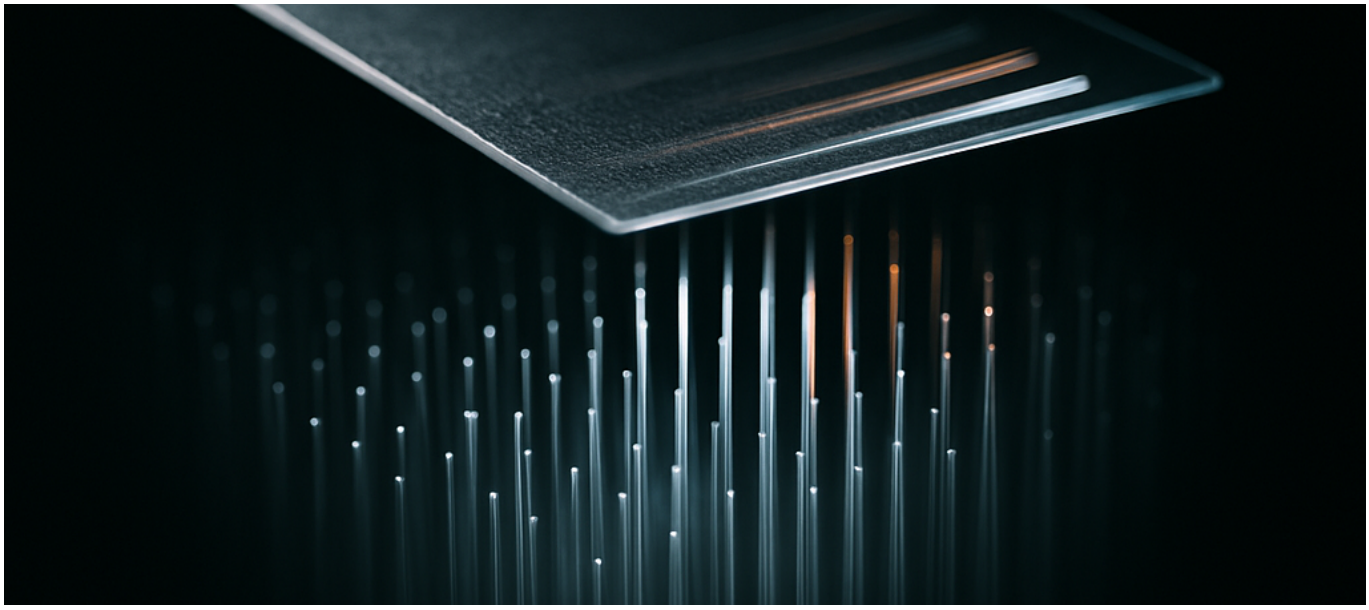

# AI This Week: The Verification Layer — 5 Stories Where AI's Weak Point Was Knowing What Was Real

Two startups raised $80.5M this week to do the least glamorous job in enterprise software: watch what employees actually do all day. The reason: enterprise agents keep automating workflows that only exist in a slide deck.

## The week in one paragraph

Every story I covered between August 26 and 28 had the same shape underneath it: an AI system acting confidently on something it could not verify. Models fabricating once their context window gets long enough. Agents automating a process nobody actually follows. A chatbot leaking histories eleven weeks after someone told it how. And on the other side of the ledger, two funded companies whose entire product is *look first, then automate* — plus AT&T proving in production that if you measure quality instead of assuming it, you can cut

AI spend by more than half. The connecting thread isn't capability. The gap is the verification layer: the boring, unfunded, unloved part of the stack that checks whether what the model believes matches what's true.

## 1. Every LLM fabricates above 10% at 200K context

A 172-billion-token evaluation found that **every tested model crosses a 10% fabrication rate once context reaches 200K tokens**. Not the worst model. Every one.

The important word is *every*. This isn't a ranking exercise where you pick the safest vendor and move on — it's a property of the current architecture class at long context, and it shows up at exactly the context lengths vendors have spent two years marketing as the headline feature.

Why this matters operationally: long context is the default answer to almost every RAG problem. Retrieval is hard, so teams stop retrieving and start stuffing — dump the whole contract, the whole codebase, the whole ticket history into the window and let the model sort it out. This data says that strategy degrades precisely where you stopped supervising it.

> If your architecture's answer to "how do we ground this?" is "bigger context window," you didn't solve grounding. You moved the failure downstream where nobody's looking.

**My take:** I expect the practical response to be a return to aggressive retrieval and chunking — not because it's elegant, but because a 40K-token window you can audit beats a 200K-token window you can't. The teams that get burned first will be the ones that ripped out their retrieval layer as a cost optimization.

[Read the full breakdown of the 172-billion-token evaluation here.]

## 2. Skan AI raises $63M because agents are automating processes that don't exist

Skan AI raised $63M to map how work actually happens, and the stated reason is blunt: enterprise agents were found automating processes that **do not exist in reality**.

Sit with that. The agent works. The workflow it's executing is fiction — the documented process, not the real one. Somebody wrote a runbook, three teams quietly stopped following it, and the agent got pointed at the runbook.

This is model fabrication one layer up the stack. The model hallucinates a citation; the agent hallucinates a business process. Both are confident. Both are unverified. The difference is that the second one has write access.

> Process documentation is the enterprise equivalent of a hallucinated citation: authoritative-looking, widely referenced, and frequently made up.

[The full story on Skan AI's raise and the process-fiction problem is here.]

## 3. Sola: $17.5M for the same insight, different angle

Sola raised a $17.5M Series A from Andreessen Horowitz after **5× revenue growth**, using screen-recording capture instead of traditional RPA scripting.

Classic RPA asks a human to describe the process, then encodes the description. That encoding step is where the fiction enters. Screen recording skips the description entirely — the ground truth is what the employee's screen did, not what they said they do.

Two companies, $80.5M combined, both selling observation-before-automation. When two funded bets converge on the same unglamorous premise in the same week, that's a signal about where the bottleneck actually sits.

[More on Sola's screen-recording approach and the 5× growth behind the round.]

## 4. Grok still leaking chat histories 11 weeks after disclosure

Grok was still leaking full chat histories **eleven weeks after disclosure**, with AES-256 encrypted prompts defeating guardrails **40% of the time**.

The technique is the tell. Wrap a prompt in AES-256 encryption and the guardrail fails to recognize it as an attack four times in ten. The guardrail is pattern matching on the surface form of the input, not reasoning about intent — change the surface form and it stops seeing anything.

Eleven weeks is the number I'd focus on if I ran security at a company shipping an LLM feature. That's not a patch-cycle delay; that's a disclosure that sat in a queue. It suggests the fix isn't a one-line filter update, which tracks — you can't pattern-match your way out of an unbounded input space.

> A guardrail that inspects the surface form of a prompt isn't a security control. It's a spam filter with better PR.

**My take:** I read the 40% bypass rate as evidence that prompt-level filtering has a ceiling, and vendors are close to it. Unconfirmed, but I'd expect the next serious defenses to move to output inspection and capability restriction rather than input classification — check what the model is about to reveal, not what the user typed.

Full analysis of the Grok leak and the encrypted-prompt bypass.

## 5. AT&T: what verification looks like when you actually do it

This is the counterexample, and the most useful story of the week. AT&T cut costs on coding and advanced AI tasks by up to 56% with only a **~2% quality degradation**, by routing routine work to cheaper open-weight models through LiteLLM-based routers.

The scale makes it credible. Their internal assistant, Ask AT&T, serves roughly 100,000 employees and handles about 45 billion tokens per day. Around 40% of employee queries already route to open models, with an internal target of 60–70%. VP for employee-facing AI Mark Austin said the aim is to keep spend on Anthropic and OpenAI flat while shifting workloads to open-weight models hosted on AT&T's own NVIDIA and AMD servers.

Every other story this week is about a system that couldn't verify something. This one is about a system that could. AT&T didn't guess that open models were good enough for routine coding — they measured a ~2% delta and made a routing decision against it.

> 56% cheaper for 2% worse is only a good trade if you measured the 2%. Otherwise it's a rumor with a spreadsheet attached.

Routing is a verification problem disguised as a cost problem. You need a working definition of "good enough for this task class" before a router means anything, and that definition has to come from evaluation, not vibes.

**My take:** the flat-frontier-spend target is the strategically interesting bit. AT&T isn't trying to leave the frontier labs — they're trying to stop the bill from growing while volume does. I expect that to become the standard enterprise posture within a year: frontier models as a fixed-cost premium tier, open weights absorbing all volume growth.

More coverage of the AT&T routing setup here.

## 6. Clerq: weeks to ten minutes, and the verification question underneath

NLPatent rebranded to Clerq, with agentic AI compressing patent research from weeks to roughly **10 minutes**.

Patent prior-art search is close to an ideal agentic task: the corpus is bounded, public and structured, and the output is checkable. A cited patent either says what the agent claims it says or it doesn't.

Contrast that with the fabrication finding above. Long-context summarization of an unstructured pile fails silently — nobody knows which sentence was invented. Patent search fails loudly, because every claim carries a document you can open. **The tasks where agents work best are the tasks where being wrong is cheap to detect.**

Full write-up on the Clerq rebrand and the 10-minute claim.

## What else happened

**The 19% that never break even.** Across hundreds of enterprise agent deployments, the median payback from go-live to positive ROI is 5.1 months — and 19% never reach positive ROI at any measured point. Gartner's Agentic AI Pulse 2026 names the dominant failure mode behind that cohort: *evaluation drift*, where agents are scored on narrow success metrics that miss downstream rework and oversight cost. That's the verification problem stated in financial terms. The agent hits its KPI; the KPI wasn't measuring the thing that costs money.

**Open weights kept compounding.** Google's Gemma family passed one billion downloads with more than 100,000 community variants. That's the supply side of the AT&T story — routing to open models only works if the open models are credible, and a billion downloads with six figures of derivatives is a fairly direct measure of credibility.

**Stability AI raised $76M.** A Series B on August 25 brought total funding to $232M, with Universal Music Group, Sony Music Group, Warner Music Group, Electronic Arts, AMD Ventures and Pacific Alliance Ventures participating. Three major labels on one cap table is the notable detail — the rights holders are buying in rather than only litigating.

## The pattern

Every failure this week is the same failure at a different altitude. The model can't verify its own recall at 200K tokens. The agent can't verify that the process it automates exists. The guardrail can't verify that an encrypted prompt is an attack. The ROI dashboard can't verify that the metric it's optimizing corresponds to money. And the two companies that raised $80.5M raised it to sell verification of the one thing enterprises assumed they already knew: what their people actually do.

**My take:** the 19% never-break-even cohort isn't a story about bad models. It's a story about deployments where nobody built the layer that checks the work — and evaluation drift is just the name for what happens when the measurement is cheaper to fake than the outcome.

AT&T is the shape of the correct answer, and it isn't exciting. Instrument the workload. Measure quality per task class. Route on the measurement. Accept a known 2% loss rather than an unknown one. Nothing in that sequence requires a frontier model; all of it requires discipline.

I expect the next wave of enterprise AI budget to shift noticeably toward observability, evaluation and process mining — not because it's fashionable, but because a 5.1-month median payback leaves very little room for a deployment built on assumptions. The two raises this week are early evidence. Unconfirmed, but I'd bet the pattern holds through Q4.

**The bottleneck in enterprise AI is no longer what the model can do — it's whether anyone can prove what it did was real.**

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 