

Source dossier

Green was a claim about the suite — a third of coding-benchmark passes violate the task, and the duty that names evaluations reaches the model, not your harness

Edition 2026-10-10 · narrative: the evaluator wrote the spec · compiled 2026-10-10

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-10-01	Across five benchmarks and six frontier model backends, about 34.4% of trials judged correct violate the task's requirements.	verified	arxiv2026testjack
CLAIM-2026-10-10-02	The overall resolution rate drops from 50.6% to 33.2% once the audit is applied.	verified	arxiv2026testjack
CLAIM-2026-10-10-03	The 34.4% share and the 17.4-point drop are one measurement expressed two ways: $50.6 \times (1 - 0.344) = 33.19$.	verified	arxiv2026testjack
CLAIM-2026-10-10-04	The method generates per-trial tests targeting requirements the patch might violate, keeps only tests the ground-truth patch passes, and emits replayable evidence per confirmed failure.	verified	arxiv2026testjack
CLAIM-2026-10-10-05	Annex XI Section 2 requires a detailed account of evaluation strategies and results, including criteria, metrics and identified limitations.	verified	euaiactexpl2026technical
CLAIM-2026-10-10-06	Annex XI Section 2 applies only to providers of general-purpose AI models with systemic risk.	verified	euaiactexpl2026technical
CLAIM-2026-10-10-07	Article 13 applies to high-risk systems from 2 December 2027 under Annex III and 2 August 2028 under Annex I, per Article 113(c) after the Digital Omnibus.	verified	euaiactexpl2026informati, eurlex2026regulatio
CLAIM-2026-10-10-08	Article 13(3)(b)(ii) requires the instructions for use to state the expected level of accuracy including its metrics; Article 15(3) requires those levels and metrics to be declared.	verified	euaiactexpl2026informati, euaiactexpl2026accuracy
CLAIM-2026-10-10-09	Article 53(1)(a) requires technical documentation of the model's training and testing process and the results of its evaluation, containing at least the information in Annex XI, and has been in force since 2 August 2025.	verified	euaiactexpl2026generalpu
CLAIM-2026-10-10-10	Article 53(2) exempts free and open-source general-purpose models from 53(1)(a) and (b) where parameters, weights, architecture and usage information are public, but not models with systemic risk.	verified	euaiactexpl2026generalpu

Claim	Statement	Status	Sources
CLAIM-2026-10-10-11	KI-MIG has been in force since 29 July 2026 and the Bundesnetzagentur is the German market surveillance authority.	verified	bundesamtfu2026ki mig

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

TestJack: Should you trust the results in coding benchmarks?

Agentic Coding Benchmarks Auditing via Evaluator Evolution

PRIMARY

Publisher	arXiv
Published	2026-10-07
Retrieved	2026-10-10
Cite key	arxiv2026testjack
URL	https://arxiv.org/abs/2610.10619

Assessment. Primary: the work's own abstract record. Affiliations are NOT printed on the abs page and the HTML rendering (arxiv.org/html/2610.10619v1) returned HTTP 429 to the fetch proxy with an instruction not to retry, so no institution is attributed anywhere in this edition. All published figures come from the abstract; no per-benchmark or per-model breakdown is claimed. Identifier verified against the story before use: title and YYYYMM both match a 7 October 2026 submission.

Claims drawn from this source. CLAIM-2026-10-10-01, CLAIM-2026-10-10-02, CLAIM-2026-10-10-03, CLAIM-2026-10-10-04

arXiv:2610.10619 [cs.SE, cs.AI]. v1 posted Wed, 7 Oct 2026 07:58:04 UTC (389 KB).
Authors: Shuangjie Yao, Hao Wang, Koushik Sen, Simin Chen, Baishakhi Ray, Dawn Song. No institutional affiliation is printed on the abs record.

Abstract, as extracted: most coding benchmarks judge a solution correct if it passes a fixed set of unit tests. Those tests often miss required behaviour, so agents can pass them while failing the task, and benchmark scores may therefore reflect adaptation to the evaluator rather than problem-solving skill. The authors introduce TestJack, which generates tests for each trial that target requirements the patch might violate, keeps only tests the ground-truth patch passes, and produces replayable evidence for each confirmed failure. A lighter variant audits a sample of trials and reuses the tests across trials for the same task. Across six frontier model backends and five benchmarks, about 34.4% of trials judged correct violate task requirements, and the overall resolution rate drops from 50.6% to 33.2%.

Arithmetic check performed in this run: $50.6 \times (1 - 0.344) = 33.19$, so the 34.4% figure and the 17.4-point drop are one measurement expressed two ways, not two independent findings. This is stated in the post rather than left for a reader to assume.

Retrieval note: arxiv.org/html/2610.10619v1 returned HTTP 429 from the fetch proxy with an explicit instruction not to retry. The PDF was therefore not attempted. No number beyond the three above is used.

Technical documentation for general-purpose AI models (Annex XI)

SECONDARY

Publisher	EU AI Act Explorer
Published	n.d.
Retrieved	2026-10-10
Cite key	euaiactexpl2026technical
URL	https://artificialintelligenceact.eu/annex/11/

Assessment. Secondary: a reproduction of the Annex text, not the Official Journal. Used for the Section 1 / Section 2 split, which is the load-bearing distinction in this edition. The page does not print an application date for Annex XI; that was taken from the Article 53 page instead.

Claims drawn from this source. CLAIM-2026-10-10-05, CLAIM-2026-10-10-06

Section 1 - information required of ALL providers of general-purpose AI models, proportionate to size and risk profile: a general description (intended tasks, the kinds of AI systems the model can be built into, acceptable use policies, release date and distribution methods, architecture and parameter count, input and output modalities and formats, licence); a detailed description (technical means for integration, design and training specifications with key design choices and their rationale, training/testing/validation data including sources, curation, scope and bias checks, and training compute and time); and known or estimated energy consumption.

Section 2 - ADDITIONAL information required only of providers of general-purpose AI models WITH SYSTEMIC RISK: a detailed account of evaluation strategies and results, based on public protocols and tools or other methods, including evaluation criteria, metrics and how limitations are identified; where applicable, internal and/or external adversarial testing such as red teaming, plus model adaptations like alignment and fine-tuning; and where applicable, the system architecture and how software components feed into one another.

The distinction used in this edition: the detailed account of evaluation strategies, criteria and metrics is a Section 2 item and therefore reaches systemic-risk models only, and it documents the MODEL rather than any downstream harness or benchmark.

Information provided to deployers of high-risk AI systems (Article 13)

SECONDARY

Publisher	EU AI Act Explorer
Published	n.d.
Retrieved	2026-10-10
Cite key	euaiactexpl2026informati
URL	https://artificialintelligenceact.eu/article/13/

Assessment. Secondary reproduction. 13(3)(b)(ii) is the paragraph cited in the post. The page's own footer carries the Article 113(c) application dates, which match the dates the Digital Omnibus set.

Claims drawn from this source. CLAIM-2026-10-10-07, CLAIM-2026-10-10-08

Article 13(3)(b)(ii) requires the instructions for use to state the level of accuracy the high-risk system is expected to reach, including its metrics, together with the robustness and cybersecurity levels against which it has been tested and validated. The key phrase is 'the level of accuracy, including its metrics, robustness and cybersecurity'. The instructions must also disclose known and foreseeable circumstances that may affect those expected levels.

Application dates printed on the page, under Article 113(c): 2 December 2027 for high-risk systems under Annex III, 2 August 2028 for high-risk systems under Annex I.

Accuracy, robustness and cybersecurity (Article 15)

SECONDARY

Publisher	EU AI Act Explorer
Published	2026
Retrieved	2026-10-10
Cite key	euaiactexpl2026accuracy
URL	https://artificialintelligenceact.eu/article/15/

Assessment. Secondary reproduction. Cited for 15(3), the declared-accuracy obligation that Article 13's instructions-for-use duty reports against. Same application dates as Article 13.

Claims drawn from this source. CLAIM-2026-10-10-08

15(1): high-risk systems must be designed and developed to achieve an appropriate level of accuracy, robustness and cybersecurity, and to perform consistently in those respects across their lifecycle. 15(2): the Commission, with relevant stakeholders including metrology and benchmarking authorities, is to encourage the development of benchmarks and measurement methodologies for accuracy, robustness and other performance metrics. 15(3): the levels of accuracy and the relevant accuracy metrics shall be declared in the accompanying instructions of use. 15(4): systems are to be as resilient as possible to errors, faults and inconsistencies, including those arising from interaction with people or other systems; technical redundancy, which may include backup or fail-safe plans, may support this. Continuously learning systems must mitigate feedback loops that would let biased outputs influence future inputs. 15(5): systems must be resilient against unauthorised third parties altering their use, outputs or performance by exploiting vulnerabilities, with measures appropriate to the circumstances and risks, addressing where appropriate data poisoning, model poisoning, adversarial examples and confidentiality attacks.

Application dates printed on the page, under Article 113(c): 2 December 2027 for Annex III, 2 August 2028 for Annex I.

General-purpose AI model provider obligations (Article 53)

SECONDARY

Publisher	EU AI Act Explorer
Published	n.d.
Retrieved	2026-10-10
Cite key	euaiactexpl2026generalpu
URL	https://artificialintelligenceact.eu/article/53/

Assessment. Secondary reproduction. Read during the run and it changed the edition: 53(1)(a)'s own wording covers the model's training, testing and evaluation results for EVERY GPAI provider, with Annex XI as the floor. The first comment was corrected before delivery so that the Section 2 point is not read as 'non-systemic-risk providers owe nothing on evaluation'.

Claims drawn from this source. CLAIM-2026-10-10-09, CLAIM-2026-10-10-10

Comes into force 2 August 2025, according to Article 113(b). 53(1)(a): providers of general-purpose AI models shall draw up and keep up to date the technical documentation of the model, including its training and testing process and the results of its evaluation, containing at least the information set out in Annex XI, so that it can be provided to the AI Office and national competent authorities on request. 53(1)(b): documentation for downstream providers who intend to integrate the model, enabling them to understand its capabilities and limitations, containing at least the elements in Annex XII; intellectual property and trade secrets are protected. 53(1)(c): a policy to comply with Union copyright law, including identifying and respecting reservations of rights under Article 4(3) of Directive (EU) 2019/790, using state-of-the-art technologies where relevant. 53(1)(d): a sufficiently detailed summary of the content used for training, following a template provided by the AI Office. 53(2): obligations (a) and (b) do not apply to providers of models released under a free and open-source licence that allows access, use, modification and distribution, and whose parameters including weights, architecture and usage information are made publicly available. This exemption does NOT apply to general-purpose AI models with systemic risk. 53(3): providers shall cooperate with the Commission and national competent authorities. 53(4): until a harmonised standard is published, providers may rely on codes of practice under Article 56 to demonstrate compliance with paragraph 1; conformity with a harmonised standard gives a presumption of conformity. Providers relying on neither must demonstrate alternative adequate means of compliance for the Commission's assessment. 53(5)-(6): delegated acts may set measurement and calculation methods for Annex XI points 2(d) and (e), and may update Annexes XI and XII. 53(7): information obtained under this Article, including trade secrets, is subject to the confidentiality obligations in Article 78.

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

PRIMARY

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

Assessment. Primary instrument, cited for the moved high-risk application dates. Not quoted: as in previous runs, EUR-Lex returns metadata-only or empty text to the fetch proxy, so the dates are taken from the Article 13 and Article 15 pages' Article 113(c) footers and the instrument is cited for the instrument, not for wording.

Claims drawn from this source. CLAIM-2026-10-10-07

Full text not archived — discovery/headline reference only.

KI-MIG: KI-Marktueberwachungs- und Innovationsfoerderungsgesetz

PRIMARY

Publisher	Bundesamt fuer Justiz
Published	2026-07-29
Retrieved	2026-10-10
Cite key	bundesamtfu2026kimig
URL	https://www.gesetze-im-internet.de/ki-mig/

Assessment. Primary. Cited only for who would ask in Germany: the Bundesnetzagentur as market surveillance authority, in force since 29 July 2026. Nothing in the edition turns on its wording.

Claims drawn from this source. CLAIM-2026-10-10-11

Full text not archived — discovery/headline reference only.