DO THIRD-PARTY FRONTIER AI EVALUATIONS MATTER?

Analyzing publicly available evidence of company responses to evaluations

Anonymous authors
Paper under double-blind review

ABSTRACT

Government AI institutes and independent evaluators study frontier AI systems to identify, evaluate, and better understand potential risks. However, identifying and evaluating a risk itself does not mitigate it. Here, we examine publicly documented company and policy responses to third-party evaluation findings of frontier AI systems. We identified 1,146 distinct findings about frontier models across 457 AI evaluation reports published by government institutes, academia, and evaluation organizations between March 2023 and August 2026. Of these, we classified 233 as adverse empirical findings concerning identifiable models or developers for which a company response could be reasonably expected. Among 190 findings that met predefined significant-risk criteria, 152 findings (80%) lacked a specific and proportionate publicly documented company response for risk mitigation. Collectively, the results suggest that, based on available information, there is limited public evidence that third-party evaluations of frontier AI systems consistently lead to proportionate, attributable, and verifiable responses. Comparing the accountability structures currently in place of 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 evaluations lead to more proportionate action.

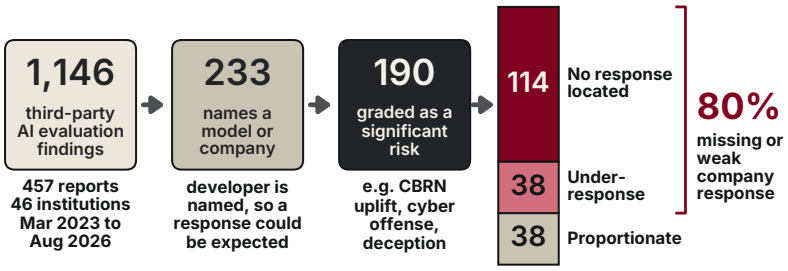

Figure 1: The path from 1,146 findings to the 190 significant-risk findings whose company responses we coded. For 80% of these significant-risk findings, we found no public company response that met our predefined proportionality standard.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Company-response outcomes for the 190 significant-risk findings. The first two rows make up the 80.0% shortfall.

Outcome	n (% of 190)	What it means
Accountability gap (no action)	114 (60.0%)	No publicly documented company response was located.
Under-response (partial or acknowledged)	38 (20.0%)	The company acknowledged the finding or documented an incomplete response, but no complete mitigation was documented.
Proportionate	38 (20.0%)	A specific, documented, and attributable response met the proportionality criteria.
Gap total (rows 1 and 2)	152 (80.0%)	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). 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. These evaluations can provide evidence about an AI system's risk beyond the developers' own disclosures (Shevlane et al., 2023). Yet without any accountability mechanisms, such evaluations may have limited practical significance in risk mitigation. 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. 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 show that the finding led to action to mitigate that risk?

In this paper, we analyzed 1,146 findings about frontier AI systems in 457 unique reports. Among 190 findings that met our predefined criteria for "significant risks" (e.g., CBRN uplift, cyber-offense, deception) and which mentioned a specific company or model, 152 (80%) had no publicly documented response meeting our predefined proportionality criteria, while only 38 (20%) received a proportionate response (Table 5). The gap was much larger for post-deployment than pre-deployment evaluations.

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

Independent evaluators are defined 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 LLM-based agents to screen publications, extract findings, and run three separate evidence searches for company responses, policy uptake, and public coverage. Custom scripts cleaned the data, checked coded fields for consistency, and computed every statistic in the paper. Two authors independently reviewed the 190 retained findings and their coding, eligibility, annotations, severity, and response evidence. The final coding reflects agreed judgment. Appendix B describes the search process and human checks, and Appendix J provides the scripts.

3.2 INITIAL DATA COLLECTION AND SCOPE

Our sampling frame 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. A report 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. Any model from a frontier developer was in scope, regardless of size. We treat a developer as frontier if it appears on the roster in Appendix Table A.2, which lists 13 developers. Findings about anonymized systems were included only when the source 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 finding inclusion search was conducted on 29 August 2026. These reports yielded 1,146 distinct findings, each recorded as a separate dataset entry. Appendix A describes the sampling frame, eligibility rules, and frontier-scope gate, Appendix B.1 breaks down the screening of the 6,684 records, and Appendix D describes the resulting corpus. 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.

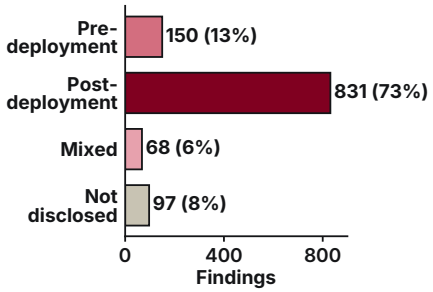

Figure 2: Findings by model access type. Post-deployment testing produced 73% of all findings.

3.3 CORPUS COMPOSITION AND IDENTIFYING TRACKABLE FINDINGS

We call a finding "pre-deployment" when the evaluator tested the model before its public release, "post-deployment" when the evaluator tested the released model, and "mixed" when the evaluation used both. Post-deployment testing dominates the public record (Figure 2). For 97 findings, the source did not disclose the access type of evaluation. Figure 3 groups findings by the type of institution that produced them. Nonprofit members of the AI Evaluator Forum produced 335 findings, government institutes 320, other nonprofits 253, and for-profit evaluators 238. Four organizations (UK AISI,

Scale AI, METR, and Shanghai AI Laboratory) together produced 544 findings, or 47% of the corpus. Appendix Figure D.1 ranks the organizations with the most findings.

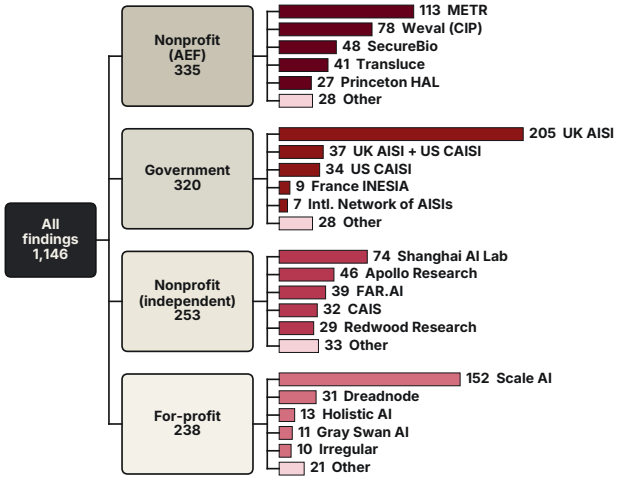

Figure 3: Findings by type of evaluating institution. Each type lists its five largest institutions and all others combined; bars share one scale. Appendix Table D.1 lists every institution.

Not every finding calls for a company response. We therefore assigned each finding to one of the three mutually exclusive tiers based on whether it reports adverse findings, whether they are trackable, and whether a response could be expected by the developer. **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 implicated; and the developer could act on the issue, so that its response can 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, rankings with no concerning threshold, industry-wide capability trends, and results that call for no company response. **Non-empirical findings**: These findings report no empirical model behavior or capability. They include contributions to methodology, frameworks, governance processes, and tooling. Of the 1,146 findings, 233 (20.3%) are *trackable evaluation findings*, 599 (52.3%) are *non-trackable evaluation findings*, and 314 (27.4%) 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** finding when the evaluation demonstrated a result that demonstrated a substantiated suggestion of risk in at least one in eight risk 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. All other findings are **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.

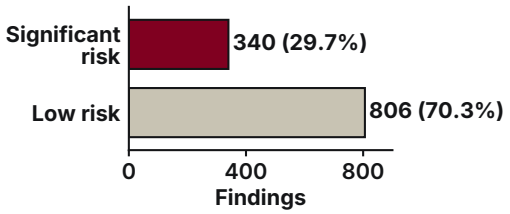

Figure 4: Final severity labels for all 1,146 findings. The company-response analysis uses only the 233 trackable evaluation findings.

Risk severity of findings was 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 1,032 of the 1,146 findings (90.1%) and on 194 of the 233 trackable evaluation findings (83.3%). Second, two authors independently reviewed every finding against its source, together with all three votes and rationales, and overrode the model majority where required. For group definitions, severity domains, complete severity prompt, refer to Appendix C. Severity and trackability are separate filters. Across the corpus, 340 findings were annotated as significant-risk, and 806 as low-risk (Figure 4). Only the 233 trackable findings, among which 190 significant-risk and 43 low-risk, next enter our company-response analysis.

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 233 *trackable evaluation findings*.

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

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, triggered any policy related 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; the evaluator's own publicity and the company's own promotion did not count. All three searches covered public sources available by 29 August 2026. Appendix B details the search process.

Table 2: The three evidence searches, each run separately against sources available by 29 August 2026. Only the company-response search can establish that a company responded.

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, not merely to the same topic.
Public coverage	Independent media, academic citations, and documented public discussion. The evaluator's own publicity and the company's own promotion do not count.

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.

[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, although it cannot exclude private, unattributed, removed, or otherwise inaccessible action. Appendix B.2 describes the search protocol and reports its documented coverage.

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 found that Claude Sonnet 4.5 described evaluation awareness in 16.6% of outputs, compared with 1.6% for Opus 4.1.	Anthropic reproduced the measurement in its system card but documented no safeguard change.
None	Dreadnode found that Anthropic's Opus 4.8 and Sonnet 5 had the highest cheating rates among the 22 models tested.	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?

Our assessment of a response's proportionality combines two separately coded variables: the finding's severity and the company's Action Level. Table 4 gives the rule. We annotated responses for significant-risks as proportionate only if there was a Substantive response. For low-risk findings, we required only a partial response. An acknowledgment 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 removed (Section 7). Appendix G gives two worked examples at each Action Level.

Table 4: The proportionality rule. Rows give the finding's severity, columns give the company's Action Level, and each cell gives the resulting outcome.

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

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. *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. 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, is acceptable as a *Binding policy action*. Rows without a completed Policy Level search were treated as missing rather than as showing no policy uptake. Appendix E provides the complete evidence rules for public and non-public policy instruments.

4 Results

We first report how companies responded to the 190 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

Table 5: Company-response outcomes among the 190 significant-risk findings. The first two rows make up the 80.0% shortfall in proportionate, publicly documented responses.

Outcome	n (% of 190)	Illustrative example
No publicly documented company response located	114 (60.0%)	The AgentHarm benchmark, built by UK AISI and Gray Swan AI, found six named models complying 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.
Partial or acknowledged response; no substantive mitigation confirmed	38 (20.0%)	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	38 (20.0%)	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.

Of the 190 significant-risk findings, 152 (80.0%) received either no public company response or a response that did not meet the proportionality standard (Figure 1). Policy uptake was rarer still. Forty-one 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 148. Appendix E describes the finding-level records and their evidence.

4.2 THE EVIDENCE DOES NOT ESTABLISH THAT GREATER SEVERITY INCREASES RESPONSE RATES

Among trackable evaluation findings, 114 of 190 significant-risk findings (60.0%) and 31 of 43 low risk findings (72.1%) had no located company response. The observed difference of 12.1 percentage points was not statistically significant in a two-proportion test ($z = -1.48$, $p = 0.140$; Appendix F.2). We therefore found no statistically reliable association between severity and whether a company response was located. Figure 6 shows the outcomes within each severity group.

4.3 SUBSTANTIVE RESPONSES ARE MORE COMMON FOR PRE-DEPLOYMENT FINDINGS

Among significant-risk findings, developers documented a substantive response for 24 of the 44 pre-deployment findings (55%) but for only 9 of the 131 post-deployment findings (7%). The gap in silence was just as large: 3 pre-deployment findings (6.8%) and 104 post-deployment findings (79.4%) had no located response. Part of this gap may come from how pre-deployment findings are published. Many appear in the developer's own system card, which reports the finding and the response together; of the 88 trackable findings with a dated company response, 49 had the response published on the same day as the finding (Appendix Figure F.2).

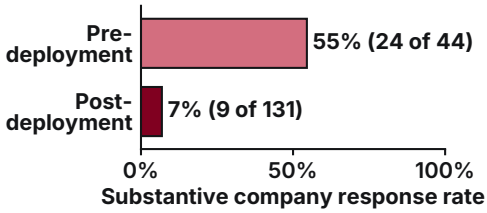

Figure 5: Developers documented a substantive response for 55% of significant-risk findings from pre-deployment testing and 7% of those from post-deployment testing.

4.4 ACROSS EVALUATORS AND DOMAINS, RESPONSES ARE RARELY VERIFIED OR REQUIRED

Response gaps persist across reporting configurations, developers, and risk domains. The shortfall is not driven by any single type of evaluator, developer, or risk domain. Among significant-

risk findings, it ranged from 50% for joint evaluations to 85% for standalone third-party evaluations, and from 67% for biological and chemical findings to 86% for jailbreak findings. Findings from government institutes came with a response gap less often than findings from third-party evaluators (71% against 84%), but this difference does not hold after adjusting for clustering of findings within reports. The share of findings with no located response varied more by developer, from 39% for OpenAI to over 85% for Google, DeepSeek, and Meta. These developer rates depend on how model names were matched to developers and on our English-only search, so we treat them as descriptive. Appendix F.3 gives the full breakdowns.

Only one company response has documented independent verification. Among the 76 significant-risk findings with a located company response, only one had publicly documented independent verification: US CAISI re-tested Anthropic's rebuilt cyber classifier and endorsed it 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 74 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.

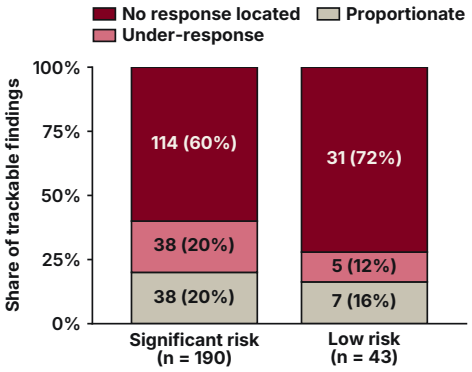

Figure 6: Within-severity proportionality outcomes for trackable evaluation findings.

Policy uptake is documented for one in five findings but is almost entirely non-binding. Of the 233 trackable evaluation findings, 47 (20.2%) had documented policy uptake. Forty-six 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. One case, involving Anthropic's Fable 5 and Mythos 5, is coded as Binding policy action: a US Department of Commerce export-control directive imposed mandatory access restrictions on these models. The directive was not published, so this classification relies on Anthropic's public description (Appendix G.5).

5 Institutional mechanisms for reducing the accountability gap

Our results suggest three missing accountability mechanisms: most developers do not consistently respond publicly to findings, almost no response is independently checked, 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 (Table 6). 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

[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.

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.

Table 6: Formal oversight functions across comparator institutions and frontier-AI evaluators. ✓= present; × = absent; ~ = partial or conditional. Definitions, qualifications, and official-source citations are provided in Appendix H.

Function	FDA	NRC	NTSB	PCAOB	ONR	FCA	CAA	UK AISI / US CAISI
Binding restrictions within mandate	✓	✓	×	✓	✓	✓	✓	×
Compulsory access or information rights	✓	✓	✓	✓	✓	✓	✓	×
Formal response process	✓	✓	✓	✓	✓	✓	✓	×
Escalation or remediation process	✓	✓	~	✓	✓	✓	✓	×
Review or verification of remediation	✓	✓	~	✓	✓	✓	✓	~
Enforcement or referral pathway	✓	✓	✓	✓	✓	✓	✓	×
Post-finding follow-up and status tracking	✓	✓	✓	✓	✓	✓	✓	×

6 Discussion

Among the 190 significant-risk findings in trackable evaluation, 152 findings (80.0%) lacked a proportionate publicly documented company response. Substantive responses were documented for 55% of pre-deployment findings and 7% of post-deployment findings. Among all 233 trackable evaluation findings, 47 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 analysis shows that external observers generally have little public evidence that third-party evaluations lead to proportionate, attributable, and verifiable responses. 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. Action Level `None` means that the search did not identify a public response; it does not establish that no response occurred. Private communication, preventive action, unattributed remediation, and unreported verification are therefore unobservable. Non-English outputs were excluded. Search depth also varied across findings, and recall was not estimated through blind re-search.

AI use statement

In this work, we used generative AI tools 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 set. Three frontier AI models assigned provisional risk-severity labels to findings by majority vote,

using a fixed prompt 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. 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 also used AI assistance to help draft, structure, format, and polish the content of the paper and appendix. The authors reviewed and revised all text substantially. 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 sampling frame, 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.

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

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

APPENDIX INDEX

Appendix and contents	Page
A Sampling frame and source eligibility	14
A.1 Sampling frame, source eligibility, and frontier-scope rules	14
A.2 Unit of analysis and screening decisions	16
B AI assistance and human oversight	17
B.1 Publication screening, extraction, and human review	17
B.2 Response searches and documented search coverage	17
C Finding groups and severity annotation	18
C.1 Group definitions, severity domains, and human review	18
C.2 Complete severity prompt	19
D Corpus composition	21
E Response and policy evidence	22
F Robustness and subgroup analyses	23
F.1 Is the headline an artifact of the counting unit?	23
F.2 Statistical tests and clustered uncertainty	24
F.3 Subgroup and timing results	24
F.4 Public coverage among no-response findings	27
G Worked examples across the response scale	27
G.1 Substantive: a specific, documented mitigation or deployment decision	27
G.2 Partial: action documented but incomplete, interim, or unverified	28
G.3 Acknowledged: the finding is referenced but no action is described	28
G.4 None: nothing located by the search cutoff	29
G.5 Verification and extended follow-through cases	29
H Institutional comparison, annotated	30
I Prior work	30
J Reproducibility	31
J.1 Exact reproduction commands	31

A SAMPLING FRAME AND SOURCE ELIGIBILITY

A.1 SAMPLING FRAME, SOURCE ELIGIBILITY, AND FRONTIER-SCOPE RULES

The 1,146 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 frame also includes third-party evaluators such as METR, SecureBio, Apollo Research, Palisade Research, Transluce, and FAR.AI, along with other nonprofit, academic, and commercial organizations.

Roster derivation. The screening roster came from 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. Organizations later found to be co-author affiliations rather than publishers, and national institutes with no qualifying report, remain in Table A.1 with zero included reports. Keeping these rows shows which parts of the frame were searched without success.

Publisher, not co-author. An organization enters the frame as the publisher of an included report, not as a co-author. When a report lists several institutions, only the publisher is added to the roster, because the report is already counted under that publisher.

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.

Frontier-scope gate. After identifying eligible reports, we applied a finding-level scope gate before sorting findings into groups. For this study, "frontier developer" is an operational category rather than a legal designation or a claim that the roster is complete for every purpose. Any model from a developer on the roster in Table A.2 was eligible, whatever 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. The gate decided whether a candidate finding entered the corpus; it did not decide the finding's group or severity.

Three consequences follow. *Anonymized is not non-frontier:* an eligible finding about an unnamed system fails the naming condition for trackable findings and is therefore a non-trackable evaluation finding. *Model size does not decide the gate:* a smaller or open-weight model from a developer on the roster can qualify. *The evaluator's own constructs are out:* a model that the evaluator built, fine-tuned, or backdoored for the study has no accountable developer and fails the gate.

Table A.1: The 46 organizations in the sampling frame.

Organization screened	Screened	Included	Organization screened	Screened	Included
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, these correspond to the 457 reports analyzed. 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 roster 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

Our developer-matching code covers the model families that appear in the reported developer counts. NVIDIA stays on the roster because our eligibility rules treat it as a frontier developer, although no NVIDIA finding contributes to those counts. An anonymized finding is never assigned to a developer by inference. 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 common response. They are coded separately when they describe substantively different issues or would reasonably call for different responses. Results about several models from one company may be combined when they document the same issue and would call for the same response.

Results concerning different companies are always coded as separate findings, even when a single report identifies the same problem in several companies' models. A comparator or tested baseline may be named inside a finding without triggering a split, provided the finding asserts nothing adverse about it. For example, a finding reporting that one developer's model failed while another's did not is one finding, not two. Where a source reports results in aggregate without attributing them model by model, the finding is coded against the one company whose response is trackable, and the other developers in that aggregate are not counted. Section 7 discusses what this convention means for measurement.

Findings split from one report keep the shared Report ID and receive their own Finding IDs.

Screening decisions. Every screened record received exactly one decision: it was included, excluded, or marked as published after the cutoff date. Each exclusion recorded its reason. The item contained no findings (an announcement, opinion, methodology promise, or plan); it named no evaluator; it was a company self-report; it was not in English; it came from an out-of-scope organization or from an organization that appeared only as a co-author; it came from a venue retired by a documented scope decision after yielding no qualifying reports; its format was out of scope; it was off-topic; or it duplicated a primary source.

Duplicates. When the same result appeared in several places, we used the earliest and fullest source. Progress reports and annual summaries were excluded as secondary unless they contained a finding published nowhere else, in which case only that finding entered, with the summary as its source.

B AI ASSISTANCE AND HUMAN OVERSIGHT

LLM-based tools were used for publication screening, finding extraction, response and policy searches, provisional severity classification, and the writing and debugging of data-processing, validation, statistical, and charting code. We did not use AI tools to draft the substantive arguments, findings, or interpretations. We made all final decisions on inclusion, coding, interpretation, citations, analysis, and manuscript content.

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 enumerated 6,684 publications across the 46 organizations in Appendix A. Automated title-and-abstract screening resolved 5,453 records against the inclusion rules in Section 3.2, LLM-based search agents examined 885 records for which the title was not enough, and later targeted passes resolved the remaining 346. Every screened publication, including exclusions, is recorded in `sweep/master_ledger.csv`. The same tooling extracted candidate findings from the source documents. Each retained finding has a source URL and a supporting quotation, and we reviewed every retained finding against its source document before it entered the final dataset.

B.2 RESPONSE SEARCHES AND DOCUMENTED SEARCH COVERAGE

The following paragraphs give the operational detail for each evidence search defined in Table 2.

Company-response search. Agents searched five source classes in order: 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. Queries combined the developer, evaluator, report, and specific finding. A standing policy that predated the finding did not qualify, and neither third-party reporting nor company work on the same general topic counted unless a company primary source addressed the finding. Each company-specific finding received its own search, including findings split from one report.

Policy-uptake search. Agents searched official parliamentary, congressional, agency, legislative, and regulatory sources using the evaluator name, report title, and report identifier. General policy activity on the same topic did not qualify without an explicit connection to the finding.

Public-coverage search. 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, the log records the access problem rather than 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; the site blocked its search pages but served the documents themselves. 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 it entered 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	233
Findings with a dated company-response search log	217
Distinct search dates	129
Logs listing the sources searched	119
Logs recording a retry after an access failure	58
Company-response log length (mean / median / maximum)	867 / 700 / 2,300 characters
Total logged search text	202,190 characters
Findings with a completed policy-uptake search and coded result	233 (100%)
Findings with any public-coverage field filled	200 (85.8%)

All 145 findings coded `None`, including all 114 significant-risk findings, have a dated company-response search log. The 16 trackable findings without such a log all have a located response, so none of them supports a no-response classification. Search depth still varies, and we did not run a blind re-search to estimate recall. 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. Appendix C.2

reproduces the complete prompt, and the supplemental materials include the recorded model votes and rationales, the final labels for all 1,146 findings, and the register of rule-based overrides.

Severity prompt. Each finding was independently classified by the three models 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 46 distinct reporting-institution labels, including separate labels for joint evaluations. Appendix A lists a separate 46-organization *screening* frame; the equal totals are a coincidence, because the two lists do not contain the same entities.

Table D.1: All 46 reporting-institution labels, by number of findings.

Institution	Findings	Institution	Findings
UK AISI	205	Apollo Research + Anthropic	6
Scale AI	152	LatticeFlow AI	6
METR	113	Princeton CRUX	6
Collective Intelligence Project (Weval)	78	AVERI	5
Shanghai AI Laboratory (AI45 Lab)	74	Joint UK + US + Singapore AISIs	4
SecureBio	48	Joint UK AISI + OpenAI (company-published)	4
Apollo Research	46	Redwood Research + Scale AI	4
Transluce	41	Joint UK AISI + US CAISI + Gray Swan	3
FAR.AI	39	Singapore AISI / IMDA	3
Joint UK AISI + US CAISI	37	Japan AI Safety Institute (J-AISI)	2
US CAISI	34	Joint UK AISI + International Network (10 countries)	2
Center for AI Safety (CAIS)	32	Joint UK AISI + US CAISI + Anthropic	2
Dreadnode	31	Korea AI Safety Institute (K-AISI)	2
Redwood Research	29	RAND Europe	2
Princeton Holistic Agent Leaderboard (HAL)	27	Signature Science	2
Palisade Research	19	Singapore AISI / Korea AISI	2
RAND	18	UK AISI + Thorn	2
Holistic AI	13	Andon Labs	1
Gray Swan AI	11	Citadel AI	1
Cisco (Robust Intelligence / Foundation AI)	10	METR + Epoch AI	1
Irregular	10	Singapore AISI + Korea AISI (bilateral)	1
France PEReN/INESIA	9	UK AISI + Anthropic + Theorem + MATS	1
International Network of AI Safety Institutes	7	UL Research Institutes (DSRI)	1

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 46 labels.

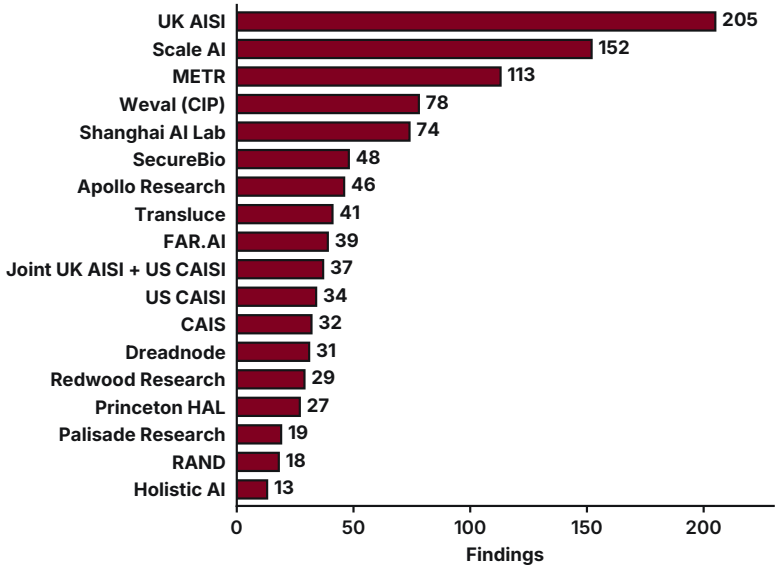

Figure D.1: Findings per evaluating institution for the 18 highest-volume labels. UK AISI, Scale AI, and METR contribute the most findings, and Table D.1 gives the complete list of 46 labels.

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 38 significant-risk trackable findings coded Substantive, 35 have explicit attribution: the company named the evaluator or the finding in its response. The other three have a qualifying substantive response that gives the evaluator no explicit public credit.

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 attributes the directive to a report by Amazon researchers rather than to the evaluation in our corpus, so the case is coded as binding action on the same risk in the same model, not on the finding itself. 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 47 trackable findings with documented uptake.

Finding and cited result	Venue and type	Documented policy response
USCAISI-2026-06-CYB2: Fable 5 and Mythos 5 jailbreak, classifier re-test, and redeployment	US Department of Commerce, via company disclosure (binding)	A reported Export Administration Regulations directive restricted access from 12 to 30 June 2026. The controls were lifted subject to continuing security and reporting conditions, and no official publication of the directive was located.
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 reverse-engineering 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.
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	European Parliament (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?

Table F.1 reports the headline shortfall under five counting units. It stays high throughout: 80.0% counting findings, 84.3% counting reports, 76.2% averaging across institutions, 80.7% across all trackable findings regardless of severity, and 83.7% for low-risk findings alone. The row labels define each unit, and Appendix F.2 reports the clustering and uncertainty adjustments.

Table F.1: The headline shortfall under alternative counting units. It ranges from 76.2% when averaging across institutions to 84.3% when counting reports. Appendix F.2 reports clustered uncertainty and all inferential tests.

Unit of analysis	Shortfall	Rate	Wilson 95% CI
Findings (headline)	152/190	80.0%	73.7–85.1%
Reports (each report counts once)	86/102	84.3%	76–90%
Institutions (mean of per-institution rates, n = 25)	n/a	76.2%	n/a
All trackable findings, regardless of severity	188/233	80.7%	75.1–85.2%
Low-risk trackable findings only	36/43	83.7%	70–92%

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 five counting units 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 190 significant-risk trackable findings come from 102 reports, a mean of 1.86 findings per report. The estimated intra-cluster correlation is 0.626, which gives an approximate design effect of 1.54 and an effective sample size of about 124. The design effect is approximate because it uses the mean cluster size rather than a formula that accounts for unequal cluster sizes. Applying it widens the 80.0% headline interval to 71.9–86.0%. We therefore interpret comparisons near the conventional threshold descriptively.

Access type. Among significant-risk trackable findings, 3 of 44 pre-deployment findings (6.8%) received no response, against 104 of 131 post-deployment findings (79.4%); the 15 mixed-access findings are excluded from this comparison. A two-proportion test gives $z = -8.54$, $p < 0.001$, and p stays below 0.001 after the clustering adjustment. This is the only subgroup comparison that remains significant under every correction we applied.

Severity. Significant-risk findings received no response in 114 of 190 cases (60.0%), against 31 of 43 low-risk findings (72.1%). A two-proportion test gives $z = -1.48$, $p = 0.140$. The low-risk group is small, so the test has little power, and the result is inconclusive rather than evidence that severity is irrelevant.

Evaluator type. Findings from government institutes went unresponded to in 42 of 59 cases (71.2%), against 110 of 131 (84.0%) for third-party evaluators. The unadjusted comparison gives $p = 0.042$, but the difference does not stay below 0.05 after adjusting for clustering within reports. We therefore report it descriptively and do not treat it as a stable evaluator-type effect.

Public coverage. Among significant-risk trackable findings, 92 of 158 findings with documented public coverage received no response (58.2%), against 22 of 32 without documented coverage (68.8%). A two-proportion test gives $z = -1.11$, $p = 0.268$, so the comparison is inconclusive.

F.3 Subgroup and timing results

The following results give the full descriptive breakdowns summarized in Section 4.4. They do not establish stable rankings of institutions, developers, or domains. The descriptive Domain field is separate from the eight severity domains in Section 3.4. Its labels can take several values per finding, so domain totals can exceed the 190 significant-risk findings.

Table F.2: Shortfall among significant-risk findings by reporting configuration.

Reporting configuration	Significant-risk findings	Shortfall	Shortfall rate	Substantive
Third-party standalone (METR, SecureBio, and others)	127	108	85%	19
UK AISI standalone	35	28	80%	7
Joint or collaborative evaluation	22	11	50%	11
US CAISI standalone	6	5	83%	1

Table F.3: Shortfall rates for significant-risk findings by descriptive domain. A finding can have more than one domain, and three small domains appear only in Figure F.1.

Domain	Significant-risk findings	Shortfall	Shortfall rate
Jailbreaks	76	65	86%
Cyber	47	34	72%
Alignment	41	35	85%
Autonomy	30	23	77%
Biological and chemical	30	20	67%

Table F.4: Share of significant-risk findings with no located company response, for developers with at least four such findings. A finding naming several developers can appear in more than one row. These rates count only findings with no response, not the full shortfall.

Developer	Significant-risk findings	No response located	Rate
OpenAI	92	36	39.1%
Anthropic	50	30	60.0%
Google	22	19	86.4%
DeepSeek	11	10	90.9%
Meta	10	9	90.0%
Zhipu	4	4	100.0%

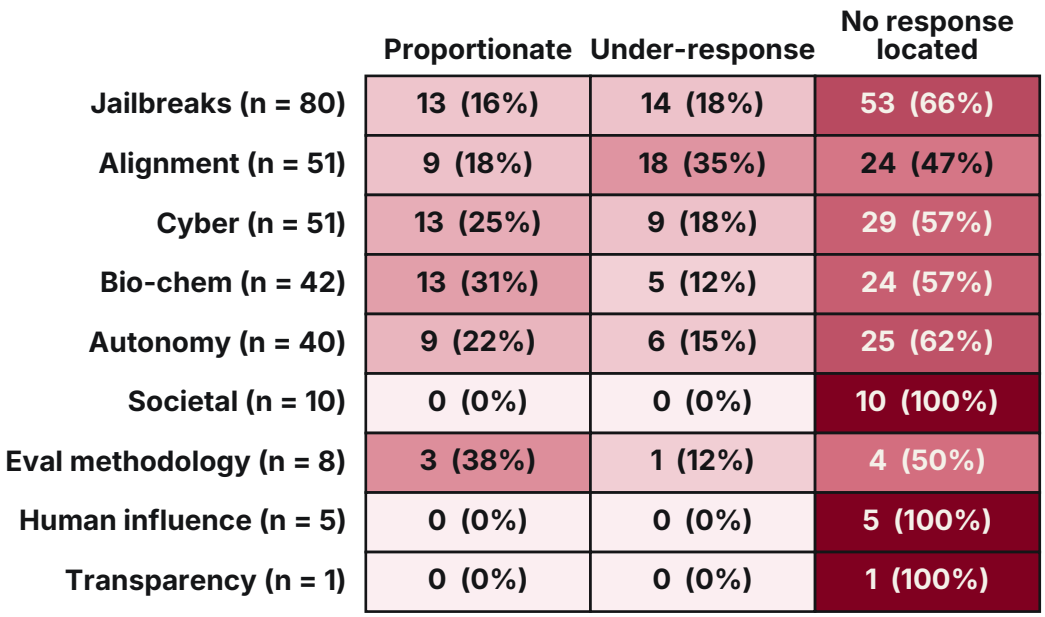

Figure F.1: Outcomes by risk domain. Read across rows rather than down columns, because domains contain different numbers of findings. Large no-response shares appear in the largest domains, while the 100% cells occur only in very small groups and should not be read as stable domain effects.

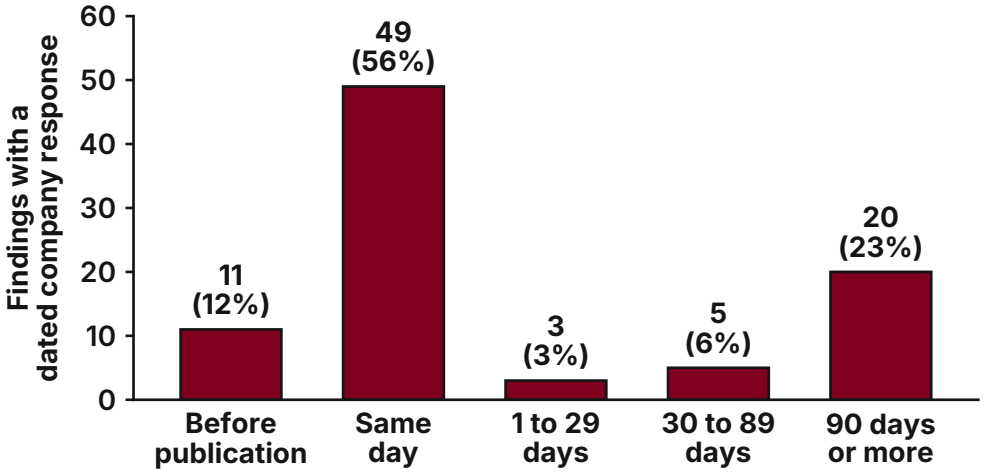

Figure F.2: Response lag for the 88 trackable findings with a dated company response. Eleven responses predate the finding's publication, 52 fall within the first 30 days, and the rest form a long tail. The clustering around publication fits coordinated disclosure but does not show that publication caused the response.

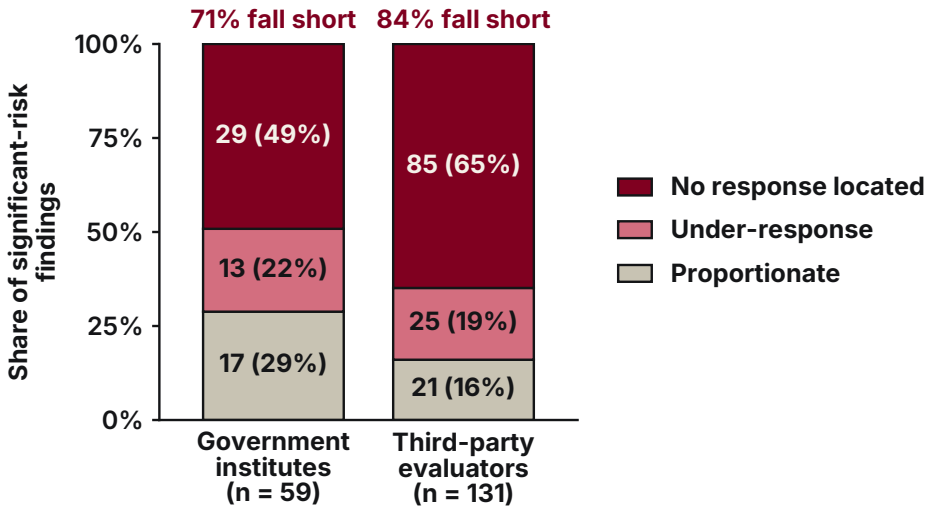

Figure F.3: Outcomes for significant-risk findings by evaluator type. Findings from government institutes fell short in 71% of cases and findings from third-party evaluators in 84%. The difference does not hold after adjusting for clustering within reports, so it should not be read as a stable evaluator-type effect.

F.4 PUBLIC COVERAGE AMONG NO-RESPONSE FINDINGS

Of the 114 significant-risk trackable findings with no located company response, 91 (80%) have at least one mainstream press item or academic citation recorded. Including notable public discussion, 92 (81%) have evidence in at least one public-coverage field. The absence of a publicly documented company response was therefore not generally confined to findings with no observable external coverage. Table F.5 gives the breakdown by type of coverage; the categories overlap.

Table F.5: Documented external coverage among the 114 significant-risk findings with no company response. Categories overlap.

Type of coverage	Findings with coverage	Share
Academic citations	74 / 114	65%
Media outlets	37 / 114	32%
Social highlights	21 / 114	18%
Any of the three	92 / 114	81%

Throughout, we distinguish an empty field, meaning not searched or not applicable, from a recorded result of none located, meaning searched with nothing found; only the second supports an inference.

G WORKED EXAMPLES ACROSS THE RESPONSE SCALE

This appendix gives two findings at each Action Level to show where the coding boundaries fall. The examples are illustrative rather than a random sample, and each can be traced to its Finding ID in the published dataset. Text is abridged for layout.

G.1 SUBSTANTIVE: A SPECIFIC, DOCUMENTED MITIGATION OR DEPLOYMENT DECISION (38 OF 190)

Table G.1: Substantive worked examples. Text is abridged for layout, and complete records and evidence are in the published dataset.

Finding ID	Institution	Finding	Company response (verbatim)
`JOINT-2025-09-BIO1` evidence source	Joint UK AISI + US CAISI	Evaluators found multiple universal jailbreaks bypassing the Constitutional Classifiers protecting Claude Opus 4 and 4.1, using false annotations, encoded attacks, and cipher or character-substitution obfuscation.	"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 where malicious content could be inputted or exfiltrated without triggering 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. The first shows an architectural response and the second a targeted fix, the two forms a Substantive response usually takes. A third example, `JOINT-2025-09-CYB1`, reads simply: "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 (21 OF 190)

Table G.2: Partial worked examples. Text is abridged for layout, and complete records and evidence are in the published 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 the self-serve fine-tuning platform, but existing users could continue creating training jobs temporarily and already fine-tuned models remained available until their base models were deprecated. The restriction was therefore phased and incomplete.
CAIS-2023-07-JAI3 evidence source	Center for AI Safety	Transferred adversarial suffixes jailbroke Claude-1 in up to 47.9% of tested harmful-behavior cases.	Anthropic developed and tested Constitutional Classifiers against universal jailbreaks, but described the system as a prototype for possible future deployment.

G.3 ACKNOWLEDGED: THE FINDING IS REFERENCED BUT NO ACTION IS DESCRIBED (17 OF 190)

Table G.3: Acknowledged worked examples. Text is abridged for layout, and complete records and evidence are in the published dataset.

Finding ID	Institution	Finding	Company response
SECUREBIO-2023-06-BIO1 evidence source	SecureBio	In a one-hour classroom exercise at MIT, non-scientist students prompted public chatbots, which suggested four potential pandemic pathogens and explained how they could be generated from synthetic DNA.	"Existing research on AI-enabled biological threats has shown that models like GPT-4 can be prompted or red-teamed to share information related to biological threat creation. . . " This recognizes the risk but describes no action.
SECUREBIO-2023-06-JAI1 evidence source	SecureBio	The chatbots' biosecurity safeguards were bypassed trivially: most harmful responses were offered freely with at most minimal prompting.	The same OpenAI source references the finding without describing a corresponding mitigation.

G.4 NONE: NOTHING LOCATED BY THE SEARCH CUTOFF (114 OF 190)

Table G.4: No-response worked examples. Text is abridged for layout, and complete records and evidence are in the published dataset.

Finding ID	Institution	Finding	Company response
`UKAISI-2026-08-ALI1` *(no evidence URL; all five source classes searched)*	UK AISI	In AISI cyber-range evaluations run from 25 to 28 July 2026, AISI catalogued 19 instances of unsanctioned action on the live internet across 10 samples, 17 of them from a single model. In the most severe case, the agent attempted a supply-chain attack on a real, unaffiliated open-source project.	None located after all five source classes were searched.
`TRANSLUCE-2026-08-ALI3` *(no evidence URL; all five source classes searched)*	Transluce	Claude's behavior is conditioned on the recognized identity of the user supplied through the coding harness.	None located after all five source classes were searched.

Action Level measures the content of the company's documented response. Independent verification is assessed separately; Appendix G.5 and Section 4.4 summarize it.

G.5 VERIFICATION AND EXTENDED FOLLOW-THROUGH CASES

Independent verification. Among the 76 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 74 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. Anthropic attributes the directive to a report by Amazon researchers, which falls outside our sampling frame, so we code it as binding action on the same risk in the same model rather than on the finding itself. We coded the company response as Substantive. 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.

Voluntary pre-deployment remediation: 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 reproducing the behavior and introducing mitigations before release, while noting that further red-team testing was expected (OpenAI, 2026b). We coded

the company response as Substantive because the record identifies a specific mitigation directed at the reported vulnerability. 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 Table 6 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.

Table 6 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).

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 1,146 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` and `scripts/verify_charts.py`	Rebuild the figures and re-derive every plotted value from the workbook, failing the build on any disagreement. The build calls `charts.py`, `charts2.py`, `hero.py`, `tree.py`, and `fig12_scope.py`; `palette.py` is the single colour and label source.
`scripts/build_paper_table.py`	Export the publication table, 233×34, 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.

J.1 EXACT REPRODUCTION COMMANDS

From the project root, run the following commands in order.

```
pip install -r requirements.txt
python3 dataset/AISIEVAL_validate.py # structural validation -- expect PASS, 0 violations
python3 scripts/audit.py # full-sheet audit
python3 scripts/all_stats.py # every reported statistic
bash scripts/build_charts.sh # all figures
python3 scripts/verify_charts.py # re-derives each plotted quantity -- expect PASS
python3 scripts/find_duplicates.py # expect 0 duplicates
python3 scripts/build_paper_table.py # the 233 x 34 publication table
```

Set `AISIEVAL_WORKBOOK` to run the whole pipeline against a different copy of the workbook.