# From Claim to Check: Auditing the Verifiability of Public AI Disclosures

**Archie Chaudhury** [1]

## Abstract

Frontier AI labs frequently publish disclosures intended to validate the performance, safety, and reliability of their foundation models. These claims can shape governance, procurement, and external oversight because, absent rigorous independent validation, they are often one basis on which outside actors assess frontier AI systems. We ask whether the claims made in these disclosures can be validated using only the information released alongside them. As a proof of concept, we audit 61 claims from 16 public artifacts across 12 model creators using a custom six-field rubric for claim reproducibility. Only 21% are directly verifiable on our rubric; the dominant failure modes are ambiguous wording, incomplete protocol specification, and missing access to the relevant evaluation context. We provide a full analysis of the rubric and results, and propose a disclosure standard targeted at the failures we observe.

## 1. Introduction

As foundation models are deployed in increasingly high-stakes settings, ensuring that they are not only performant but also safe has become more urgent (Tabassi, 2023; Autio et al., 2024; Shevlane et al., 2023; Phuong et al., 2024). Public disclosures such as model cards, system cards, reports, and technical papers can shape governance, procurement, and external oversight (Bender & Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2021; Pushkarna et al., 2022; Anthropic, 2024; OpenAI, 2024; Google DeepMind, 2024). Yet these artifacts often omit the harnesses, prompt sets, or checkpoints needed to generate the reported results, making reproduction difficult for independent organizations and technical governance researchers seeking to inform policymakers (Liang et al., 2023; Casper et al., 2024; White et al., 2024; Liu et al., 2023). In this work, we propose an audit rubric for analyzing the reproducibility and verifiability of public AI disclosures.

Prior work addresses adjacent pieces of this problem. Documentation proposals call for structured reporting of model context, intended use, data, and limitations, while audit and transparency research shows that meaningful external scrutiny depends on operational detail, evidence access, and reporting quality (Bender & Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2021; Pushkarna et al., 2022; Raji et al., 2020; Bommasani et al., 2024; Casper et al., 2024; McCaslin et al., 2025; Staufer et al., 2025). We study a narrower question: **can the claims in public AI disclosures actually be reproduced?** We answer it with a claim-level audit of 61 claims from 16 public artifacts across 12 model creators. Only 21% of claims are directly verifiable as disclosed. Our contributions are a rubric for claim-level verifiability, an empirical proof-of-concept audit across multiple creators, and a minimum disclosure standard for more rigorous external oversight.

We do not assume that every evaluation detail should always be public. In some cases, reduced transparency may reflect information hazards, benchmark integrity concerns, security constraints, or other legitimate access limits. Our narrower concern is whether public disclosures clearly distinguish between claims that are independently checkable, claims that are assumption-dependent, and claims that rely primarily on internal-only evidence.

## 2. Methodology

Our method focuses on a narrow question: whether a public AI claim is independently verifiable *as disclosed*. We do not rerun provider evaluations. Instead, we ask whether the public artifact contains enough information for a competent external party to understand the claim, identify the evaluated system, and reproduce a close version of the reported test. This claim-centered framing follows work that treats model reporting, traceability, and disclosure quality as preconditions for accountability rather than optional communication layers (Mitchell et al., 2019; Arnold et al., 2019; Gebru et al., 2021; Raji et al., 2020; Tabassi, 2023; Autio et al., 2024; Winecoff & Bogen, 2025; Bhat et al., 2023).

[1]Independent. Correspondence to: Archie Chaudhury <archchaudhury02@gmail.com>.

*Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026*, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

*Table 1.* Field-wise rubric coverage across all 61 claims. These counts are not a sequential pipeline.

| Stage | Count | % |
|---|---|---|
| All claims | 61 | 100.0 |
| Measurable | 52 | 85.2 |
| Protocol sufficient | 37 | 60.7 |
| Access sufficient | 39 | 63.9 |
| Directly verifiable | 13 | 21.3 |

**Corpus.** We assembled a purposive corpus of 16 recent public artifacts from 12 model creators, spanning system cards, model cards, technical reports, release notes, and one government evaluation. The sample was chosen for variation in provider, artifact type, and access regime rather than for statistical representativeness. We included only public artifacts that make at least one concrete behavioral, safety, or capability claim about a deployed or released system. From these artifacts we extracted 61 claims, with between one and six claims per artifact, prioritizing statements likely to matter for governance, procurement, or external oversight. To make the heterogeneity of the corpus explicit, we also grouped claims into descriptive buckets covering architecture and deployment facts, capability claims, safety and robustness claims, and mitigation or governance claims.

**Claim coding.** Each claim was coded with a six-field rubric covering specificity of wording, disclosure of inputs, metric, judge or scoring procedure, inference setup, and external access to the relevant system or version. The six fields are coded independently rather than as a strictly sequential pipeline, and the final outcome label identifies the first decisive blocking failure. We then assigned one final label: *Directly verifiable*, *Verifiable with reasonable assumptions*, or one of four non-verifiable categories: *ambiguous wording*, *missing protocol inputs*, *missing metric or judge information*, and *missing access or version stability*. We define *measurable* claims as claims whose wording is precise enough that a third party could tell what would count as confirming or refuting them. We use *verifiable with reasonable assumptions* for claims that a serious external auditor could check only by filling in limited, low-impact gaps without substantially changing the evaluation target. Because this is a proof-of-concept audit, a single annotator graded the sample. Future work should use multiple annotators to increase confidence in both the rubric and individual scores. The appendix reports aggregate decision rules, corpus summaries, representative examples, and the full per-claim audit table used for the pilot audit.

## 3. Results and Discussion

Table 1 reports field-wise coverage rather than a sequential pipeline. While $52/61$ claims are at least measurable, only $37/61$ disclose enough about inputs and metrics for a plausible reconstruction, and only $13/61$ are directly verifiable. Many claims therefore appear empirical while still omitting the details needed for an external check.

*Table 2.* Final verifiability outcomes across all 61 claims.

| Label | Count | % |
|---|---|---|
| Direct | 13 | 21.3 |
| Assumption-burdened | 6 | 9.8 |
| Not verifiable: inputs/protocol | 9 | 14.8 |
| Not verifiable: metric/judge | 7 | 11.5 |
| Not verifiable: access/version | 11 | 18.0 |
| Not verifiable: ambiguity | 15 | 24.6 |
| **Verifiable subtotal** | **19** | **31.1** |
| **Non-verifiable subtotal** | **42** | **68.9** |

Table 2 gives the headline distribution: 13 claims (21.3%) are directly verifiable and another 6 (9.8%) are verifiable with reasonable assumptions; the remaining 42 (68.9%) are not independently verifiable from the public record. Public artifacts are often treated as evidence about model safety, robustness, and deployment readiness, yet in our corpus most such claims cannot be checked by a third party using the disclosed information alone.

The dominant failure mode is ambiguous claim wording. Fifteen of the 42 non-verifiable claims (35.7%) fail before protocol questions even arise. Typical examples include comparative or threshold-style statements such as Meta's claim that Llama 4 does not introduce catastrophic cyber risk, Anthropic's claim that Claude Sonnet 4.6 is its most capable Sonnet model, and OpenAI's claim that GPT-5.3-Codex is faster without a fully specified baseline (Meta, 2025; Anthropic, 2026; OpenAI, 2026a). Such claims may be informative as developer positioning, but they are weak inputs to external oversight because no outside actor can tell what test would confirm or refute them.

The remaining failures come from missing evidence or context rather than from wording alone. Nine claims fail because inputs or protocol are underspecified, seven because the metric or judging procedure is incomplete, and eleven because the relevant system, version, or supporting evidence is not realistically available to outsiders. High performance on a mathematics benchmark, for instance, can depend on evaluation choices such as multiple samples or best-of-$n$ selection, even though most users interact with models in a one-shot manner. Access constraints matter especially for system-level safety and agentic evaluations, where claims depend on internal tooling, private connectors, unpublished prompts, or deployment-specific mitigations.

The governance-relevant end of the corpus is especially hard to verify. Claims tied to public benchmarks and externally accessible releases, such as DeepSeek-R1's reported 79.8% pass@1 on AIME 2024, are much easier to audit than broad safety, capability, or policy-compliance statements

that depend on private evaluation infrastructure (DeepSeek-AI, 2025). Regime differences point in the same direction. Open-weight technical reports are 50.0% non-verifiable, while closed-API system cards are 81.2% non-verifiable. When claims are public-facing and governance-relevant, disclosures should state clearly whether they are independently checkable, assumption-dependent, or based on internal-only evidence. Recent corrections in xAI's Grok 4.1 documentation illustrate the cost of leaving that distinction implicit (xAI, 2025).

**Discussion and Disclosure Standard** These results motivate a simple disclosure standard for public capability claims. Any claim framed as a capability comparison or benchmark result should be tied to an independently accessible benchmark or explicitly labeled as an internal-only evaluation rather than presented as externally comparable evidence. Disclosures should also specify the scoring methodology and evaluation conditions in enough detail to permit replication, including the metric definition, pass criteria, judge configuration, and the exact model checkpoint, version, or release snapshot used to produce the reported number. We believe that such a disclosure standard would support more rigorous third-party validation of claims made by frontier labs. A fuller version of the standard appears in the appendix.

**LLM Usage Statement** We utilized LLMs to help with research, gathering examples of recent model releases, public disclosures and claims, assist with reviewing the corpus, and to help draft the paper. The idea itself and core writing were done by the author.

## References


Anthropic. Responsible Scaling Policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2024.

Anthropic. Claude Sonnet 4.6 System Card. https://www.anthropic.com/claude-sonnet-4-6-system-card, 2026. Published February 2026.

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., and Varshney, K. R. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13, 2019.

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., and Roberts, K. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. National Institute of Standards and Technology, NIST AI 600-1, https://doi.org/10.6028/NIST.AI.600-1, 2024.

Bender, E. M. and Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Bhat, A., Coursey, A., Hu, G., Li, S., Nahar, N., Zhou, S., K"astner, C., and Guo, J. L. C. Aspirations and Practice of Model Documentation: Moving the Needle with Nudging and Traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–17. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581518. URL https://doi.org/10.1145/3544548.3581518.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The Foundation Model Transparency Index. *arXiv preprint arXiv:2407.12929*, 2024.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948, 2025.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for Datasets. In *Communications of the ACM*, volume 64, pp. 86–92, 2021.

Google DeepMind. Frontier Safety Framework. https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/, 2024.

Liang, P., Bommasani, R., Lee, T., et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023.

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., et al. LLM360: Towards Fully Transparent Open-Source LLMs. https://arxiv.org/abs/2312.06550, 2023.

McCaslin, T., Alaga, J., Nedungadi, S., Donoughe, S., Reed, T., Painter, C., Bommasani, R., and Righetti, L. STREAM: A Standard for Transparently Reporting Evaluations in AI Model Reports. arXiv:2508.09853, https://doi.org/10.48550/arXiv.2508.09853, 2025.

Meta. Llama 4 Model Card. https://github.com/meta-llama/llama-models, 2025.

MiniMax. MiniMax-M1 Technical Report. https://arxiv.org/abs/2506.13585, 2025.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 220–229, 2019.

OpenAI. Preparedness Framework. https://openai.com/index/preparedness-framework/, 2024.

OpenAI. GPT-5.3-Codex System Card. https://openai.com/index/gpt-5-3-codex-system-card/, 2026a. Published February 2026.

OpenAI. GPT-5.4 Thinking System Card. https://openai.com/index/gpt-5-4-thinking-system-card/, 2026b. Published March 2026.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., et al. Evaluating Frontier Models for Dangerous Capabilities. https://arxiv.org/abs/2403.13793, 2024.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. https://arxiv.org/abs/2204.01075, 2022.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 33–44, 2020.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., et al. Model Evaluation for Extreme Risks. https://arxiv.org/abs/2305.15324, 2023.

Staufer, L., Yang, M., Reuel, A., and Casper, S. Audit Cards: Contextualizing AI Evaluations. arXiv:2504.13839, https://arxiv.org/abs/2504.13839, 2025.

Tabassi, E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, NIST AI 100-1, https://doi.org/10.6028/NIST.AI.100-1, 2023.

White, M., Haddad, I., Osborne, C., et al. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. https://arxiv.org/abs/2403.13784, 2024.

Winecoff, A. A. and Bogen, M. Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–18. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713814. URL https://doi.org/10.1145/3706598.3713814.

xAI. Grok 4.1 Model Card. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf, 2025.

# A. Appendix

## B. What the 61 claims are

The audited claims were extracted from 16 public artifacts spanning system cards, model cards, technical reports, release notes, and one government evaluation. Table 3 records that corpus at a glance.

*Table 3.* Corpus from which the 61 audited claims were extracted.

| Developer or evaluator | Artifact family | Public artifact(s) named in draft |
|---|---|---|
| OpenAI | System cards | GPT-5.4 Thinking; GPT-5.3-Codex |
| Anthropic | System cards | Sonnet 4.6; Haiku 4.5; Opus 4.6 |
| Google DeepMind | Model / system documentation | Gemini 3.1 Pro |
| Meta | Model card | Llama 4 |
| xAI | Model cards | Grok 4; Grok 4.1 |
| Mistral AI | Model card | Mistral Large 2 |
| DeepSeek | Technical report | DeepSeek-R1 |
| Z.AI | Technical report | GLM-4.5 |
| Qwen | Technical report | Qwen3 |
| MiniMax | Technical report | MiniMax-M1 |
| Cohere | Technical report | Command A |
| UK AI Security Institute | Government evaluation | Frontier Trends report |

At the developer level, the audited set contains OpenAI (16 claims), Anthropic (10), Google DeepMind (6), DeepSeek (4), Z.AI (4), Qwen (4), MiniMax (4), Cohere (3), Meta (3), xAI (3), UK AI Security Institute (3), and Mistral AI (1), summing to 61 claims overall. As noted in the main text, each artifact contributed between one and six claims, selected for governance relevance and concreteness rather than for statistical representativeness.

## C. Rubric used for claim coding

Each extracted claim was evaluated on six fields before a final outcome label was assigned. The rubric is designed to capture whether an external party could operationalize the claim *as disclosed*, not whether the claim itself is true.

*Table 4.* Six-field rubric used to audit claim verifiability.

| Field | Coding question |
|---|---|
| Specificity | Is the wording precise enough that a third party could tell what would count as confirming or refuting the claim? |
| Inputs | Are the prompts, tasks, datasets, or test conditions identified well enough to reconstruct the evaluation target? |
| Metric | Is the outcome measure, threshold, or scoring function stated clearly enough to reproduce the reported result? |
| Judge | If human or model-based grading is involved, is the judging procedure described well enough to replicate it? |
| Inference setup | Are the model version, sampling settings, tools, and surrounding runtime conditions specified sufficiently? |
| Access | Can an external party realistically access the relevant model, version, benchmark materials, or evidence trail? |

The final labels are interpreted as follows: *Direct* means the claim is concrete and all major reproduction conditions are sufficiently disclosed; *Assumption-burdened* means a serious external check is plausible but would require filling in one or more reasonable gaps; the four non-verifiable labels identify the first blocking failure as ambiguity, missing protocol inputs, missing metric or judge information, or missing access / version stability.

## D. Additional descriptive breakdowns

Appendix Figure 1 complements the main outcome table by showing where claims clear or fail individual fields of the rubric. Most claims are at least measurable, but many fail before a third party can reconstruct the relevant protocol. Appendix Table 5 then disaggregates the 42 non-verifiable claims by primary failure mode.

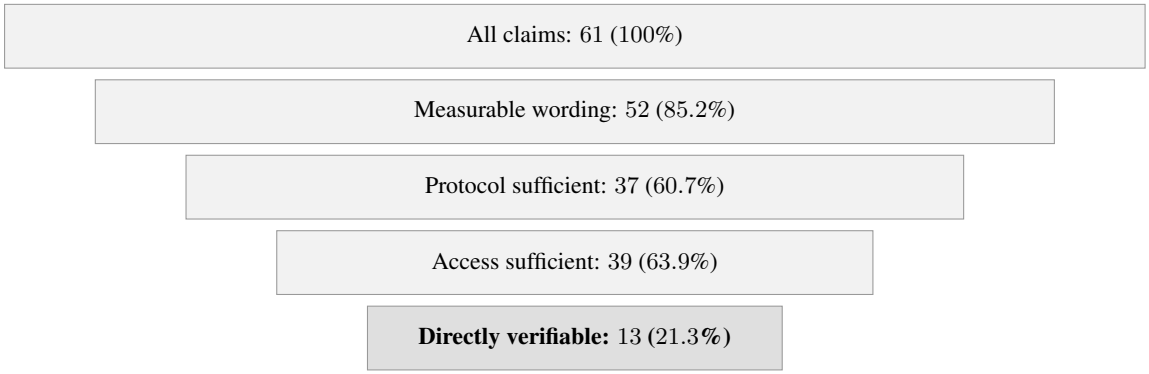


*Figure 1.* Field-wise rubric coverage for the full claim set. The counts are reported independently rather than as a sequential funnel.

*Table 5.* Primary failure modes among non-verifiable claims.

| Failure mode | Count | % of non-verifiable |
|---|---|---|
| Ambiguous claim wording | 15 | 35.7 |
| Underspecified inputs | 9 | 21.4 |
| Underspecified metric or threshold | 4 | 9.5 |
| Underspecified judge or grading | 3 | 7.1 |
| Missing access / unstable version | 5 | 11.9 |
| Internal-only evidence | 6 | 14.3 |

*Table 6.* Verifiability outcomes by developer regime. Percentages are within regime.

| Regime | $N$ | Direct | Assumpt. | Non-verif. | % non-verif. |
|---|---|---|---|---|---|
| Open-weight technical reports | 22 | 9 | 2 | 11 | 50.0 |
| Closed-API system cards | 32 | 2 | 4 | 26 | 81.2 |
| xAI late-transparency releases | 3 | 1 | 0 | 2 | 66.7 |
| Government evaluation | 3 | 0 | 0 | 3 | 100.0 |
| Other | 1 | 1 | 0 | 0 | 0.0 |

## E. Representative claim examples

The examples below illustrate how the labels were applied in practice.

**Direct.** "MiniMax-M1 contains 456B total parameters with 45.9B activated per token" is coded as Direct because the architecture statement is concrete and tied to an open release with inspectable technical documentation (MiniMax, 2025).

**Assumption-burdened.** "DeepSeek-R1 achieves 79.8% pass@1 on AIME 2024" is close to externally checkable because the benchmark, metric, and model family are public, but replication still requires assumptions about generation settings and the exact evaluation procedure (DeepSeek-AI, 2025).

**Not verifiable: ambiguity.** "Claude Sonnet 4.6 is Anthropic's most capable Sonnet model yet" is non-verifiable because "most capable" is not operationalized in the public artifact, leaving the comparison target and scoring criterion unclear (Anthropic, 2026).

**Not verifiable: access / version.** "GPT-5.4 Thinking achieved 0.998 on prompt-injection attacks against email connectors" is numerically precise, but the relevant connector setup, attack suite, and grading environment are not externally available, so the claim cannot be independently checked from public materials alone (OpenAI, 2026b).

## F. Full per-claim audit table

Table 7 provides the full per-claim coding table synchronized to the final frozen audit spreadsheet used to generate the reported aggregate counts. The table includes all 61 claims in the audited corpus, along with the six rubric fields and the final effective verifiability label.

*Table 7.* Full per-claim audit table for the pilot audit sheet. Columns report the six rubric fields and the final label used in the audit spreadsheet.

| ID | Org. | Artifact | Claim text | Claim type | Spec. | Inputs | Metric | Judge | Setup | Access | Final label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking is the first general-purpose model to have implemented mitigations for high capability in cybersecurity. | governance/risk classification | Semi-specific | Insufficient | Partial | Partial | Partial | Partial | Not verifiable: ambiguity |
| 2 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking scores 62.6% on HealthBench. | capability / safety-adjacent evaluation | Specific | Partial | Sufficient | Partial | Partial | Partial | Assumption-burdened |
| 3 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking scores 96.6% on HealthBench Consensus. | capability / safety-adjacent evaluation | Specific | Partial | Sufficient | Partial | Partial | Partial | Assumption-burdened |
| 4 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking achieved 0.998 on prompt-injection attacks against email connectors. | robustness / security | Specific | Partial | Sufficient | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 5 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking achieved 0.978 on prompt-injection attacks against function-call attacks. | robustness / security | Specific | Partial | Sufficient | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 6 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking achieved 0.86 destructive-action avoidance, 0.18 perfect reversion, and 0.53 user-work preserved. | reliability / agentic safety | Specific | Partial | Sufficient | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 7 | OpenAI | GPT-5.3-Codex System Card | GPT-5.3-Codex is the most capable agentic coding model to date. | capability | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 8 | OpenAI | GPT-5.3-Codex System Card | GPT-5.3-Codex is 25% faster. | capability / deployment | Semi-specific | Insufficient | Partial | Sufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 9 | OpenAI | GPT-5.3-Codex System Card | Apollo found overall low rates of covert deceptive behaviors, comparable to GPT-5 and GPT-5.2 Codex. | safety / deception | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 10 | OpenAI | GPT-5.3-Codex System Card | The suite includes non-zero instances of covert action. | safety / deception | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: metric/judge |
| 11 | OpenAI | GPT-5.3-Codex System Card | The model sometimes sandbags on capability Q&A tasks such as biology or chemistry. | safety / deceptive alignment | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: metric/judge |
| 12 | OpenAI | GPT-5.3-Codex System Card | The model demonstrates stronger AI R&D sabotage capabilities than previous models, achieving near-perfect scores on Apollo's tasks and outperforming human baselines. | safety / sabotage risk | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 13 | Anthropic | Claude Sonnet 4.6 System Card | Claude Sonnet 4.6 is Anthropic's most capable Sonnet model yet. | capability | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 14 | Anthropic | Claude Sonnet 4.6 System Card | Sonnet 4.6 is a full upgrade across coding, computer use, long-context reasoning, agent planning, knowledge work, and design. | capability | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 15 | Anthropic | Claude Sonnet 4.6 System Card | Sonnet 4.6 features a 1M token context window in beta. | deployment / model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Partial | Direct |
| 16 | Anthropic | Claude Opus 4.6 System Card / Release Materials | Opus 4.6 shows an overall safety profile as good as, or better than, any other frontier model in the industry. | safety | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 17 | Anthropic | Claude Opus 4.6 System Card / Release Materials | Opus 4.6 has enhanced cybersecurity abilities, motivating six new cybersecurity probes. | safety / cybersecurity | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: inputs/protocol |
| 18 | Google DeepMind | Gemini 3.1 Pro Model Card | Gemini 3.1 Pro is Google's most advanced model for complex tasks as of publication. | capability | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Partial | Not verifiable: ambiguity |
| 19 | Google DeepMind | Gemini 3.1 Pro Model Card | Gemini 3.1 Pro has a token context window of up to 1M and a 64K token output. | deployment / model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Partial | Direct |
| 20 | Google DeepMind | Gemini 3.1 Pro Model Card | Gemini 3.1 Pro significantly outperforms Gemini 3 Pro across a range of benchmarks requiring enhanced reasoning and multimodal capabilities. | capability | Semi-specific | Partial | Partial | Partial | Partial | Partial | Assumption-burdened |
| 21 | Google DeepMind | Gemini 3.1 Pro Model Card | Gemini 3.1 Pro produced +0.10% text-to-text safety and +0.11% multilingual safety versus Gemini 3.0 Pro. | safety | Specific | Partial | Partial | Partial | Partial | Partial | Not verifiable: access/version |
| 22 | Google DeepMind | Gemini 3.1 Pro Model Card | Gemini 3.1 Pro achieved a human-normalized average score of 1.27 on RE-Bench versus Gemini 3 Pro's 1.04. | capability / agentic evaluation | Specific | Partial | Sufficient | Partial | Partial | Partial | Assumption-burdened |
| 23 | Google DeepMind | Gemini 3.1 Pro Model Card | On three situational-awareness challenges, Gemini 3.1 Pro achieved a success rate of almost 100%. | safety / situational awareness | Semi-specific | Partial | Partial | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 24 | DeepSeek | DeepSeek-R1 Technical Report | DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. | capability / reasoning | Semi-specific | Partial | Partial | Partial | Partial | Partial | Assumption-burdened |
| 25 | DeepSeek | DeepSeek-R1 Technical Report | Reasoning abilities can be incentivized through pure reinforcement learning, without supervised fine-tuning as a preliminary step. | methodology / capability | Semi-specific | Partial | Partial | Sufficient | Partial | Partial | Not verifiable: inputs/protocol |

| ID | Org. | Artifact | Claim text | Claim type | Spec. | Inputs | Metric | Judge | Setup | Access | Final label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | DeepSeek | DeepSeek-R1 Technical Report | DeepSeek-R1-Zero exhibits emergent behaviors such as self-verification, reflection, and generating long chains of thought. | capability / reasoning behavior | Semi-specific | Insufficient | Insufficient | Partial | Partial | Partial | Not verifiable: metric/judge |
| 27 | Z.AI | GLM-4.5 Technical Report | GLM-4.5 is an open-source MoE LLM with 355B total parameters and 32B activated parameters. | model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 28 | Z.AI | GLM-4.5 Technical Report | GLM-4.5 supports both thinking and direct response modes through a hybrid reasoning method. | model capability / mode | Specific | Sufficient | Sufficient | Sufficient | Partial | Accessible | Direct |
| 29 | Z.AI | GLM-4.5 Technical Report | GLM-4.5 achieves strong performance across agentic, reasoning, and coding tasks. | capability | Semi-specific | Partial | Partial | Partial | Partial | Accessible | Not verifiable: ambiguity |
| 30 | Qwen | Qwen3 Technical Report | Qwen3 is the latest version of the Qwen model family. | release / model family | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 31 | Qwen | Qwen3 Technical Report | Qwen3 advances performance, efficiency, and multilingual capabilities. | capability | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Accessible | Not verifiable: ambiguity |
| 32 | Qwen | Qwen3 Technical Report | The Qwen3 series includes models of both dense and MoE architectures, with parameter scales ranging from 0.6B to 235B. | model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 33 | Qwen | Qwen3 Technical Report | Qwen3 expands multilingual support from 29 to 119 languages and dialects. | multilingual capability / deployment | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 34 | MiniMax | MiniMax-M1 Technical Report | MiniMax-M1 is the world's first open-weight, large-scale hybrid-attention reasoning model. | model positioning | Semi-specific | Insufficient | Insufficient | Insufficient | Insufficient | Accessible | Not verifiable: ambiguity |
| 35 | MiniMax | MiniMax-M1 Technical Report | MiniMax-M1 contains 456B total parameters with 45.9B activated per token. | model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 36 | MiniMax | MiniMax-M1 Technical Report | MiniMax-M1 natively supports a context length of 1 million tokens. | deployment / model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 37 | MiniMax | MiniMax-M1 Technical Report | MiniMax-M1 scales test-time compute efficiently with lightning attention. | efficiency / capability | Semi-specific | Partial | Partial | Partial | Partial | Accessible | Not verifiable: ambiguity |
| 38 | Cohere | Command A Technical Report | Command A can deliver up to 156 tokens/sec, 1.75× higher than GPT-4o and 2.4× higher than DeepSeek V3. | efficiency / deployment | Specific | Partial | Sufficient | Sufficient | Partial | Partial | Not verifiable: access/version |
| 39 | Cohere | Command A Technical Report | Command A and Command R7B perform best-in-class across established benchmarks for their respective model sizes. | capability | Semi-specific | Partial | Partial | Partial | Partial | Partial | Not verifiable: ambiguity |
| 40 | Cohere | Command A Technical Report | Weights for both models have been released for research purposes. | access / release | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 41 | Anthropic | Claude Sonnet 4.6 System Card | When connecting to external data sources and tools via Model Context Protocol... | robustness / security | Specific | Partial | Sufficient | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 42 | Anthropic | Claude Sonnet 4.6 System Card | Claude Sonnet 4.6 was released under the AI Safety Level 3 (ASL-3) standard f... | governance / risk classification | Semi-specific | Insufficient | Insufficient | Insufficient | Insufficient | Inaccessible | Not verifiable: metric/judge |
| 43 | Anthropic | Claude Opus 4.6 System Card / Release Materials | Claude Opus 4.6 demonstrated autonomous discovery of hundreds of zero-day vul... | safety / cybersecurity | Semi-specific | Insufficient | Partial | Partial | Insufficient | Inaccessible | Not verifiable: inputs/protocol |
| 44 | OpenAI | GPT-5.3-Codex System Card | In Apollo Research's sabotage suite, GPT-5.3-Codex achieved a mean best-of-10... | safety / sabotage capability | Specific | Partial | Sufficient | Partial | Partial | Inaccessible | Not verifiable: access/version |
| 45 | Meta | Llama 4 Model Card (Scout / Maverick) | Llama 4 models do not introduce risk plausibly enabling catastrophic cyber ou... | safety / cybersecurity | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Accessible | Not verifiable: ambiguity |
| 46 | Meta | Llama 4 Model Card (Scout / Maverick) | Llama 4 was evaluated with adversarial datasets against CBRNE (chemical, biol... | safety / CBRN | Semi-specific | Insufficient | Insufficient | Insufficient | Insufficient | Accessible | Not verifiable: inputs/protocol |
| 47 | xAI | Grok 4 Model Card | Third-party testing shows that Grok 4's end-to-end offensive cyber capabiliti... | safety / cybersecurity | Semi-specific | Insufficient | Insufficient | Partial | Insufficient | Inaccessible | Not verifiable: metric/judge |
| 48 | xAI | Grok 4 Model Card | Warning Grok 4 against jailbreak attacks in the system prompt serves to signi... | robustness / safeguards | Vague | Insufficient | Insufficient | Insufficient | Partial | Partial | Not verifiable: ambiguity |
| 49 | xAI | Grok 4.1 Model Card | Results in previous xAI model cards were reported with an error in the evalua... | methodology / corrigendum | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Partial | Direct |
| 50 | UK AI Security Institute | Frontier AI Trends Report 2025 | In the cyber domain, AI models can now complete apprentice-level tasks 50% of... | capability / cybersecurity | Specific | Partial | Sufficient | Partial | Insufficient | Inaccessible | Not verifiable: access/version |
| 51 | UK AI Security Institute | Frontier AI Trends Report 2025 | AISI has found universal jailbreaks — techniques that override safeguards acr... | safety / safeguards | Semi-specific | Insufficient | Insufficient | Insufficient | Insufficient | Inaccessible | Not verifiable: inputs/protocol |
| 52 | UK AI Security Institute | Frontier AI Trends Report 2025 | Open-source models are catching up, with the capability gap narrowing to arou... | capability / ecosystem | Semi-specific | Insufficient | Partial | Insufficient | Insufficient | Inaccessible | Not verifiable: metric/judge |
| 53 | Mistral AI | Mistral Large 2 Announcement / Model Card | Mistral Large 2 has a 128k context window. | deployment / model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 54 | OpenAI | GPT-5.4 Thinking System Card | GPT-5.4 Thinking does not have a plausible chance of reaching a High threshol... | governance / risk classification | Semi-specific | Insufficient | Insufficient | Insufficient | Insufficient | Inaccessible | Not verifiable: metric/judge |
| 55 | OpenAI | GPT-5.4 Thinking System Card | Based on the observed distribution of conversations with GPT-5.2 Thinking, 99... | safety / policy compliance | Specific | Insufficient | Sufficient | Partial | Insufficient | Inaccessible | Not verifiable: inputs/protocol |
| 56 | DeepSeek | DeepSeek-R1 Technical Report | DeepSeek-R1 achieves 79.8% pass@1 on AIME 2024. | capability / reasoning benchmark | Specific | Sufficient | Sufficient | Sufficient | Partial | Accessible | Assumption-burdened |

| ID | Org. | Artifact | Claim text | Claim type | Spec. | Inputs | Metric | Judge | Setup | Access | Final label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | Anthropic | Claude Haiku 4.5 System Card | Claude Haiku 4.5 showed signs of awareness that the model was in an evaluatio... | safety / evaluation awareness | Specific | Insufficient | Sufficient | Partial | Insufficient | Inaccessible | Not verifiable: inputs/protocol |
| 58 | Z.AI | GLM-4.5 Technical Report | GLM-4.5 ranks 3rd overall and 2nd on agentic benchmarks among tested propriet... | capability / ranking | Semi-specific | Insufficient | Partial | Partial | Partial | Accessible | Not verifiable: inputs/protocol |
| 59 | Anthropic | Claude Opus 4.6 System Card / Release Materials | Claude Opus 4.6 performed at or below the level of Claude Opus 4.6 on automat... | governance / comparative safety | Vague | Insufficient | Insufficient | Insufficient | Insufficient | Inaccessible | Not verifiable: ambiguity |
| 60 | Meta | Llama 4 Model Card (Scout / Maverick) | Llama 4 Scout has 17B activated parameters and 16 experts. | model specification | Specific | Sufficient | Sufficient | Sufficient | Sufficient | Accessible | Direct |
| 61 | OpenAI | GPT-5.4 Thinking System Card | On cyber range scenarios testing end-to-end cyber operations, GPT-5.4 Thinking achieved a combined success rate of 73.33%, compared to GPT-5.3-Codex at 80%. | safety / cybersecurity | Specific | Insufficient | Sufficient | Partial | Insufficient | Inaccessible | Not verifiable: inputs/protocol |

## G. Disclosure standard

This appendix section summarizes the minimum disclosure standard implied by the audit results. The standard is intentionally narrow: it does not require developers to disclose every internal evaluation detail, but it does require that any public claim presented as externally meaningful be labeled and documented in a way that lets readers distinguish between independently checkable evidence, assumption-sensitive evidence, and internal-only evidence.

**1. Label the evidence type.** Mark each public claim as independently checkable, assumption-dependent, or internal-only.

**2. Tie capability claims to a benchmark.** If a claim is comparative or benchmark-based, name the benchmark and make clear whether an external party can access it.

**3. Specify how the score was produced.** Report the metric, threshold, pass criteria, judge configuration, and any material aggregation or selection rule such as best-of-$n$ sampling.

**4. Identify the evaluated system.** Record the exact checkpoint, release snapshot, tools, and deployment configuration used for the reported result.

**5. State access limits explicitly.** If core prompts, connectors, datasets, or harnesses are unavailable, say so plainly rather than implying external reproducibility.

*Figure 2.* Minimum disclosure standard implied by the audit. The goal is not maximal transparency in every case, but clear labeling and enough procedural detail for outside readers to understand what is and is not independently checkable.

In practical terms, the standard has five parts. First, every public claim should identify its evidence type: independently checkable, assumption-dependent, or internal-only. Second, capability claims should be tied to an independently accessible benchmark whenever possible; if this is impossible, the claim should say so directly. Third, the disclosure should specify the scoring methodology in enough detail to support replication, including metrics, thresholds, grader setup, and material sampling or selection rules. Fourth, the reported result should identify the exact model version, checkpoint, or deployment snapshot evaluated. Fifth, any binding access constraints should be stated explicitly so that readers do not confuse developer attestation with externally reproducible evidence.

This standard is deliberately modest. It does not require releasing every proprietary harness or sensitive safety procedure. It does require that developers distinguish claims that invite external verification from claims that rest primarily on internal evaluation. On the evidence in this audit, making that distinction explicit would already improve the quality and interpretability of public AI disclosures.

## H. Limitations and release plan

We note several limitations. First, we did not rerun the underlying evaluations or benchmarks for the claims in our corpus; our goal is to assess whether those claims are independently verifiable *as disclosed*. Second, although the appendix now includes the full per-claim coding table used to generate the reported aggregate outcomes, corpus summaries, and representative examples, claim extraction and coding still require judgment. Third, the ratings were assigned in a single-rater pass, so the paper should be read as a proof of concept for the rubric rather than as a statistically robust estimate of industry-wide disclosure quality. A more rigorous follow-on study should incorporate multiple raters, report formal agreement statistics, and test sensitivity to claim sampling.