

Source dossier

The reviewer was big enough — a weak model can audit a strong coding agent when the evidence is checkable, and the Article 14 that binds today is not the one you reach for

Edition 2026-10-06 · narrative: oversight-needs-a-decisive-check · compiled 2026-10-06

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-06-01	Groundability, Not Scale Alone: When Weak Reviewers Can Audit Strong Coding Agents is arXiv:2610.01023, v1 submitted 1 October 2026, by researchers at the University of California, Berkeley and Virginia Tech.	verified	arxiv2026groundabi, arxiv2026html
CLAIM-2026-10-06-02	The study covers 411 execution-labelled traces from three agents (GPT-5.4, Gemini-2.5-Pro, Claude-Sonnet-5) plus 101 controlled cases, reviewed by six models: Llama-3.1-8B, Qwen-2.5-Coder-32B, Qwen-2.5-72B, GPT-OSS-120B, Qwen3-235B and GPT-4.1.	verified	arxiv2026groundabi, arxiv2026html
CLAIM-2026-10-06-03	On 122 held-out traces supplied with official execution evidence, five of six reviewers improved both catch and over-rejection versus structured evidence, and two — GPT-OSS-120B and GPT-4.1 — classified all 122 correctly.	verified	arxiv2026groundabi, arxiv2026html
CLAIM-2026-10-06-04	On 154 GPT-5.4 traces, structured but unchecked evidence raised both defect catch and over-rejection.	verified	arxiv2026groundabi
CLAIM-2026-10-06-05	The paper states that reviewer size is not a consistent predictor of review quality; parameter count showed no monotonic correlation across the model ladder tested.	verified	arxiv2026groundabi, arxiv2026html
CLAIM-2026-10-06-06	The frozen cascade — patch-caused static errors plus generated tests that first fail on the unpatched repository — scored coverage 0.89, risk 0.33, catch 0.76 and over-rejection 0.66 on 121 scored held-out GPT-5.4 traces, and coverage 0.86, risk 0.26, catch 0.80 and over-rejection 0.67 on 59 Gemini traces.	verified	arxiv2026groundabi, arxiv2026html
CLAIM-2026-10-06-07	In the paper, coverage is the fraction of traces receiving a decision rather than abstention, risk is the error rate among non-uncertain decisions, catch rate is $P(\text{reject} \mid \text{defect})$ and over-rejection rate is $P(\text{reject} \mid \text{acceptable})$.	verified	arxiv2026html
CLAIM-2026-10-06-08	The paper reports that most false rejections occur when unresolved cases reach the reviewer.	verified	arxiv2026groundabi

Claim	Statement	Status	Sources
CLAIM-2026-10-06-09	AI Act Article 14 (human oversight) sits in Chapter III, Section 2 and applies to high-risk AI systems; its application moved to 2 December 2027 for Annex III systems and 2 August 2028 for Annex I under Regulation (EU) 2026/1744.	verified	euaiactexpl2026human, eurlex2026regulatio
CLAIM-2026-10-06-10	Cyber Resilience Act Article 14 reporting has applied to manufacturers since 11 September 2026: an early warning within 24 hours of becoming aware of an actively exploited vulnerability, a full notification within 72 hours, and a final report no later than 14 days after a corrective measure is available; open-source software stewards are not in scope until 11 December 2027, and reports go to the ENISA Single Reporting Platform under Article 16.	verified	europeancom2026cra, eurlex2024regulatio, cyberresili2026cyber
CLAIM-2026-10-06-11	AI Act Article 14(4)(a) requires that oversight personnel be enabled to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance.	verified	euaiactexpl2026human

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

Groundability, Not Scale Alone: When Weak Reviewers Can Audit Strong Coding Agents

PRIMARY

Publisher	arXiv
Published	2026-10-01
Retrieved	2026-10-06
Cite key	arxiv2026groundabi
URL	https://arxiv.org/abs/2610.01023

Assessment. Primary. The work itself. v1 submitted 1 October 2026 04:10:59 UTC; the submission timestamp is arXiv's own record, not a coverage date. Subject classes cs.SE, cs.AI, cs.CL. Identifier checked against the story: 2610 matches an October 2026 submission and the title on the abs page is the title used here.

Authors. Junyu Guo, Shangding Gu, Ming Jin, Javad Lavaei.

Abstract (verbatim). Coding agents can return plausible patches that omit required behavior. These failures are hard to review because long traces and confident summaries often hide what was missed. We ask when a nominally weaker reviewer can reliably decide whether a patch solves its issue. We study 411 execution-labeled traces from three agents and 101 controlled cases. On 154 GPT-5.4 traces, structured but unchecked evidence raises both defect catch and over-rejection. We then provide official execution evidence as an upper-bound diagnostic. After choosing and freezing one of two formats per reviewer, five of six reviewers improve both rates on 122 held-out traces; two classify every trace correctly. Reviewer size is not a consistent predictor of quality. Because official tests are unavailable in deployment, we also evaluate a frozen cascade with patch-caused static errors and generated tests that first fail on the unpatched repository. On 121 scored held-out GPT-5.4 traces and 59 Gemini traces, its coverage is 0.89 and 0.86, risk is 0.33 and 0.26, catch is 0.76 and 0.80, and over-rejection is 0.66 and 0.67. Most false rejections occur when unresolved cases reach the reviewer. Official execution evidence shows the potential of weak review when decisive checks are available. Producing equally reliable checks without official tests remains the main bottleneck.

HTML full text — Groundability, Not Scale Alone

PRIMARY

Publisher	arXiv
Published	2026-10-01
Retrieved	2026-10-06
Cite key	arxiv2026html
URL	https://arxiv.org/html/2610.01023v1

Assessment. Primary, and the only place the operative definitions appear. Affiliations were taken from this rendering of the work's own front matter, which gives University of California, Berkeley and Virginia Tech with the corresponding author at berkeley.edu; the per-author mapping is not reproduced here, so the edition says only that the researchers are at those two institutions. The PDF was not readable this run — the fetch proxy rate-limited it (HTTP 429) and was not retried.

Metric definitions. Coverage: fraction of traces receiving a decision rather than abstention. Risk: error rate among non-uncertain decisions. Catch rate: $P(\text{reject} \mid \text{defect})$. Over-rejection rate: $P(\text{reject} \mid \text{acceptable})$.

Evidence formats. Structured (non-grounded) evidence is pipeline-selected but independently unchecked — changed code and heuristic signals. Grounded evidence is backed by independent verification such as static analysis or test execution.

Composition of the 411 real traces. 154 GPT-5.4 (32 design, 122 held-out), 59 Gemini, 78 multilingual GPT-5.4, 120 Claude-Sonnet. Plus 101 synthetic/controlled cases: 37 reference patches, 34 ablations, 30 naturalistic pilot cases.

Agents audited. GPT-5.4, Gemini-2.5-Pro, Claude-Sonnet-5.

Reviewer models. Llama-3.1-8B, Qwen-2.5-Coder-32B, Qwen-2.5-72B, GPT-OSS-120B, Qwen3-235B, GPT-4.1.

Findings. On 122 held-out traces with official execution evidence, five reviewers improved both catch and over-rejection versus structured evidence; GPT-OSS-120B and GPT-4.1 classified all 122 correctly. Frozen cascade on held-out sets: GPT-5.4 coverage 0.89, risk 0.33, catch 0.76, over-rejection 0.66; Gemini coverage 0.86, risk 0.26, catch 0.80, over-rejection 0.67. Reviewer size is not a consistent predictor of quality — parameter count showed no monotonic correlation across the model ladder.

CRA reporting obligations

PRIMARY

Publisher European Commission

Published 2026-09-11

Retrieved 2026-10-06

Cite key europeancom2026cra

URL <https://digital-strategy.ec.europa.eu/en/policies/cra-reporting>

Assessment. Primary and dispositive on who owes the CRA reporting duty and from when. Cite key already exists in the vault; reused, not duplicated. Note the vault also carries europeancom2026cyber against the same URL — a pre-existing duplicate, not created here.

Who must report. Manufacturers and open-source software stewards, per Article 24(3) CRA.

What must be reported. Actively exploited vulnerabilities; severe incidents having an impact on the security of products with digital elements.

When the obligations apply. Manufacturers: 11 September 2026. Open-source software stewards: 11 December 2027.

Deadlines. Early warning within 24 hours of becoming aware. Full notification within 72 hours. Final report no later than 14 days after a corrective measure is available (for vulnerabilities); within one month of the 72-hour notification (for severe incidents).

Where. The CRA Single Reporting Platform established by ENISA under Article 16 CRA. Notifications reach the CSIRT of the manufacturer's main establishment and are shared with other relevant CSIRTs. Article 71(2) addresses open-source steward obligations.

Regulation (EU) 2024/2847 — Cyber Resilience Act

PRIMARY

Publisher	EUR-Lex
Published	2024-11-20
Retrieved	2026-10-06
Cite key	eurlex2024regulatio
URL	https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng

Assessment. Primary. The instrument itself, for the operative wording of the reporting duty. Cite key already in the vault; reused.

Regulation (EU) 2024/2847 of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). Entered into force 10 December 2024. Article 14 carries the reporting obligations for actively exploited vulnerabilities and severe incidents; Article 16 establishes the single reporting platform. Non-commercial open-source software developed outside a commercial activity is largely outside scope, and open-source stewards carry a lighter, tailored set of obligations.

Human oversight — Article 14

SECONDARY

Publisher	EU AI Act Explorer
Published	2026-08-02
Retrieved	2026-10-06
Cite key	euaiactexpl2026human
URL	https://artificialintelligenceact.eu/article/14/

Assessment. Secondary but faithful article text. Used for the Article 14 wording and its place in the Act's structure, not for any date; the dates come from Regulation (EU) 2026/1744.

Location. Chapter III (High-Risk AI Systems), Section 2 (Requirements for high-risk AI systems). Applies to high-risk AI systems under Annex III and Annex I.

Article 14(1). 'High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.'

Article 14(4)(a). Oversight personnel must be enabled to 'properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance.'

Article 14(4)(d). Personnel must be able to 'decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output'.

Application. 2 December 2027 for Annex III systems; 2 August 2028 for Annex I.

Regulation (EU) 2026/1744 — Digital Omnibus on AI

PRIMARY

Publisher	EUR-Lex
Published	2026-07-27
Retrieved	2026-10-06
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Assessment. Primary. The instrument that moved the Annex III dates, and with them Article 14. Cite key already in the vault; reused.

Regulation (EU) 2026/1744, the Digital Omnibus on AI, in force 27 July 2026. Moves Annex III high-risk obligations to 2 December 2027, Annex I to 2 August 2028, and member-state regulatory sandboxes to 2 August 2027. Article 50 transparency continues to apply from 2 August 2026, with the marking-of-generated-content obligation from 2 December 2026. Article 53 GPAI has been in force since August 2025.

Cyber Resilience Act explained — scope, classes and deadlines

SECONDARY

Publisher CyberResilienceAct.eu

Published 2026-09-11

Retrieved 2026-10-06

Cite key cyberresili2026cyber

URL <https://www.cyberresilienceact.eu/explained.html>

Assessment. Secondary. Used only to corroborate the article number and the timeline already established from the Commission's own page; nothing rests on it alone.

Reporting articles. Articles 14 and 16 carry the reporting duties: 24 hours for an early warning on an actively exploited vulnerability or severe incident, 72 hours for the fuller notification with corrective measures, 14 days for the final report once resolved.

Timeline. 10 December 2024 entry into force. 11 June 2026, Chapter IV applies. 11 September 2026, reporting obligations begin. 31 October / 31 December 2026, core harmonised standards due. 11 December 2027, full application of most provisions.

Open source. Non-commercial open-source software developed outside a commercial activity is largely outside scope; open-source stewards have a lighter, tailored set of obligations.