

Source dossier

It held until the restart - coding agents implement the full fix on production incidents and then undo it, and the duty that bites is the one that asks what you fixed

Edition 2026-10-03 · narrative: declaration-is-not-resolution · compiled 2026-10-03

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-03-01	Incident-Arena, arXiv:2610.00648v1, submitted 30 September 2026, is a human-built benchmark of 20 carefully selected tasks grounded in real-world deployed open source software; each task deploys a production application to an ephemeral Kubernetes cluster, injects parameterizable faults and applies a sustained load profile.	verified	arxiv2026incidenta
CLAIM-2026-10-03-02	Across 20 human-built tasks and 3 application substrates, the best model configuration passes at 64.3%.	verified	arxiv2026incidenta
CLAIM-2026-10-03-03	The study evaluates 10 agent configurations across the 20 tasks at 150 trials per task, producing 3,000 model trials.	verified	arxiv2026full
CLAIM-2026-10-03-04	Of 2,967 episodes with usable action records, 1,638 (55%) issued a full fix implementing all reference-solution actions, and of those only 59% (966 of 1,638) pass.	verified	arxiv2026full
CLAIM-2026-10-03-05	Of the 672 episodes that implemented a complete fix and still failed, 340 failed only the outcome gate and 329 failed safety checks, primarily from agents undoing their own fixes: 145 by moving the maintenance window to a second that was still unsafe and 38 by raising the connection pool past its ceiling after it had been fixed.	verified	arxiv2026full
CLAIM-2026-10-03-06	In 782 failures the change held at declaration but could not persist through a restart or past the next trigger, and in 448 failures the agent changed a setting the task did not permit, 261 of those touching a different subsystem from the fault.	verified	arxiv2026full
CLAIM-2026-10-03-07	Across eight models with five reasoning settings, pass@1 gains 4.6 points from the lowest setting to the highest, with negligible differential gains on complex versus simple tasks.	verified	arxiv2026full

Claim	Statement	Status	Sources
CLAIM-2026-10-03-08	Incident-Arena grades through two gates that must both pass: an outcome gate measuring service-level indicators under continued traffic, and a safety gate checking restart durability, re-occurrence of the fault trigger, scope compliance and data or guardrail integrity; grading is deterministic and LLM-free.	verified	arxiv2026full, arxiv2026incidenta
CLAIM-2026-10-03-09	Article 23(4)(d) of Directive (EU) 2022/2555 (NIS2) requires a final report not later than one month after the submission of the incident notification, including a detailed description of the incident, the type of threat or root cause that is likely to have triggered it, applied and ongoing mitigation measures, and where applicable its cross-border impact.	verified	nisdirectiv2022nis, eurlex2022directive
CLAIM-2026-10-03-10	Point 2 of Annex III to the AI Act covers AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity.	verified	euartificia2024annex
CLAIM-2026-10-03-11	Annex III high-risk obligations under the AI Act were moved to 2 December 2027 by the Digital Omnibus, Regulation (EU) 2026/1744, in force 27 July 2026, while Article 50 transparency has applied since 2 August 2026 and the marking-of-generated-content obligation applies from 2 December 2026.	verified	eurlex2026regulation, euartificia2026transpare
CLAIM-2026-10-03-12	Thirteen of the twenty Incident-Arena tasks sit on a single application substrate.	verified	arxiv2026full
CLAIM-2026-10-03-13	The paper identifies its best-performing configuration as a specific named vendor model at its highest reasoning setting.	single-source	arxiv2026full

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

Incident-Arena: Getting agents to the last nine of reliability

PRIMARY

Publisher arXiv
Published 2026-09-30
Retrieved 2026-10-03
Cite key arxiv2026incidenta
URL <https://arxiv.org/abs/2610.00648>

Assessment. The work's own record. Front matter read directly for authors and affiliations; the abstract read verbatim for the task count, the substrate count and the 64.3% headline. The abs page was rate-limited on a second request, so the HTML full text was used for the body figures and is archived separately.

Claims drawn from this source. CLAIM-2026-10-03-01, CLAIM-2026-10-03-02, CLAIM-2026-10-03-08

Front matter

Title. Incident-Arena: Getting agents to the last nine of reliability

Authors and affiliations, as printed in the work's own front matter. Andre Fu (Abundant AI); Malik Drabla (Adrenaline AI); Leon Liu (Carnegie Mellon University); Meji Abidoye (Abundant AI); Marek Šuppa (Comenius University in Bratislava); Lata Mishra (Massachusetts General Hospital); Adnan El Assadi (Abundant AI); Yiyuan Li (Abundant AI).

Identifier. arXiv:2610.00648v1, submitted 30 September 2026.

Abstract, verbatim

AI coding agents are ubiquitous in software engineering workflows in both academia & industry. Yet, despite their use in app coding, relatively less attention has been paid to their ability to execute on production incident response...we introduce Incident-Arena, a human-built benchmark of 20 carefully selected tasks grounded in real-world deployed open source software. Each task deploys a production application to an ephemeral Kubernetes cluster, injecting parameterizable faults, and a sustained load profile given the task requirements...Across 20 human-built tasks and 3 application substrates, the best model configuration passes at 64.3%, with failures extending from diagnosis/localization errors, through incomplete repairs and unsafe regressions.

Identifier check

The YYMM of the identifier (2610) is consistent with a 30 September 2026 submission, and the title on the record matches the story. The adjacent record 2610.00651, Agent Evaluation Reliability: More Tasks Won't (Always) Fix An Agent Leaderboard, is a different paper on a neighbouring subject and was set aside; no figure in this edition comes from it.

Full text and failure taxonomy - Incident-Arena

PRIMARY

Publisher	arXiv
Published	2026-09-30
Retrieved	2026-10-03
Cite key	arxiv2026full
URL	https://arxiv.org/html/2610.00648

Assessment. The paper's own HTML rendering, used because the abs page was rate-limited after the first read. Every body figure in this edition comes from here. The three substrate names reported by the extraction pass could not be reconciled with the abstract's 'open source software' for one of the three, so no substrate is named anywhere in the edition.

Claims drawn from this source. CLAIM-2026-10-03-03, CLAIM-2026-10-03-04, CLAIM-2026-10-03-05, CLAIM-2026-10-03-06, CLAIM-2026-10-03-07, CLAIM-2026-10-03-12

Scale

Ten agent configurations, twenty tasks, 150 trials per task: "across 20 tasks this produces 3,000 model trials".

The conditional that carries the story

Of 2,967 episodes with usable action records, 1,638 (55%) issued a full fix implementing all reference-solution actions. "Of the episodes that implement a full fix, only 59% (966 of 1,638) pass."

Where the remaining 672 went

340 failed only the outcome gate - premature declaration. 329 failed safety checks, "primarily from agents undoing their own fixes by moving the maintenance window to a second that was still unsafe (145) or raising the connection pool past its ceiling after it had been fixed (38)".

Durability and scope

"In 782 failures the change held at declaration, but could not persist through a restart, or past the next trigger." In 448 failures agents "changed a setting the task did not permit", with "261 touched a different subsystem from the fault".

The reasoning dial

Across eight models with five reasoning settings: "From the lowest setting to the highest, pass@1 gains 4.6 points", with negligible differential gains on complex versus simple tasks. The authors' conclusion: "Deliberation does not touch the failure" of premature stopping.

The two gates

Outcome gate: service-level indicators - goodput, latency, error rates - under continued traffic. Safety gate: restart durability of affected services, re-occurrence of the fault trigger, scope compliance, and data/guardrail integrity. Both must pass. Grading is deterministic and LLM-free.

Limitation carried into the edition as a hedge

Thirteen of the twenty tasks sit on a single application substrate.

NIS 2 Directive Article 23 - Reporting obligations

SECONDARY

Publisher NIS-2-Directive.com

Published 2022-12-27

Retrieved 2026-10-03

Cite key nisdirectiv2022nis

URL https://www.nis-2-directive.com/NIS_2_Directive_Article_23.html

Assessment. Existing vault key, reused for the same document. Used this run because the official EUR-Lex rendering returned only page metadata to the fetcher, twice, on two URL forms. The paragraph 4 text below is what was read; the OJ link is published alongside it so a reader can check the official wording.

Claims drawn from this source. CLAIM-2026-10-03-09

Article 23(3), verbatim

"An incident shall be considered to be significant if: (a) it has caused or is capable of causing severe operational disruption of the services or financial loss for the entity concerned; (b) it has affected or is capable of affecting other natural or legal persons by causing considerable material or non-material damage."

Article 23(4)(a)-(d), verbatim

"(a) without undue delay and in any event within 24 hours of becoming aware of the significant incident, an early warning, which, where applicable, shall indicate whether the significant incident is suspected of being caused by unlawful or malicious acts or could have a cross-border impact; (b) without undue delay and in any event within 72 hours of becoming aware of the significant incident, an incident notification, which, where applicable, shall update the information referred to in point (a) and indicate an initial assessment of the significant incident, including its severity and impact, as well as, where available, the indicators of compromise; (c) upon the request of a CSIRT or, where applicable, the competent authority, an intermediate report on relevant status updates; (d) a final report not later than one month after the submission of the incident notification under point (b), including the following: (i) a detailed description of the incident, including its severity and impact; (ii) the type of threat or root cause that is likely to have triggered the incident; (iii) applied and ongoing mitigation measures; (iv) where applicable, the cross-border impact of the incident"

The limb this edition turns on

Point (d)(iii) - "applied and ongoing mitigation measures" - is a statement of fact about a repair, made one month after the notification. A repair that was true at the agent's declaration and false after a restart makes that statement untrue.

Directive (EU) 2022-2555 (NIS2) - official text

PRIMARY

Publisher	EUR-Lex
Published	2022-12-27
Retrieved	2026-10-03
Cite key	eurlex2022directive
URL	https://eur-lex.europa.eu/eli/dir/2022/2555/oj

Assessment. Published as the official text for the article quoted. Not archived with a body: the fetcher returned only navigation and metadata from both the ELI and the CELEX HTML renderings this run, so the article wording was read on the secondary and that is recorded rather than papered over.

Claims drawn from this source. CLAIM-2026-10-03-09

Full text not archived — discovery/headline reference only.

Annex III: High-Risk AI Systems Referred to in Article 6(2)

SECONDARY

Publisher	EU Artificial Intelligence Act
Published	2024-07-12
Retrieved	2026-10-03
Cite key	euartificia2024annex
URL	https://artificialintelligenceact.eu/annex/3/

Assessment. Existing vault key for this exact URL, reused rather than duplicated. Kept on this publisher spelling deliberately: the euaiactexpl* spelling of the same site already holds Annex I under euaiactexpl2026annex, and a second Annex note on that spelling would have collided.

Claims drawn from this source. CLAIM-2026-10-03-10

Point 2, verbatim

"AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity."

The eight Annex III headings

1. Biometrics.
2. Critical infrastructure.
3. Education and vocational training.
4. Employment, workers' management and access to self-employment.
5. Access to and enjoyment of essential private services and essential public services and benefits.
6. Law enforcement.
7. Migration, asylum and border control management.
8. Administration of justice and democratic processes.

Why it matters here

The operative words are "safety components". An agent that triages and remediates incidents on an ordinary SaaS backend is not a safety component in critical digital infrastructure, so for most teams this category does not reach them at all - and it is also the category that was deferred.

Transparency obligations under Article 50

SECONDARY

Publisher	EU Artificial Intelligence Act
Published	2026-09-30
Retrieved	2026-10-03
Cite key	euartificia2026transpare
URL	https://artificialintelligenceact.eu/article/50/

Assessment. Existing vault key, reused for the same document. No fresh body archived this run: the Article 50 application date and the 2 December 2026 marking date are carried from claims already verified in the vault, and no new reading of this page was needed for them.

Claims drawn from this source. CLAIM-2026-10-03-11

Full text not archived — discovery/headline reference only.

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

PRIMARY

Publisher	EUR-Lex
Published	2026-07-27
Retrieved	2026-10-03
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Assessment. Existing vault key for the same instrument, reused. The deferral dates it carries were verified in earlier editions and are not re-derived here.

Claims drawn from this source. CLAIM-2026-10-03-11

Full text not archived — discovery/headline reference only.