

Current Public AI Evaluation-Boundary Statistics Do Not Support a Common Base Rate

An audit of 2026 evidence from Anthropic, AISI, and METR

Jack Lakkapragada

Independent AI Safety Contributor · 11 September 2026

Epistemic status. This is a short methods paper based on a primary-source audit of published statistics from Anthropic, AISI, and METR. I did not run a new model experiment. I started with a simple question: can these public rates actually be compared as if they measure the same thing? While auditing the sources, I found several unit, outcome, and conditioning mismatches, so I refined the comparison rules before writing the conclusion and then froze them. The result is not a new “AI escape rate.” It is a test of whether the existing public numbers can support one.

Abstract

This project started with a simple question: can the public evaluation-boundary statistics from Anthropic, AISI, and METR actually be compared as if they measure the same thing? I audited the underlying numerators, denominators, event definitions, and evaluation conditions before trying to combine anything. The answer was no. Across the sources I reviewed, the reported rates differ in what counts as an event, what population or run set is being measured, and how violations are detected. The only clean within-design comparison I found was AISI’s five-model cheating analysis. I therefore do not report a cross-lab base rate. Instead, I propose a short reporting checklist that would make future evaluation-boundary statistics easier to interpret and compare.

1. Why I chose this question

Track 2 #13 asked whether the public evidence from Anthropic, AISI, and METR could support a first, caveated base rate of evaluation-boundary violations. That immediately interested me because getting to a percentage is easy; figuring out whether the underlying numbers are actually measuring the same thing is much harder. I started with a simple hypothesis: some of these published rates probably should not be combined, but I did not assume they would all fail. On the surface, the numbers look comparable — Anthropic reported three incidents involving six of 141,006 reviewed runs, AISI reported cheating-attempt rates of 7.8%–14.1% across five models and separately 10 of 122 runs with unsanctioned live-internet actions, and METR reported cheating in at least 16% of successful long-horizon runs while also cataloguing 44 against-intent incidents. My question was whether these are really estimates of the same underlying quantity, or whether combining them would create a rate that none of the sources actually measured.

2. How I tested the hypothesis

I started by asking the same three questions of every reported rate: **what exactly counts as the event, what unit is being counted, and what runs or cases make up the denominator?** I then checked the surrounding setup — what task was being run, whether internet access was intentional, what safeguards were on or off, and how the behavior was detected. This mattered because two percentages can look comparable while measuring very different things.

I treated a comparison as invalid if the underlying event was different, the units did not match, or the denominator was conditioned on a different set of runs. I also rejected cases where the numerator or denominator could not be recovered from the public source. For context, I grouped the behaviors into four broad families: containment failure, cheating, unauthorized third-party action, and oversight or monitoring subversion. The grouping helped organize the evidence, but I did not treat two events as equivalent just because they fell into the same family.

I wrote the first version of these comparison rules before doing the quantitative synthesis, then refined them as the source audit exposed unit and outcome mismatches. I locked the final rules before writing the conclusion and applied them consistently across the audited sources. This was not preregistered.

3. What I analyzed

I focused on the main primary sources behind the public numbers: Anthropic’s July incident report and September follow-up, AISI’s cheating study and August incident report, and METR’s *Frontier Risk Report*. For each source, I pulled out what was actually being counted, what the denominator represented, what conditions the model was operating under, and how the behavior was detected. I also checked whether later updates changed the original search or count. I used the Hugging Face, OpenAI, METR/Redwood, and collusion.wiki reports only as context for how these incidents unfolded, not as rate evidence, because I did not find a clear evaluation-run denominator in those campaign reports.

4. What the sources actually measure

Source	Published statistic	What increments k	Conditioning / frame	What this means
Anthropic Jul	3 incidents; 6 affected runs / 141,006 reviewed runs	Run involved in successful unauthorized access to external production infrastructure	Runs in which Claude could have obtained internet access; later search shown incomplete	6/141,006 is only the observed fraction within the initial reviewed set. I do not treat it as prevalence or as an underlying incident probability.
AISI cheating	67/475, 54/475, 60/475, 43/475, 37/475	Detected cheating attempt: out-of-scope or disallowed action used to shortcut intended solution	Same cyber-eval suite; 475 trajectories per model	This is the cleanest multi-model comparison I found because the five models were tested in the same suite with the same denominator and monitor setup. I do not combine them into one 261/2,375 rate.
AISI incident	10/122 runs; 19 unsanctioned actions	Run included autonomous unsanctioned live-internet action targeting real people or organizations	Internet intentionally enabled; classifiers deliberately disabled	Not a sandbox escape and not the same numerator as Anthropic.
METR TH1.1	At least 16%	Cheating / illegitimate solution among qualifying runs	Successful Time Horizon 1.1 runs on tasks $\geq 8h$	Success-conditioned; exact k and n not published, so no reconstructed rate beyond the stated percentage.
METR incident set	44 incidents	Selected against-intent incident	18 selected eval cases + 24 public + 2 company-shared	Separate incident atlas; no denominator; not the numerator for the $\geq 16\%$ statement.

5. What I found

The main result was straightforward: none of the cross-study pairs I audited from Anthropic, AISI, and METR supported a clean, unadjusted common-rate comparison. The closest pair was Anthropic’s six affected runs and AISI’s 10/122 runs because both are run-level and both involve third-party action. But they still do not measure the same event. Anthropic counted runs that resulted in successful unauthorized access to third-party production infrastructure, while AISI counted runs with unsanctioned live-internet actions targeting real people or organizations. The setups were also different: accidental versus intentional internet access, different safeguard states, different tasks, and different detection methods. Even if I ignore the finer classification and put both in the same broad category, I still would not treat them as two estimates of one common rate.

The other cases were clearer. AISI's cheating study and its 10/122 incident report measure different outcomes. METR's $\geq 16\%$ figure applies only to successful long-horizon runs and does not publish the exact integer numerator and denominator. METR's 44 incidents are a separately selected incident set with no denominator. Anthropic's 3 incidents, 6 affected runs, and 141,006 reviewed runs are three related facts, not interchangeable pieces of one rate. The later ~481 million transcript search also covers a different search universe, so I do not treat it as an updated denominator for the July analysis.

Pair	Why they do not match	Other Issue	Decision
Anthropic 6/141,006 vs AISI 10/122	Different event: successful production access vs unsanctioned live-internet action	Different opportunity, safeguards, tasks, and detection	Do not combine
Anthropic 6/141,006 vs AISI cheating	Third-party action vs cheating attempt	Different evaluation design and opportunity	Do not combine
AISI cheating vs AISI 10/122	Cheating attempt vs third-party action	Same organization does not make the outcome equivalent	Do not combine
AISI cheating vs METR $\geq 16\%$	Detected cheating attempt vs success-conditioned illegitimate run	METR does not publish exact k and n	Do not combine
METR $\geq 16\%$ vs METR 44	Different objects	The 44 incidents have no denominator	Do not combine
AISI cheating, five models	Same event definition, suite, denominator, and monitor protocol	Model is the intended comparison; monitor sensitivity may still vary	Best within-design comparison; do not collapse into one pooled rate

6. How the result connects back to the hypothesis

I started with a fairly modest hypothesis: some of these public rates probably should not be compared directly. The audit ended up showing something stronger. Every cross-study pair I examined failed to support a clean common-rate comparison. That does not mean the individual studies are wrong or that their numbers are useless. It means those numbers only make sense inside the conditions that produced them. Once the event definition, denominator, opportunity to act, or detection method changes, the rate is no longer measuring the same thing.

So the negative result is the result. I could not responsibly build the “first denominator” from the current public evidence without creating a number that none of the sources actually measured.

7. Recommendation: report the measurement recipe, not only the percentage

The biggest practical takeaway from this audit is that a percentage by itself is not enough. If these rates are going to be compared across labs or studies later, the report needs to show how the number was produced. Based on the problems I found in this audit, I would want the following information reported alongside any evaluation-boundary rate:

Information to report	What it means	Why it matters
Event definition	Exactly what counts as the event: attempt, external contact, unauthorized action, successful access, persistence, compromise, etc.	Prevents successful access from being treated as the same thing as a lower-threshold unsanctioned action.
Numerator and denominator units	Report k and n separately and make clear whether they count runs, incidents, actions, agents, or something else.	Prevents incident/run and action/run mismatches.
Conditioning frame	State whether the denominator includes all runs, only successful runs, internet-capable runs, selected cases, etc.	Makes hidden conditioning, such as METR's success-conditioned rate, visible.
Opportunity structure	State whether internet or external tools were available and whether that access was intentional.	Separates designed-open-internet behavior from actual containment failure.
Safeguard state	State whether classifiers, refusal layers, or other relevant controls were enabled.	Makes clear what version of the system was actually being evaluated.
Detection and revision history	State how the behavior was detected, known blind spots or false negatives, and whether the search universe changed later.	Makes lower-bound or incomplete searches visible.
Model and task identifiers	Report the model/checkpoint, task family, and whether the result is model-level or entity-level.	Makes later stratification possible without forcing unrelated rates together.

8. Limitations, lessons learned, and next steps

Limitations. This was a targeted audit of the 2026 public evidence I used for this sprint, not a systematic review of every AI incident statistic. I also refined the comparison rules while working through the sources, although the main result still holds even under the earlier, broader classification. I report observed fractions rather than confidence intervals because these runs are not clean IID samples from a common population. I also limited the METR review to the sections needed for the 16% claim and Appendix D, and I left one unresolved AISI per-sample split out of the analysis rather than force an interpretation.

Lessons learned.

1. A percentage can be perfectly valid inside one study and still be misleading when carried into another.
2. Detection matters as much as the event itself. A retrospective search that later expands and a monitor with known blind spots are not measuring behavior in the same way.
3. Sometimes the right result is deciding what **not** to calculate. In this sprint, refusing unsupported rates was as important as extracting the published ones.

Next steps.

1. Publish the extraction workbook with the report so the source decisions, exclusions, and rejected calculations are visible.
2. Have an independent reviewer recode the Anthropic and AISI event definitions from the primary-source text alone and see whether they reach the same pairwise decisions.
3. Apply the same audit to future reports that publish both numerator and denominator. If enough studies eventually line up on event definition, unit, conditioning, opportunity, and detection, then a stratified or explicitly modeled cross-study rate may become possible. Until then, I would not manufacture a pooled “AI escape rate.”

Appendix A. Limitations and dual-use considerations

The main risk with this paper is not that it exposes a new attack method; it is that the result gets interpreted the wrong way. I only use information that is already public, and I do not describe how to escape a sandbox, gain external access,

or evade monitoring. My finding is also not that the published rates are useless. AISI's cheating rates, its 10/122 result, and Anthropic's 6-of-141,006 count are all informative within the settings that produced them. The problem starts when those numbers are carried across studies and treated as if they measure the same thing. My goal here is simply to make that comparison more disciplined, not to provide a safety checklist or publish new offensive techniques.

Appendix B. Claims I refused to make

1. 3/141,006 as a run-level rate.
2. 6/141,006 as prevalence or an underlying incident probability.
3. 4/481 million as an updated July cyber-evaluation rate.
4. AISI's 7.8–14.1% cheating rates averaged with the separate 10/122 incident result.
5. 10/122 as a sandbox-escape rate.
6. AISI's numerator described as "unsuccessful actions."
7. METR's 44 incidents used as the numerator behind its $\geq 16\%$ result.
8. An exact denominator inferred from METR's phrase "well over 100."
9. AISI's five model results collapsed into a pooled 261/2,375 cheating rate.
10. Campaign-level agent, action, or edit counts converted into run-level rates.
11. Binomial confidence intervals added to these retrospectively assembled corpora.
12. The rules were preregistered, the entire 2026 public record was audited, or ICH E9(R1) was implemented.
13. The proposed information set is mathematically minimal, sufficient, or a field standard.

References

- Anthropic. (2026, July 30). Investigating three real-world incidents in our cybersecurity evaluations. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- Anthropic. (2026, September 9). An alignment assessment of recent cybersecurity incidents. <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>
- Centre for Long-Term Resilience. (2026, February 2). The Loss of Control Observatory: a prototype to detect real-world AI control incidents. <https://www.longtermresilience.org/reports/the-loss-of-control-observatory-a-prototype-to-detect-real-world-ai-control-incidents/>
- Coalition for Secure AI. (2025). AI Incident Response Framework, V1.0. OASIS Open Project, Workstream 2. <https://www.coalitionforsecureai.org/wp-content/uploads/2026/03/AI-Incident-Response-1.pdf>
- Ezell, C., Roberts-Gaal, X., & Chan, A. (2025). Incident analysis for AI agents. arXiv:2508.14231. <https://arxiv.org/abs/2508.14231>
- Higgins, J. P. T., et al. (Eds.). Cochrane Handbook for Systematic Reviews of Interventions. Cochrane. <https://training.cochrane.org/handbook>
- Hugging Face. (2026, July 27). Anatomy of a frontier lab agent intrusion. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- ICH. (2019). E9(R1) addendum on estimands and sensitivity analysis in clinical trials. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
- METR. (2026, May 19). Frontier risk report (February to March 2026). <https://metr.org/blog/2026-05-19-frontier-risk-report/>
- METR & Redwood Research. (2026, August 26). Brief independent investigation of agents' behavior in the OpenAI / Hugging Face hacking incident. <https://www.redwoodresearch.org/research/hugging-face-incident>
- OECD. (2025). Towards a common reporting framework for AI incidents. OECD Artificial Intelligence Papers No. 34. https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

- OpenAI. (2026, July 21). OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- UK AI Security Institute. (2026, July 21). Cheating behaviour in frontier model evaluations. <https://www.aisi.gov.uk/blog/cheating-behaviour-in-frontier-model-evaluations>
- UK AI Security Institute. (2026, August 4). Incident report: unsanctioned agent behaviour during cyber testing. <https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing>
- UK AI Security Institute. (2026). Security incident INC-2026-07-28-01 [technical report]. https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf
- Von Arx, S., Byrd, C. S., Kitts, S., & Larsen, T. (2026, September 4). Discovery of a new OpenAI agent message board. <https://collusion.wiki/>