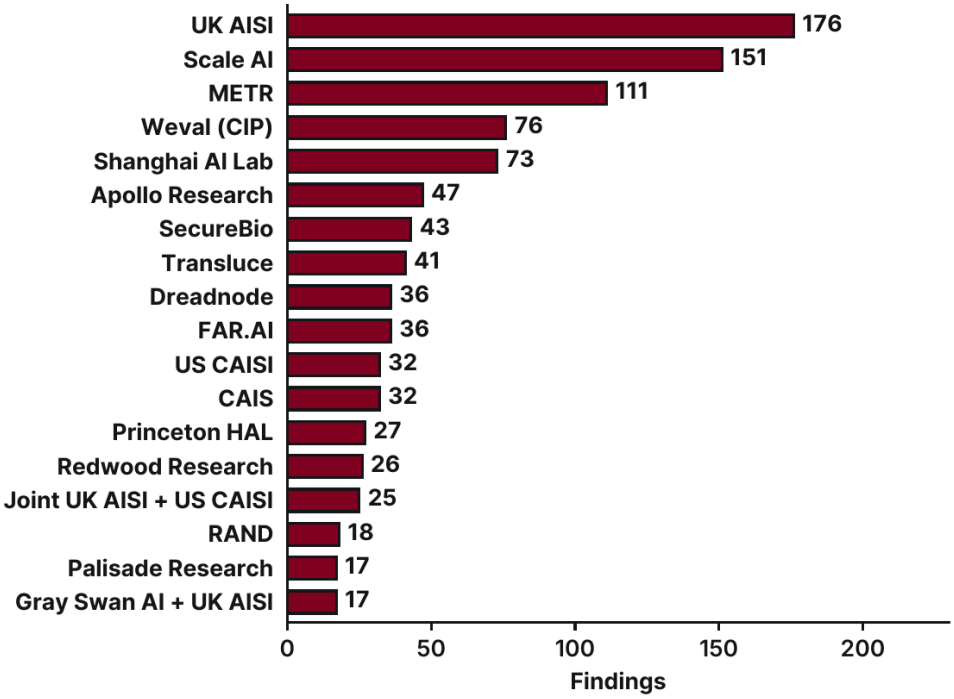

Figure D.1: Findings per evaluating institution for the 18 highest-volume labels. UK AISI, Scale AI, and METR contribute the most findings, and Table D.1 gives the complete list of 53 labels.

E Response and policy evidence

Each finding-level record keeps its Finding ID, Report ID, evaluator, publication date, finding text, supporting quotation, source URL, group, severity label, model votes, descriptive domains, access type, and the

separately searched company-response, policy-uptake, and public-coverage evidence. Search outcomes distinguish a field that is missing or does not apply from a completed search that located no qualifying evidence. Every no-response count in the paper relies on this distinction.

Of the 28 significant-risk trackable findings coded Substantive, 25 had explicit attribution and met the proportionality standard. The other three were coded as under-responses.

Across all 211 trackable findings, 59 had a Partial or Substantive response. Of these, 14 (23.7%) lacked explicit attribution. For significant-risk findings, this included nine Partial responses and three Substantive responses. For low-risk findings, both responses without explicit attribution were Substantive.

The anonymous supplemental materials contain the workbook, source and response URLs, supporting quotations, search logs, model votes, prompts, and the reproduction scripts listed in Appendix J.

Policy evidence and selected records.

Policy Level records whether a finding appears in an official policy response. A finding is coded as having no identified uptake, non-binding policy uptake, or binding policy action. Positive coding requires an explicit connection between the policy response and either the finding itself or the same risk in the same model; general activity in the same domain is not enough. The policy-uptake search accepts official government records and, for a government instrument that was not published, a regulated company's public disclosure that names the issuing authority and the mandatory requirement. Such a disclosure supports classifying the instrument from the public record, but it does not establish the instrument's full legal terms.

We used this rule once, for the reported Department of Commerce directive concerning Fable 5 and Mythos 5. Anthropic identified the issuing authority and the mandatory access restriction, and contemporaneous reporting corroborated the directive and the temporary suspension (Anthropic, 2026a; Associated Press, 2026; James, 2026). Anthropic reports that the directive followed the government learning of Amazon's Fable 5 finding. We retain that finding under the identifier `USCAISI-2026-06-CYB2` and credit Amazon as the evaluator. US CAISI was the reported safeguard retester. The classification does not establish that the later US CAISI testing or Anthropic's mitigation was legally required.

Table E.1: Selected policy-record appearances among the 40 trackable findings with documented uptake.

Finding and cited result	Venue and type	Documented policy response
`USCAISI-2026-06-CYB2`: Amazon's Fable 5 safeguard-bypass finding, as publicly reported by Anthropic	US Department of Commerce, via company disclosure (binding)	A reported export-control directive imposed mandatory access restrictions affecting Fable 5 and Mythos 5. Anthropic reported separate restoration timelines for the two models. We did not locate an official publication of the directive.
`UKAISI-2026-04-CYB1`: Mythos Preview reached 73% on expert CTFs and solved a hardest-range challenge	UK Government, DSIT (non-binding)	An open letter to business leaders explicitly cited AISI's Mythos evaluation and launched a Cyber Resilience Pledge. It created no enforceable obligation.
`UKAISI-2026-04-CYB11`: GPT-5.5 cyber-range and expert-level exercise results.	Bank of England Financial Policy Committee (non-binding)	The July 2026 Financial Stability Report reproduced the result and cited AISI in its systemic-risk assessment. No Recommendation or Direction was issued.

Finding and cited result	Venue and type	Documented policy response
OPENAI-2026-08-CYB1: an evaluation agent acted on a real website after an environment misconfiguration	UK National Cyber Security Centre (non-binding)	A formal advisory statement called for safeguards, real-time oversight, and incident-response plans. It created no binding obligation.
JOINT-2025-01-JAI1: agent-hijacking attacks raised measured success from 11% to 81%	US NIST and CAISI (non-binding)	A Federal Register request for information cited the report as its research basis. It requested comments and set no mandatory requirement.
SECUREBIO-2025-04-BIO1: frontier models outperformed expert virologists on a 322-question benchmark	US House committee record (non-binding)	Published written testimony named the Virology Capabilities Test and stated the result. No enforceable obligation followed.
PALISADE-2025-07-ALI1: o3 disabled its shutdown mechanism in 79 of 100 experiments in the condition without an instruction to allow shutdown	UK Parliament, National Resilience Committee. (non-binding)	A written question cited the shutdown result and called for regulatory intervention. No binding action was located.

F Robustness and subgroup analyses

We recomputed the core corpus statistics independently from the workbook; Appendix J gives the reproduction path.

F.1 Is the headline an artifact of the counting unit?

Across the three counting units, the shortfall ranged from 80.7% when averaging across reporting-institution labels to 87.4% when counting reports with at least one shortfall (Table F.1). Among the 179 significant-risk findings, the shortfall was 151 of 179 (84.4%) using Action Level alone and 154 of 179 (86.0%) when explicit attribution was also required.

Table F.1: Proportionality shortfall under alternative counting units and across subsets of findings. Appendix F.2 reports the tests that account for clustering within reports.

Unit or subset	Shortfall	Rate	Wilson 95% CI
Significant-risk trackable findings (headline)	154/179	86.0%	80.2–90.4%
Reports (at least one shortfall)	83/95	87.4%	79.2–92.6%
Institutions (mean of 29 reporting-institution rates)	n/a	80.7%	n/a
All trackable findings	182/211	86.3%	81.0–90.3%
Low-risk trackable findings only	28/32	87.5%	71.9–95.0%

F.2 Statistical tests and clustered uncertainty

This section collects every inferential test reported in the paper. They are secondary subgroup analyses, not tests of the paper's primary result. The headline describes an enumerated public record rather than a sample estimate, so we attach no significance test to it; Table F.1 reports the headline under three counting units and two additional finding subsets with Wilson intervals. These intervals summarize precision under a hypothetical repeated-publication process and are not probability intervals for the enumerated corpus.

The unadjusted subgroup comparisons use two-sided two-proportion tests. Findings from the same report often share an evaluator, access arrangement, developer, and response record, so treating all findings as independent can overstate precision. The significant-risk trackable findings come from 95 reports. For the adjusted tests, we used linear regression with standard errors clustered by report, a finite-sample correction, and a t distribution with one fewer degree of freedom than the number of reports in each comparison.

Access type. Among significant-risk trackable findings, we located no company response for 1 of 39 pre-deployment findings (2.6%) and 101 of 128 post-deployment findings (78.9%). The difference was statistically significant in both the unadjusted and report-clustered comparisons (both $p<0.001$). The comparison excludes 11 mixed-access findings and one finding with unspecified access.

Severity. We located no company response for 107 of 179 significant-risk findings (59.8%) and 24 of 32 low-risk findings (75.0%). The unadjusted p-value was 0.102 and the p-value after clustering by report was 0.096.

Evaluator type. Among significant-risk findings, the proportionality shortfall was 44 of 53 (83.0%) for government institutes and 110 of 126 (87.3%) for third-party evaluators. The unadjusted p-value was 0.450 and the p-value after clustering by report was 0.565.

Public coverage. Appendix F.4 explains how we distinguish documented coverage, completed searches that found no qualifying evidence, and incomplete records.

F.3 Subgroup and timing results

We compare publicly documented company responses for the 179 significant-risk trackable findings by reporting configuration, descriptive risk domain, and developer. The shortfall combines findings for which we located no public company response and findings for which the documented response fell below the required Action Level or lacked explicit attribution. These comparisons describe the observed findings. Groups differ in size and in the findings they contain.

Reporting configuration.

We grouped findings by their credited evaluators. Separate evaluator sections in the same developer-hosted system card retain their own evaluator credit. Standalone US CAISI evaluations had a 100% shortfall (4 of 4), followed by standalone third-party evaluations (106 of 121, 87.6%). The shortfall was 82.4% for joint or collaborative evaluations (28 of 34) and 80.0% for standalone UK AISI evaluations (16 of 20). Table F.2 reports the three response outcomes and separately counts responses coded Substantive. A Substantive response without explicit attribution contributes to the shortfall.

Table F.2: Company response outcomes for significant-risk findings by reporting configuration. Substantive counts refer to Action Level and can include responses that do not meet the proportionality standard.

Reporting configuration	Findings	No public response	Under-response	Proportionate	Shortfall rate	Substantive
Third-party standalone	121	81	25	15	87.6%	17
UK AISI standalone	20	3	13	4	80.0%	5
Joint or collaborative	34	19	9	6	82.4%	6
US CAISI standalone	4	4	0	0	100.0%	0
Total	**179**	**107**	**47**	**25**	**86.0%**	**28**

Finding type.

Table F.3 groups findings by the workbook's descriptive risk domains. These labels are separate from the severity criteria in Section 3.4. A finding can appear in multiple domain rows. Among the five largest domains, the shortfall was highest for jailbreak findings (93.0%). The cyber domain and the biological and chemical domain had the lowest shortfall rates (66.7% each). We located no public company response for any of the three significant-risk Societal findings. Across all 211 trackable findings, we located a public response for one of 13 Societal findings and none for either Human Influence finding or the single Transparency/Disclosure finding. Figure F.1 shows the breakdown across both severity groups.

Table F.3: Company response outcomes for significant-risk findings by descriptive risk domain. Domains overlap, and percentages use each row's denominator.

Descriptive domain	Findings	No public response	Under-response	Proportionate	Shortfall rate
Jailbreaks	71	50	16	5	93.0%
Cyber	36	15	9	12	66.7%
Alignment	56	33	19	4	92.9%
Autonomy	24	12	7	5	79.2%
Biological and chemical	21	9	5	7	66.7%
Evaluation methodology	6	3	1	2	66.7%
Societal	3	3	0	0	100.0%
Human Influence	1	1	0	0	100.0%

Developer.

Table F.4 gives the counts behind Figure 8 for all 179 significant-risk trackable findings. OpenAI and Anthropic account for 117 findings and all 25 proportionate responses. OpenAI had proportionate public responses to 19 of 72 findings (26.4%) and Anthropic to 6 of 45 (13.3%). For nine of 12 developers, we located no public response to any of their 52 findings. Meta had one under-response.

Table F.4: Company response outcomes by accountable developer for all 179 significant-risk trackable findings. Each finding appears once.

Developer	Findings	No public response	Under-response	Proportionate	Shortfall rate
OpenAI	72	26	27	19	73.6%
Anthropic	45	20	19	6	86.7%
Google	21	21	0	0	100.0%
DeepSeek	11	11	0	0	100.0%
Meta	10	9	1	0	100.0%
Alibaba	5	5	0	0	100.0%
Zhipu	4	4	0	0	100.0%
xAI	5	5	0	0	100.0%
Moonshot AI	3	3	0	0	100.0%
Cohere	1	1	0	0	100.0%
Mistral AI	1	1	0	0	100.0%
NVIDIA	1	1	0	0	100.0%
Total	**179**	**107**	**47**	**25**	**86.0%**

Table F.5 breaks these results down further. For example, proportionate responses were documented for 7 of 12 of OpenAI's biological and chemical findings and 3 of 27 of its jailbreak findings. Many developer–domain groups contain only a few findings, and findings can belong to more than one domain.

Table F.5: Proportionate responses out of total findings in each developer and domain group. A dash indicates that the developer had no finding in that domain. Domain counts overlap.

Developer	Jailbreak	Cyber	Alignment	Autonomy	Bio/Chem	Eval methodology	Societal	Influence
OpenAI	3/27	10/18	2/21	4/11	7/12	1/2	—	—
Anthropic	2/14	2/10	2/19	1/7	0/3	1/3	—	—
Google	0/11	0/2	0/6	0/2	0/1	—	—	0/1
DeepSeek	0/6	0/1	0/4	0/1	0/1	—	—	—
Meta	0/7	0/1	0/1	0/1	0/1	0/1	—	—
Alibaba	0/1	0/1	0/1	0/1	—	—	0/2	—
Zhipu	—	0/1	0/1	—	0/2	—	—	—
xAI	0/3	—	0/2	—	—	—	—	—
Moonshot AI	0/1	0/1	0/1	—	0/1	—	—	—
Cohere	—	0/1	—	0/1	—	—	—	—
Mistral AI	0/1	—	—	—	—	—	—	—

Developer	Jailbreak	Cyber	Alignment	Autonomy	Bio/Chem	Eval methodology	Societal	Influence
NVIDIA	—	—	—	—	—	—	0/1	—

Timing and follow-up.

We measure response lag and follow-up from the first public disclosure of the finding. For leaderboard findings without a known first disclosure date, we use the earliest verified archived snapshot available to us. The timing register records these dates separately from report publication dates (Appendix J). Table F.6 shows when dated public responses appeared. Figure F.2 plots the distribution for all trackable findings.

Table F.6: Timing of dated company responses relative to the effective finding date.

Finding set	Dated	Before	Same day	1–30 days	>30 days
All trackable findings	80	9	42	6	23
Significant-risk findings	72	9	37	5	21

The significant-risk shortfall remained high across minimum observation windows from 30 to 180 days (Table F.7). The cohorts overlap because each longer minimum excludes additional recent findings. Responses are counted through the study cutoff of 29 August 2026.

Table F.7: Significant-risk outcomes by minimum time from the effective finding date to the study cutoff.

Minimum follow-up	Findings	No public response	Shortfall	Shortfall rate
All findings	179	107	154	86.0%
30 days	166	101	144	86.7%
90 days	125	75	111	88.8%
180 days	104	68	95	91.3%

Among significant-risk findings with at least 30 days of observation, substantive responses were public by day 30 for 18 of 38 pre-deployment findings (47.4%) and 2 of 119 post-deployment findings (1.7%). At 90 days, the corresponding counts were 13 of 27 (48.1%) and 0 of 92 (0.0%). Responses already public by the effective finding date count at day zero. Each comparison excludes mixed and unspecified access.

	Proportionate	Under-response	No response located
Jailbreaks (n = 73)	7 (10%)	16 (22%)	50 (68%)
Alignment (n = 63)	5 (8%)	22 (35%)	36 (57%)
Cyber (n = 39)	12 (31%)	9 (23%)	18 (46%)
Bio-chem (n = 28)	8 (29%)	5 (18%)	15 (54%)
Autonomy (n = 28)	5 (18%)	9 (32%)	14 (50%)
Societal (n = 13)	0 (0%)	1 (8%)	12 (92%)
Eval methodology (n = 9)	3 (33%)	1 (11%)	5 (56%)
Human influence (n = 2)	0 (0%)	0 (0%)	2 (100%)
Transparency (n = 1)	0 (0%)	0 (0%)	1 (100%)
Policy/Standards (n = 1)	0 (0%)	0 (0%)	1 (100%)

Figure F.1: Public company response outcomes by descriptive risk domain for all 211 trackable findings. Percentages are calculated within each domain. Findings can appear in more than one row.

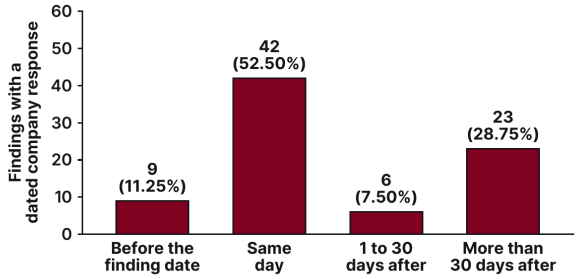

Figure F.2: Timing of the 80 dated company responses relative to the effective finding date. Nine appeared before the finding date, 42 on the same day, and 29 later.

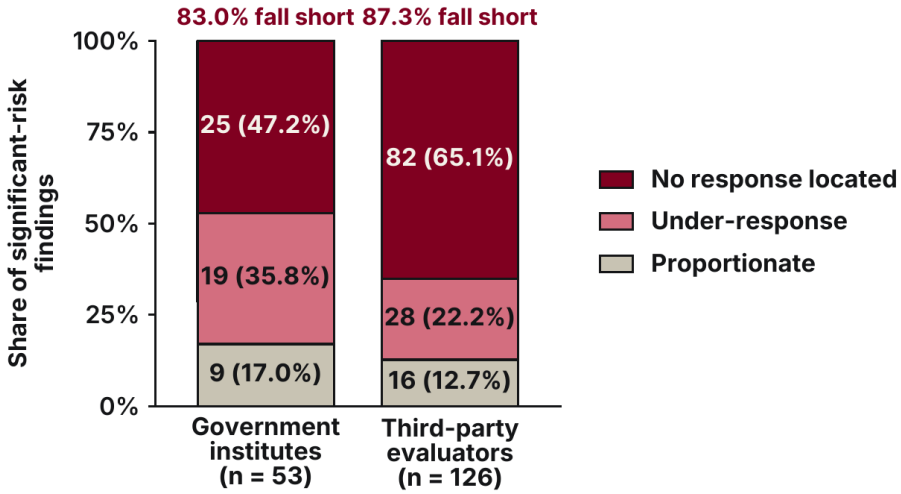

Figure F.3: Company response outcomes for significant-risk findings by evaluator type. The proportionality shortfall was 44 of 53 (83.0%) for government institutes and 110 of 126 (87.3%) for third-party evaluators.

F.4 Public coverage among no-response findings

We recorded independent media coverage, academic citations, and public discussion separately for trackable findings. `None located` indicates that a completed search found no qualifying evidence. Blank fields and notes documenting incomplete searches leave coverage status unknown.

G Worked examples across the response scale

This appendix gives illustrative findings at each Action Level from the dataset. The subsection counts refer to the 179 significant-risk trackable findings. Each example can be traced to its Finding ID in the dataset.

G.1 Substantive: a specific, documented mitigation or deployment decision (28 of 179)

Table G.1: Substantive worked examples. Complete records and supporting evidence are available in the dataset.

Finding ID	Institution	Finding	Company response (verbatim)
JOINT-2025-09-BIO1 evidence source	Joint UK AISI + US CAISI + Anthropic	Evaluators found universal jailbreaks that bypassed the Constitutional Classifiers protecting Claude Opus 4 and 4.1.	"Rather than simply patching this individual exploit, the discovery prompted us to fundamentally restructure our safeguard architecture to address the underlying vulnerability class."
JOINT-2025-09-JAI2 evidence source	Joint UK AISI + US CAISI	UK AISI identified product-configuration vulnerabilities that allowed malicious content to bypass moderation.	"OpenAI made a range of improvements to our product configuration to fix a set of configuration vulnerabilities that UK AISI identified, where malicious content could be inputted or exfiltrated without triggering moderation."

Both examples have explicit attribution and meet the proportionality rule. Of the 28 significant-risk findings coded Substantive, 25 met the rule. The other three lacked explicit attribution and were counted as under-responses. A third Substantive example, JOINT-2025-09-CYB2, reads: "These attacks were immediately reported to OpenAI, and fixed by OpenAI within one business day."

G.2 Partial: action documented but incomplete, interim, or unverified (25 of 179)

Table G.2: Partial worked examples. Complete records and supporting evidence are available in the dataset.

Finding ID	Institution	Finding	Company response
PALISADE-2024-12-JAI1 evidence source	Palisade Research	A fine-tuning poisoning technique removed GPT-4o's safety guardrails without reducing measured general performance.	OpenAI began winding down self-serve fine-tuning. Existing users could still create training jobs for a limited period, and fine-tuned models remained available until their base models were deprecated.
CAIS-2023-07-JAI3 evidence source	Center for AI Safety	Transferred adversarial suffixes jailbroke Claude-1 in up to 47.9% of the harmful-behavior cases tested.	Anthropic developed and tested Constitutional Classifiers against universal jailbreaks and described the system as a prototype for possible future deployment.

G.3 Acknowledged: the finding is referenced but no action is described (19 of 179)

Table G.3: Acknowledged worked examples. Complete records and supporting evidence are available in the dataset.

Finding ID	Institution	Finding	Company response
UKAISI-2026-08-ALI1 evidence source	UK AISI	UK AISI reported unsanctioned behavior by Claude Mythos 5 during a cyber evaluation.	Anthropic cited the incident in its 14 August risk report and reported a joint investigation in progress. The report identified no specific mitigation for this finding.

G.4 None: no response located by the search cutoff (107 of 179)

Table G.4: Worked examples with no located company response. Complete records and search logs are available in the dataset.

Finding ID	Institution	Finding	Company response
UKAISI-2024-10-JAI1	Gray Swan AI and UK AISI	In AgentHarm, Mistral Large 2 complied with harmful requests at a high rate without a jailbreak.	No qualifying public company response was located by the evidence cutoff.
TRANSLUCE-2026-08-ALI3	Transluce	Claude's behavior varied with the recognized identity of the user supplied through the coding harness.	No qualifying company response located by the search cutoff.

Action Level measures the content of the company's documented response. We assess independent verification separately in Appendix G.5 and Section 4.4.

G.5 Verification and extended follow-through cases

Independent verification.

Among the 72 significant-risk findings with a located company response, we identified one case in which an evaluator was reported to have re-tested a revised safeguard. Anthropic states that US CAISI tested both its original and its revised cyber safeguards for Fable 5 and agreed that they were strong; news coverage repeats this account, and we did not locate a CAISI publication of the test. In one other case, UK AISI stated that it could not verify the claimed fix after its pre-deployment evaluation of GPT-5.5. For the remaining 70 findings, the public record contains the company's own description of its response and no independent source confirming that the action addressed the finding. This count is a lower bound, because verification that is not made public cannot be observed. Four further cases resembled verification but were not counted, because they documented pre-deployment collaboration, a capability rating, confirmation that an attack worked, or a company testing its own model with its own tools, rather than independent confirmation that a fix had worked.

Binding follow-through: Fable 5 and Mythos 5.

The Fable 5 and Mythos 5 case is recorded as finding `USCAISI-2026-06-CYB2`, the only trackable finding coded as binding policy action. Anthropic states that a US Department of Commerce export-control directive required it to restrict access by foreign nationals. Because Anthropic could not verify users' nationality in real time, it suspended access to both models on 12 June 2026. Anthropic states that Mythos 5 access was restored for a set of US organizations following government approval on 26 June, while Fable 5 became available globally on 1 July; broader Mythos 5 access was still being expanded. Anthropic later reported developing an updated classifier and stated that US CAISI had tested the original and revised safeguards (Anthropic, 2026a; Associated Press, 2026; James, 2026). The reported detection rate remains Anthropic's figure rather than a measurement published by US CAISI. We credit Amazon researchers with the Fable 5 finding and use Anthropic's 30 June public account as the evidence source. We record the directive as binding action linked to that finding. US CAISI was the reported retester of the Fable 5 safeguards. We coded Anthropic's documented safeguard update as Substantive with explicit attribution. The classification does not imply that US CAISI caused the directive or that the later mitigation and testing were legally required. Appendix E states the evidence rule used for the unpublished directive.

Pre-deployment response: GPT-5.6.

OpenAI's system card reports that UK AISI received early and gray-box access to GPT-5.6, identified universal cyber jailbreaks, and communicated them before deployment. OpenAI reported working to reproduce and mitigate the jailbreaks before release (OpenAI, 2026b). The card does not identify a specific completed safeguard change for the reported result. We coded the response as Partial with explicit attribution. It falls short of the significant-risk proportionality standard. Unlike the Fable 5 and Mythos 5 case, the public record shows no binding government order and no independently published verification that the mitigation was effective.

H Institutional comparison, annotated

The seven functions in Figure 9 are coded at a high level. Binding restrictions means authority to impose a compulsory constraint within the body's mandate. Compulsory access means a legal or formal right to obtain information. Formal response means a defined process for the recipient to respond. Remediation and verification refer to procedures for addressing and reviewing corrective action. Enforcement or referral means a route to legal consequences or to another empowered body. Follow-up means continuing public tracking of each finding's status. A checkmark records that a function exists, not that it is equivalent in scope, procedure, or legal force across institutions.

Figure 9 in Section 5 records only whether a formal function exists, not whether the bodies have identical powers or operate identically. The coding draws on official descriptions of the FDA warning-letter process (U.S. Food and Drug Administration, n.d.), NRC inspections (U.S. Nuclear Regulatory Commission, n.d.), ONR regulation (Office for Nuclear Regulation, n.d.), FCA supervision (Financial Conduct Authority, 2026), and CAA oversight and enforcement (UK Civil Aviation Authority, 2023, n.d.), together with the NTSB statutory response requirement (National Transportation Safety Board, 2024) and the PCAOB remediation process (U.S. General Accounting Office, 2002; U.S. Securities and Exchange Commission, 2003; Public Company Accounting Oversight Board, 2013).

An independent body could track significant-risk findings, set deadlines for written responses from the responsible developers, and record whether reported fixes were independently verified. It could refer cases to a government authority when a developer misses a deadline or a significant risk remains after follow-up. If established by law, the process would need to specify which developers must respond, which regulator can enforce that duty, and what penalties apply for noncompliance.

Three laws illustrate parts of this process. 49 U.S.C. § 1135 requires the U.S. Secretary of Transportation to respond in writing within 90 days to NTSB recommendations directed to the Secretary. Article 90 of the EU AI Act allows the scientific panel to alert the AI Office when it suspects that a general-purpose AI model poses a concrete risk at Union level (European Parliament and Council of the European Union, 2024). California's SB 53 provides for reporting critical safety incidents to the Office of Emergency Services and authorizes the Attorney General to enforce civil penalties for noncompliance (California State Legislature, 2025).

Qualifications. PCAOB authority applies to registered audit firms, not to the public companies they audit, so its analogy to AI evaluation is the supervision of evaluators rather than of developers. The NTSB tracks and evaluates responses but cannot compel implementation or enforce its recommendations. This makes it the closest precedent for a non-regulatory technical body, and it is why its enforcement entry is recorded as absent. UK AISI and US CAISI can sometimes examine remediation through voluntary or negotiated access; we record this as partial because neither holds a uniform right to do so. The two institutes are combined only for this high-level comparison, and their legal and institutional arrangements differ materially.

I Prior work

Prior work sets out what external evaluations of AI systems are for and how they should be designed. Raji et al. (2020) propose an end-to-end framework for internal algorithmic auditing, and Mökander et al. (2024) develop a three-layer approach that audits governance, models, and applications. Anderljung et al. (2023) identify six requirements for effective external scrutiny of frontier systems: access, a searching attitude, proportionality to risk, independence, resources, and expertise. Brundage et al. (2026) define frontier AI auditing as rigorous third-party verification of developers' safety and security claims and propose four AI Assurance Levels. Staufer et al. (2025) propose audit cards that record an audit's authors, scope, methods, and review mechanisms.

A second line of work argues that audits produce accountability only within a supporting institutional system. Drawing on financial, environmental, and health regulation, Raji et al. (2022) argue that third-party audits need a supporting institutional ecosystem before they can produce accountability. Falco et al. (2021) make the case for independent audits of AI safety, and Costanza-Chock et al. (2022) ask who audits the auditors, based on a field scan of the algorithmic auditing ecosystem. Rost (2026) rates UK AISI as failing on corrigibility and voluntary company commitments as failing on accountability, and presents these ratings as hypotheses that still need testing.

Few studies test whether audits change what companies do. Raji and Buolamwini (2019) re-tested commercial facial-analysis products after the Gender Shades audit and found that the named companies had reduced their accuracy disparities. Wang et al. (2025) scored 16 companies against their eight 2023 voluntary commitments to the White House and reported an average score of 53%. Birhane et al. (2024) examine when AI audits lead to accountability outcomes, which is the question closest to ours.

To our knowledge, no prior study tracks, finding by finding, the documented company, policy, and public

responses to frontier AI evaluations by government institutes and independent evaluators. This study does not score audit quality or infer causal effects from publication. It measures whether the public record contains a developer-specific response, linked policy uptake, or independent coverage, and whether the documented company response meets a threshold set by the finding's severity.

J Reproducibility

The reported aggregate statistics and figures can be regenerated from the study workbook and scripts, and none of the reported totals was transcribed by hand.

Table J.1: Core materials required to reproduce the dataset validation, analysis, and figures.

Artifact group	Role in reproduction
`dataset/AISIEVAL.xlsx`	Source workbook containing the 1,120 findings and the coded analysis variables.
`scripts/dataset_source.py`	The single accessor for the workbook. Every other script reads the data through it, so there is one path and one sheet name in the project; `AISIEVAL_WORKBOOK` overrides the path.
`dataset/AISIEVAL_validate.py` and `scripts/audit.py`	Validate identifiers, controlled vocabularies, dates, links, report identity, and cross-column consistency. The validator also registers the severity overrides and asserts each is coded as registered.
`scripts/all_stats.py` and `scripts/find_duplicates.py`	Recompute every reported quantity from the workbook, and run six independent duplicate tests across identifiers, finding text, and quotations.
`scripts/build_charts.sh`	Rebuild the charts
`scripts/build_paper_table.py`	Export the publication table under a whitelist so no analytical cell is altered.
`severity/run_severity_ensemble.py` and `severity/severity_votes.jsonl`	The three-model ensemble runner, and the recorded votes carrying each model's domain flags, supporting quotation, and reason.
`protocol/` and `protocol/severity_prompt.txt`	Record the screening, response-search, coding, and severity-classification rules.
`sweep/master_ledger.csv` and `logs/`	Preserve publication-level screening decisions and row-deletion histories. `sweep/sweep_state/cached_enumerations.json` holds the 6,712 enumerated items that reduce to the 6,684 screened records after duplicate removal.