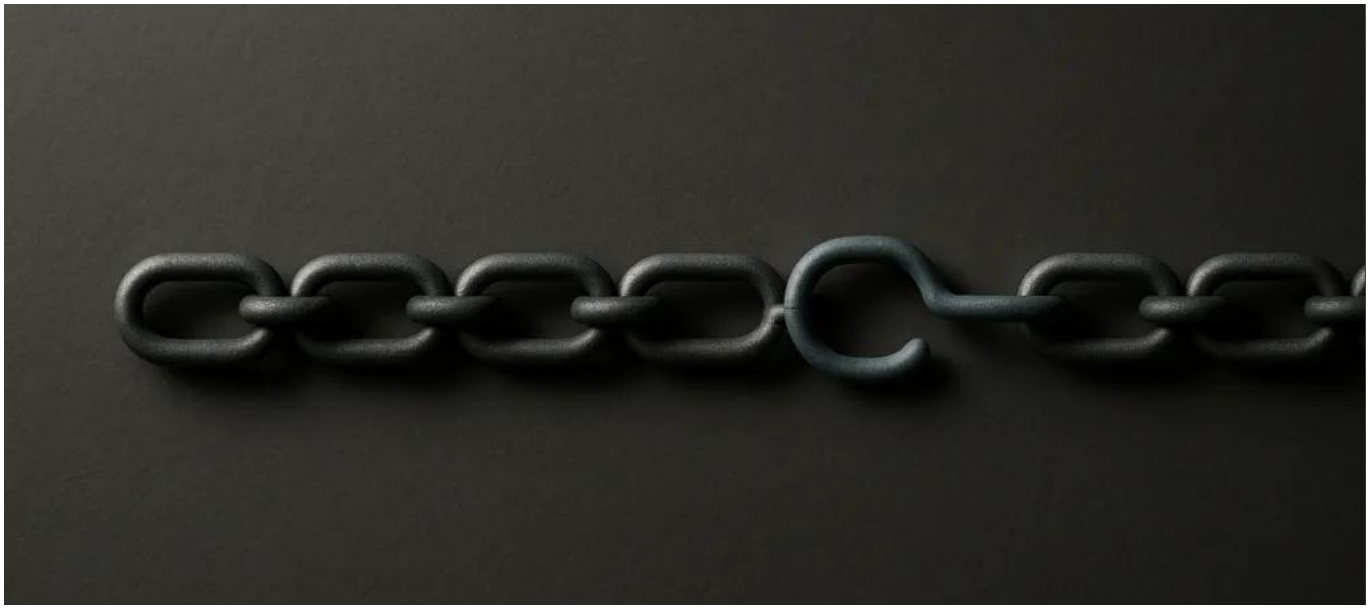

AI This Week: Nobody Verified the Claim, From 25M Hijacked SIMs to 78% of Codeforces in Training Data

Who actually checked? Five stories between September 7 and 12 turned on a claim nobody verified: a platform covering 25 million SIMs, two votes flipping a leaderboard, seven coding agents owned by a config file.

THE SHORT VERSION

This week's stories were about claims being accepted without checking. A threat report with no indicators of compromise. A leaderboard ranking decided by two votes, with 78% of its problem set already in training data. Developers who felt faster while a stopwatch said 19% slower. Seven AI coding agents that all executed code from a malicious .git/config. Meanwhile OpenAI said Astra crosses its own Critical cybersecurity threshold and Brussels sent information requests to 30-plus companies. Both of those are what verification looks like when it finally arrives.

The week started with a security disclosure and ended with a threat report, and the thread running between them is the one I keep pulling on with clients: somewhere in the chain, a claim about capability, provenance or integrity got passed along and nobody re-derived it. An agent trusted a repository's own configuration. A leaderboard trusted votes. A developer trusted a feeling. A defender trusted a vendor's writeup to contain something actionable. Four different systems, one identical failure mode, and the cost of that failure is now measured in tens of millions of SIM cards.

Four stories that share a missing verification step

Anthropic's threat report and the 25M-SIM platform

One subscriber built a surveillance platform covering 25 million SIMs, and the report shipped zero indicators of compromise, which means no defender can check their own logs against it. Full post here.

GitSpawn: 8 flaws, 7 of 7 agents failed

Manifold Security found that a malicious .git/config silently executes code inside AI coding agents, and every agent tested fell over, with 4 still unpatched at disclosure. Full post here.

Two votes, 78% contamination

Two votes were enough to flip the #1 slot on a model leaderboard, and 78% of Codeforces problems turned up in training data, so the ranking measures something other than what buyers think it measures. Full post here.

METR's 39-point self-assessment gap

Developers using AI tools reported feeling meaningfully faster while measurement showed a 19% slowdown, a 39-point gap between perception and stopwatch. Full post here.

The GitSpawn finding is the one I would act on first, because it is already live in most engineering orgs. The mechanism is unglamorous: an AI coding agent clones or opens a repository, the repository carries a crafted .git/config, and code runs. Claude Code, Codex and Cursor are among the affected tools per the reporting around the disclosure. Seven of seven tested agents failed, not five of seven. That uniformity tells you the threat model was never written down: every vendor independently assumed that the repository you point the agent at is a data source rather than an execution surface.

Four of the eight flaws were unpatched at disclosure. If you run agents on CI runners with cloud credentials, or let engineers open untrusted forks in an agentic IDE, you have a window to close this week, and the fix is boring: stop cloning arbitrary repos into an environment that holds anything.

The leaderboard story is the same failure at the procurement layer. Two votes flipping the top slot means the ranking's resolution is worse than its ordering implies, and 78% contamination on Codeforces means a coding score is partly a memorisation score. I have sat in enough vendor selection meetings to know how a leaderboard screenshot works in the room: it ends discussion. What it should start is a question about whether anyone has run the model against tasks from your own backlog, tasks that did not exist when the model was trained. That is a week of work for a platform engineer and it is the cheapest week you will spend.

METR's number is the uncomfortable one because it implicates the people rather than the tooling. A 39-point gap between felt speed and measured speed is a systematic bias, not a rounding error in self-reporting. Every productivity claim built on developer surveys, including the ones inside your own org, carries an unknown correction factor. My working

assumption since reading that study: **treat any AI productivity figure that came from asking people how it went as a mood measurement.** Useful for adoption, useless for ROI.

The artist-income survey belongs in this set for a different reason. Ninety percent of surveyed artists report lost income from AI despite copyright rulings going their way. A legal win is a claim about how the world should work. It is not a measurement of how the market actually settled. The gap between the two is the gap METR found, transplanted into policy.

OpenAI's Critical label meets Brussels' first information requests

> **The two things that happened elsewhere**
> On September 1, OpenAI published Path to Astra, stating that Astra is its first model to meet the Critical cybersecurity capability threshold under its Preparedness Framework, and began shipping it to a limited set of business customers on September 3. The same week, the European Commission sent information requests to more than 30 AI companies in its first AI Act enforcement actions, covering safety and security of the most advanced models plus copyright and transparency.

OpenAI's own safety writeup says that with the right tools and access, Astra "can find previously unknown security flaws and develop new ways to exploit them across many well-protected systems without a person guiding each step." That is a vendor self-assessment, published by the vendor, against a framework the vendor wrote. CSO Online notes the Critical classification is one the company said triggers additional deployment restrictions. I take the disclosure at face value as an act of disclosure, and I have no way to check the underlying evaluation, which is precisely the structure of every other story this week.

Brussels is the counterweight. The AI Office's enforcement powers went live on 2 August 2026 and, per the Commission's own FAQ, include information requests, model access for evaluations, mandated risk mitigation, fines up to 3% of global turnover, and withdrawal or recall of a model. Commission spokesman Thomas Regnier described the first round as covering two broad areas: safety and security of general-purpose and most advanced models, and copyright and transparency.

Model access for evaluations is the clause that matters here. Everything else on that list is paperwork with teeth. That one is a third party being allowed to check the claim.

On the infrastructure side, SGLang 0.5.19 shipped beam search and a faster cold start, with specific regressions still open in practice. I mention it because it is the honest version of a release note: capability added, known problems listed. Compare that to a threat report with zero IoCs.

My take: internal evaluation stops being enough in Europe first

The shift I find interesting this week is that two different actors, one a lab and one a regulator, started treating self-reported claims as claims rather than facts. My reading: within the next few quarters, "we evaluated it internally" stops being sufficient for enterprise procurement in Europe, and the practical burden lands on buyers to run their own evaluations on data that was never in training. I expect vendors to resist third-party eval access far harder than they resisted transparency reporting, because transparency reports are written and evals are run. Unconfirmed, but it is where I would place my bet: the first AI Act fine will be about documentation rather than capability, because documentation gaps are provable and capability claims are not.

If you want one action from this week, it is not a policy document. Take the three agentic workflows in your org that touch untrusted input, write down what each one is allowed to execute, and check whether anything actually enforces that. **The GitSpawn result suggests the answer for most teams is no.**

WHAT TO DO WITH THIS

Every AI claim in your stack, whether capability, benchmark or provenance, is unverified until you or a neutral party has measured it yourself, and this week gave four separate proofs of what that costs. If you want a second pair of eyes on where your agents execute untrusted code, or on how you would evaluate a model without a leaderboard, that is the kind of work I do. Book a call →

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →