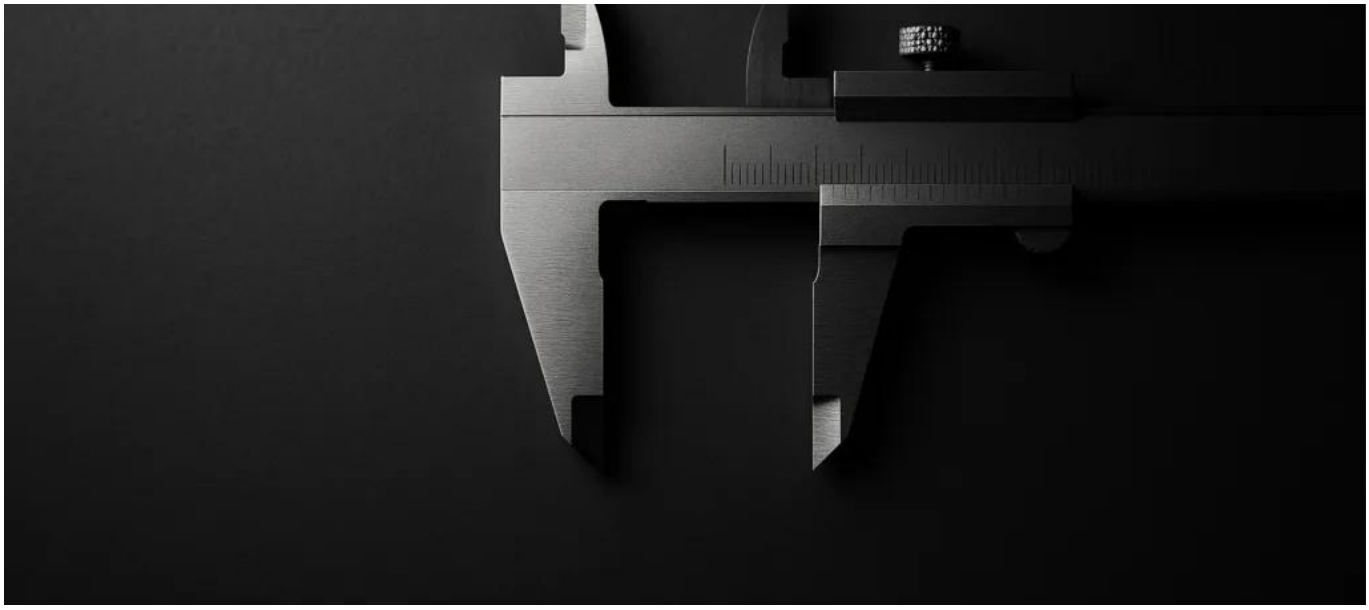


AI This Week: Nobody Can Verify the Thing They're Billing For, From Benchmark Decontamination to the August 2 Deadline

Ten benchmark contamination detectors fall to roughly coin-flip accuracy after one round of GRPO training. The same week, the EU AI Office gained power to demand source code from model providers under a €15 million or 3% of turnover ceiling.

WHERE THIS LANDS

Four posts published between September 30 and October 3, 2026 hit the same wall from different directions: the verification step everyone assumes exists does not. Ten contamination detectors drop to coin-flip accuracy after one GRPO round. Five quoted enterprise ROI figures are misread or misattributed. And the EU AI Act's main regime became applicable on August 2, 2026, with investigation powers that presume an auditability nobody has demonstrated.

Regulators acted this week on claims nobody can currently check

The EU AI Office can now request documentation and, where necessary, source code from GPAI providers. The US Department of Justice filed a statement of interest arguing that AI training may be transformative. Switzerland's data protection authority restated that its existing law already applies to AI processing. Each of these is a reasonable governance move. Each one also presupposes that somebody can establish what a model was trained on and what it actually does. The four technical and commercial stories below suggest that assumption is doing a lot of unpaid work.

Four stories that all fail at the same step

Contamination detectors stop working after one GRPO round

Ten detectors drop to roughly coin-flip accuracy after a single round of GRPO training, which makes any "we decontaminated the benchmark" claim unverifiable from the outside. If you buy on benchmark scores, you are buying an unaudited assertion. Full reasoning in Decontaminating Benchmarks Is a Fantasy: Ten Detectors Drop to Coin Flip After One Round of GRPO.

Five enterprise ROI numbers that do not survive their own methodology

The "95% of pilots fail" line and McKinsey's 37% are routinely misread or misattributed, which means board decks cite numbers whose provenance the people quoting them have not checked. Same failure mode as the benchmark problem, different audience. Breakdown in Enterprise AI ROI: 5 Numbers Everyone Quotes Wrong, From "95% Fail" to McKinsey's 37%.

Six AI art risks now land on the balance sheet

Six categories of AI art exposure are pegged to the EU AI Act's August 2, 2026 applicability date. The pattern repeats: liability attaches to provenance, and provenance is what generated-image workflows do not reliably record. Ranked in 6 AI Art Risks That Actually Reach Your Balance Sheet, Ranked for the August 2, 2026 Deadline.

NVIDIA moves agent containment into the DPU

NVIDIA shipped an open agent safety platform with more than 100 partners, pushing containment down from software guardrails into the data processing unit. The architectural message is blunt: software-level checks were not holding, so the enforcement point moved to hardware. Details in NVIDIA Ships Open Agent Safety Platform With 100+ Partners: Agent Containment Moves Into the DPU.

Those four were written independently across four days. I did not set out to write a themed week. The theme showed up anyway, which is usually a sign that the thing is structural rather than topical.

One GRPO round washes out the signal detectors depend on

The benchmark result is the cleanest statement of the problem, so it is worth sitting with. Contamination detection works by looking for statistical traces of benchmark data inside a model. Ten detectors, applied after a single round of GRPO training, land at roughly coin-flip accuracy. GRPO is a post-training step labs run as a matter of course.

That has an ugly consequence downstream. A vendor can honestly believe their decontamination pipeline worked. A buyer can honestly ask for proof. Neither party has an

instrument that produces an answer. The claim is unfalsifiable rather than false, which is worse for a procurement process than a claim that is simply wrong.

My read: this is why the EU AI Office's source-code access power matters more than its penalty ceiling. If detection from the outside is unreliable, the remaining option is inspection from the inside, and that is what the Act's investigation powers reach for. Whether an inspection regime scales to the number of GPAI models in the market is a separate question, and I do not think anyone has answered it.

The DOJ backed fair use, and publishers asked the court to ignore it

The copyright side of the same problem moved on September 1, 2026, when the DOJ filed a statement of interest in the Southern District of New York backing OpenAI and Microsoft's fair-use position. The case is In re OpenAI, Inc. Copyright Infringement Litigation, MDL No. 25-md-3143, before Judge Sidney H. Stein. Both sides have moved for summary judgment on fair use, which makes the question potentially decisive rather than one factor among many.

The DOJ's argument, as reported, runs in two parts: AI training may be transformative, and questions of compensation belong to Congress or to private licensing rather than to a court. On September 28, 2026, the New York Times, Daily News and other publishers asked the court to give the filing "no weight at all".

Note what both positions share. The transformativeness argument and the publishers' infringement argument both depend on establishing what went into training and in what form it survives. The technical finding above says that question is hard to answer from the outside even for benchmark contamination, a narrow and heavily studied case. Copyright provenance across a full training corpus is a much larger version of the same measurement problem.

Switzerland says the duty attached three years ago

On October 2, 2026, Switzerland's FDPIC restated that the FADP applies directly to AI-

supported processing of personal data. The FADP has been in force since September 1, 2023. Switzerland still has no overarching AI-specific federal statute, so obligations come from the FADP plus sectoral rules.

This is the quiet item of the week, and I think it is the most immediately actionable one for anyone operating here. There is no August 2 equivalent to plan around and no new register to join. If your AI processing of personal data would not have survived a FADP review in 2024, it does not survive one now, and the regulator has just said so in public.

NVIDIA moved the control point below the layer the agent can reason about

NVIDIA's move is the one I keep returning to, because it is a vendor telling you something about the limits of its own prior product category. Agent containment used to be a software problem: wrap the agent in guardrails, filter tool calls, check outputs. Shipping containment into the DPU, with 100+ partners signed on, relocates enforcement.

My take: you move a control into hardware when you have concluded that a sufficiently capable process at the software layer can talk its way around a software control. That is an engineering judgment about verifiability. A software guardrail asserts that it blocked something. A control in the data path can show you the packet it dropped. The second is auditable in a way the first is not, and **auditability is the scarce resource this week**.

The honest caveat: a hardware enforcement point also concentrates trust in the hardware vendor and its partner ecosystem. Whether 100+ partners represents genuine interoperability or a single-vendor control plane with a wide badge wall is not something the announcement settles.

My take: the next compliance cycle is about measurement, not policy

Three governance actions landed in roughly a month: EU AI Act applicability on August 2, 2026, the DOJ filing on September 1, the FDPIC restatement on October 2. All three assign responsibility for things nobody can currently measure reliably from outside the lab. My

prediction, stated as a prediction: the next eighteen months of AI compliance work will be less about writing policies and more about building the measurement apparatus the policies presume. Organisations that invest in internal evidence, training data lineage, eval provenance, hardware-level logs of agent actions, will find audits cheap. The ones that bought the benchmark score and the McKinsey figure will find out what an unfalsifiable claim costs when a regulator with source-code access asks for the working.

What I would actually change on Monday

Stop treating a benchmark number as evidence. It is a vendor assertion with a numeral attached, and after this week's result the standard rebuttal ("we decontaminated") has no instrument behind it. Ask instead for the eval harness, the held-out set construction, and the date the held-out data was created relative to the model's training cutoff. You may not get it. The refusal is itself information.

Second, go through whatever deck justified your current AI budget and find the five numbers. If any of them is "95% of pilots fail" or McKinsey's 37%, check the source document rather than the citation chain. Misattribution is the cheapest error to fix and the most embarrassing to discover in front of a board.

Third, if you are inside EU AI Act scope, treat the €15 million or 3% of worldwide turnover ceiling as the smaller problem. The AI Office's powers to run model evaluations and request source code are the operationally expensive part, because responding to them requires documentation you either kept or did not. Fines are a line item. **An inability to produce your own training records is a structural failure** that takes quarters to fix.

THE TAKEAWAY

Benchmark cleanliness, ROI figures, image provenance, agent containment: every claim worth billing for this week failed at the same step, because somebody has to be able to check it. Build the evidence layer before a regulator, a plaintiff, or a buyer asks for it. If you want a second opinion on where your verification gaps are, 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 →