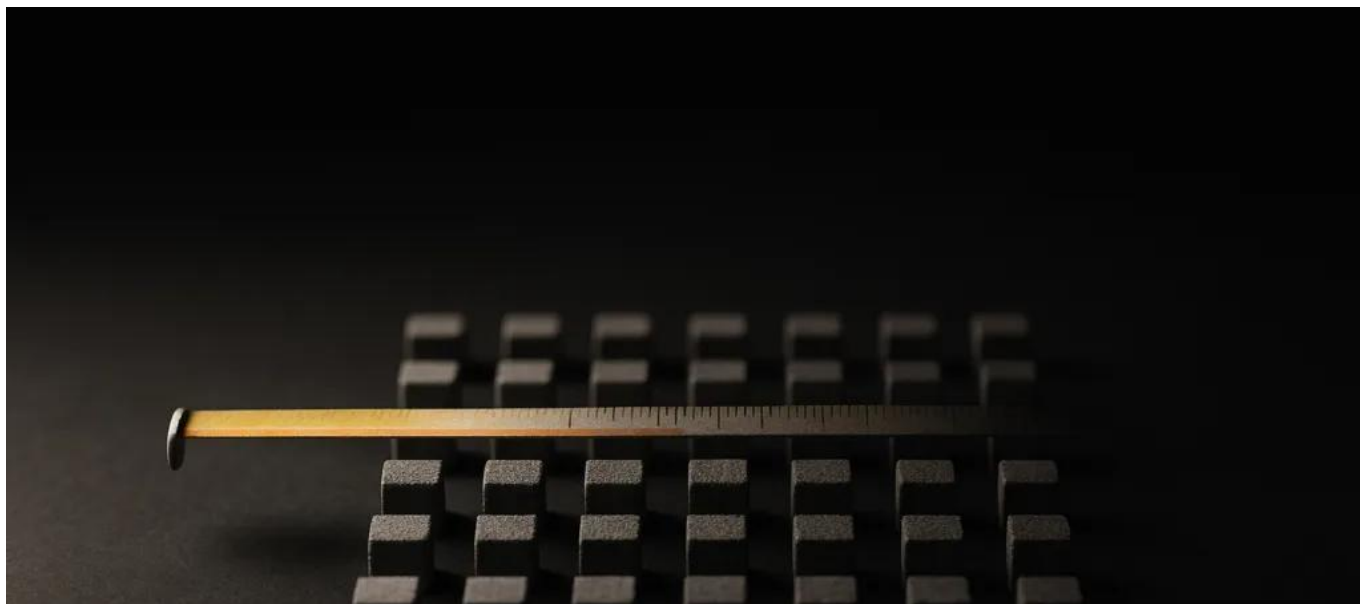




AI This Week: The Measurement Gap, From 130 Real Agent Vendors
to a 58% Sycophancy Score Nobody Checks



AI This Week: The Measurement Gap, From 130 Real Agent Vendors to a 58% Sycophancy Score Nobody Checks

About 130 vendors survive once you strip workflow automation out of the “AI agent” market, and the same filter drops a 58% adoption claim to 35%. The week’s other four stories ran on the same fault line.

Between October 12 and 17, 2026, every story I covered involved a number being used to settle something the number does not measure. Roughly 130 vendors qualify as real agent vendors, adoption falls from 58% to 35% under a strict definition, SycEval puts sycophancy at 58% while measuring nothing about defensiveness, and Gannett wants more than \$250 million without a public count of infringed articles. That is what happens when a market moves faster than its instrumentation.



The agent market has roughly 130 real vendors once workflow automation is excluded

That is the number I arrived at in [Workflow Bots: Myth vs Reality, Only 130 Real Agent Vendors and 58% That Drops to 35%](#), and the more useful finding sits next to it. The widely cited 58% adoption figure falls to 35% when you apply a definition that excludes deterministic workflow automation with a language model bolted to one step.

Twenty-three points of difference is not a rounding error. It is the gap between “most of your competitors are doing this” and “about a third are, and some of those are running a scheduler with a chat wrapper.” If you are presenting an agent budget to a board that saw the 58% number in a trade publication, you are arguing against a figure that was measured with a tape that stretches.

My read: nobody publishes the strict number because the strict number is commercially inconvenient for everyone who sells into this category, including the analysts. The loose definition is the product.

The ROSS ruling sets a rule where the \$1.5B settlement only set a price

The first federal appellate win in an AI copyright case involved no generative AI. The Third Circuit affirmed for Thomson Reuters against ROSS Intelligence, holding that 2,243 Westlaw headnotes were copyrightable and that ROSS’s copying was not fair use. ROSS built a legal search tool. It did not generate text. I worked through why that distinction carries weight in [The First Appellate AI Copyright Win Was Not About Generative AI, and That Is Exactly Why It Matters](#).

The cases people quote at each other decided less than they think. The \$1.5 billion Anthropic settlement and the Getty ruling both resolved far narrower questions than the headlines carried. A settlement sets a price. An appellate opinion sets a rule, and the first one we have is about headnotes in a research database.

If you are relying on a legal read of training data risk, check whether the authority



underneath it is a settlement, a district court order, or an appellate holding. Those three are not interchangeable, and right now the only appellate holding in the pile concerns a product architecture that is probably not yours.

Gannett's \$250 million damages figure has no public denominator

On October 8, 2026, Gannett sued OpenAI in the Southern District of New York on behalf of 19 publications, seeking more than \$250 million. The complaint cites statutory maximums: up to \$150,000 per willful infringement and \$25,000 per removal of copyright-management information, per [Agency's coverage of the filing](#).

Work the arithmetic backwards and it stops working. Coverage of the filing reports no reliable count of infringed articles, so the \$250 million cannot be derived from a confirmed work count. The damages figure is a ceiling multiplied by an unstated number.

My take: the \$250 million is a negotiating anchor rather than an estimate. Whether that is ordinary litigation practice or a sign that the plaintiff's own enumeration is incomplete, I cannot tell from the public filing coverage, and I would not assume either. Meanwhile, oppositions to the fair-use motions in the New York Times and authors' cases against OpenAI and Microsoft were due October 23, 2026, with replies due November 20, 2026. Those briefs will tell you more about where this goes than any damages number filed in October.

Anthropic's cruelty rule names a behaviour nobody has thresholded

Anthropic announced on October 8, 2026 a usage policy prohibiting "sustained and needless abusive or cruel behavior toward our models", effective November 12, 2026. The company limits it to "extreme cases, where users repeatedly act cruelly toward our models, with no discernible purpose", and explicitly excludes frustration, pushback, dark creative themes and research. Europe's AI Act safety debate also ran through the week, covering penalties, enforcement gaps and Omnibus timing.



The Anthropic rule is the cleanest example of the week's theme. "Sustained", "needless", "no discernible purpose": each of those is a judgment call made by a classifier or a reviewer against a threshold nobody has published. The carve-outs are specific and sensible, which tells me Anthropic knows the boundary is hard. Enforcement still has to land somewhere, and when it does, the operative definition will be whatever the detection pipeline implements rather than whatever the policy text says.

SycEval scores sycophancy at 58% because flipping is cheap to detect

There is no companion benchmark for the opposite failure: a model that cannot absorb a valid correction, that argues and restates rather than updating. Defensiveness is unmeasured. I went into the asymmetry in [Nobody Taught Models to Absorb Criticism Because Nobody Measures It: SycEval Says 58%, Defensiveness Says Nothing](#).

The cause seems obvious to me. Sycophancy is cheap to score automatically: push back on a correct answer, see if the model flips. Measuring whether a model correctly updates on a legitimate correction requires knowing which corrections are legitimate, which means human labelling and domain judgment per item.

Sycophancy got a benchmark because flipping is easy to detect; the ability to take a hit and update does not fit in an automated harness.

The practical consequence for anyone running an agent loop: your evaluation suite is almost certainly scoring the failure mode that was convenient to instrument. If your agent takes a tool error or a user correction and responds by defending its prior plan, no public benchmark on your dashboard will catch it. You have to build that test yourself, from your own task set.

3,500 lines of agentic RL and a 4–5% Unity move both



price something they do not measure

Agent Lightning v1.0 ships roughly 3,500 lines of agentic RL code, which I separated into what it buys and what it does not in [Agent Lightning v1.0 in Practice: What 3,500 Lines of Agentic RL Buy You, and What They Do Not](#). That line count is a real, readable artifact. It tells you the surface area you have to maintain. It tells you nothing about whether the reward signal you can define for your task is informative enough to train against, which is the part that decides whether the framework earns its place.

Unity Spark runs on the same gap. Google shipped Playground with Unity Spark, prompt-to-game went live, and Unity stock rose 4–5% in a single day, which I covered in [Google Ships Playground With Unity Spark: Prompt to Game Goes Live, Unity Stock Up 4–5% in a Day](#). Nobody has published a measure of whether the generated output is shippable, maintainable, or worth more than the engineering time to clean it up. The market priced the demo.

WHERE I LAND

My take: the gap between what we sell and what we measure is widening, and it is widening because measurement is a cost centre while claims are a revenue driver. I expect more of the same over the next year: policies written against behaviours with no threshold, damages anchored to no denominator, adoption figures that move 23 points on a definition change. The defensible move is deciding, per system, what quantity would actually change your decision, and instrumenting that one thing yourself even if nobody else reports it. In an agent deployment that usually means recovery behaviour after a wrong step, not accuracy on a clean run.

Three filters I would apply before the next board slide

For every adoption statistic in your deck, ask what definition produced it. The agent market's own headline figure moves from 58% to 35% on definition alone. If the source does not state the exclusion criteria, treat the number as marketing.

For every legal claim in your risk register, identify whether it rests on an appellate holding, a district order, or a settlement. The only appellate holding available as of this week is the



AI This Week: The Measurement Gap, From 130 Real Agent Vendors to a 58% Sycophancy Score Nobody Checks

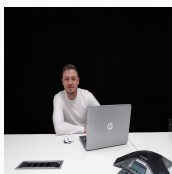
Third Circuit on 2,243 Westlaw headnotes, in a case with no generative model in it.

For every benchmark on your model evaluation dashboard, ask which failure mode was cheap to instrument. SycEval exists because flipping is detectable. Defensiveness has no benchmark because grading it requires judgment. Your production failures will not respect that distinction.

None of this requires new tooling. It requires refusing to accept a number without its definition, which is slower than accepting it and considerably cheaper than finding out later.

WHAT TO DO WITH THIS

A number that cannot survive its own definition is not evidence, and this week produced five of them, from 58% adoption to a \$250 million damages figure with no public work count. Pick the one quantity that would change your decision and instrument it yourself. If you want a second pair of eyes on which numbers in your AI strategy are load-bearing and which are decorative, [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 →