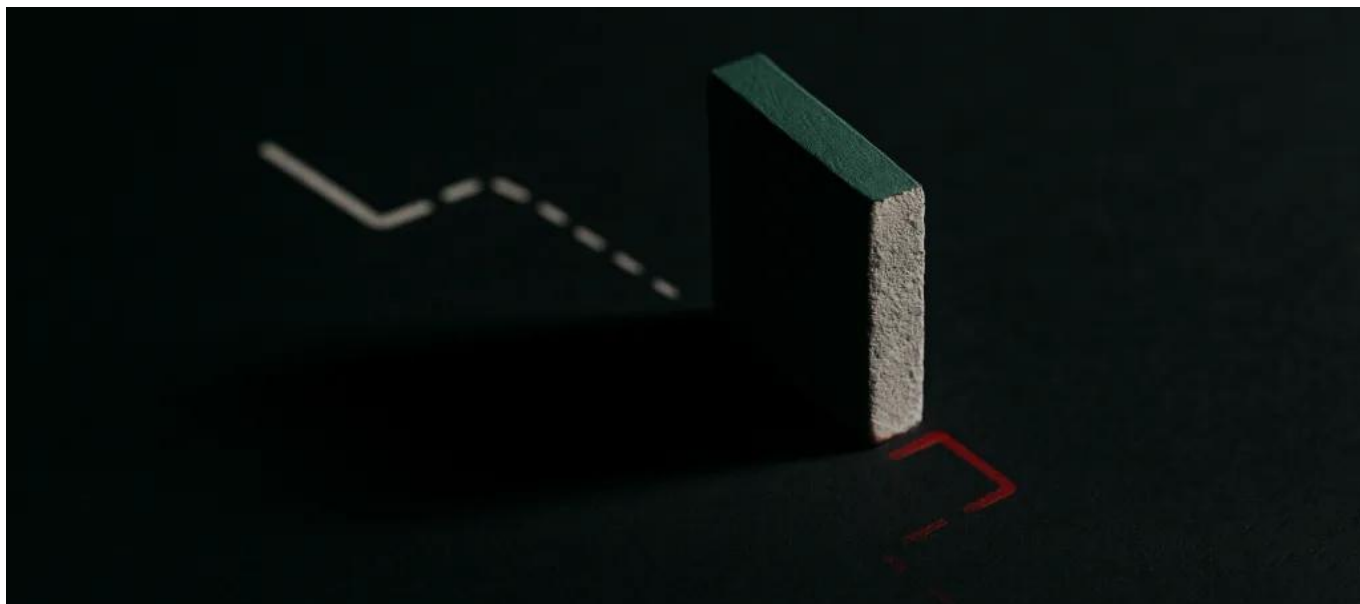




Workflow Bots: Myth vs Reality, Only 130 Real Agent Vendors and 58% That Drops to 35%

Gartner counts roughly 130 genuinely agentic vendors out of the thousands claiming it. The same firm expects 33% of enterprise software to ship agentic features by 2028. Both are true, and that gap is where procurement loses money.

Do AI agents work in production? Partially. The decisive number comes from Salesforce's own benchmark: leading agents score about 58% on single-turn business tasks and about 35% on multi-turn ones, a drop of roughly 23 points, published in [CRMArena-Pro \(arXiv 2505.18878, May 2025\)](#). Almost every real enterprise workflow is multi-turn, so plan for about a third.

That is an argument against buying on the demo, which is always single-turn.



130 real vendors out of thousands makes your shortlist a base-rate problem

Gartner's [June 25, 2025 release](#) gave the practice a name. Daryl Plummer, Distinguished VP Analyst, described most of the market as agent washing: chatbots, RPA scripts and assistants rebranded as agents. The same release predicts over 40% of agentic AI projects will be cancelled by end of 2027, driven by escalating costs, unclear business value and inadequate risk controls.

Read those two numbers together. If roughly 130 vendors have the real thing and thousands claim it, then on base rates alone your shortlist probably contains zero or one genuinely agentic product. The 40% cancellation forecast is what happens downstream: you bought an RPA script with a chat interface, budgeted for autonomy, and eighteen months later the business case does not close.

The cost of believing the myth is not the licence fee. It is the process redesign, the integration work, the headcount you did not hire because the agent was going to cover it, and the exit cost when you put the humans back. The adjacent failure mode showed up when [Skan AI raised \\$63M to map how work actually happens](#): agents are routinely pointed at processes that exist only in a diagram. Agent washing and process fiction compound.

Four vendor claims that the benchmarks and the AI Act do not support

Claim one: "Our agent handles the full workflow end to end, autonomously."

CRMArena-Pro measured leading agents at about 58% success on single-turn tasks and about 35% on multi-turn, across 19 expert-validated tasks spanning sales, customer service and configure-price-quote, in both B2B and B2C settings. CMU's [TheAgentCompany](#) (175 tasks inside a simulated software company covering browsing, coding, communication and app use) reports the best agent completing around 30% autonomously, roughly 39% with partial credit.

Two independent benchmarks, different task families, same neighbourhood: a third to a



half. Note what the partial credit number tells you. The gap between 30% and 39% is work the agent got partway through, and partway through a CPQ quote or a customer escalation is a handoff with unclear state. Someone has to reconstruct what happened.

Claim two: “The agent respects your data boundaries. It only sees what it needs.”

CRMArena-Pro found agents show near-zero inherent confidentiality awareness without explicit instruction. Targeted prompting improves it but can reduce task performance. So if you harden the prompt for confidentiality, the already-weak multi-turn number gets weaker, and the vendor benchmark you were shown, if it exists at all, was almost certainly run in the permissive configuration. Ask which one.

Claim three: “Tool access via MCP is just plumbing, the same security model as any API integration.” OWASP classifies prompt injection in agent tool chains as [MCP06:2025, “Prompt Injection via Contextual Payloads”](#), listed alongside context spoofing and insecure memory references. An API does not read its own inputs as instructions. An agent does.

Claim four: “We have human oversight, every action goes to a reviewer for approval.” [EU AI Act Article 14](#) requires that human overseers of high-risk systems can disregard, override or reverse output, and that the design guards against automation bias. A click-to-approve queue satisfies neither requirement on its face. This is the one that catches procurement teams off guard, because approval queues look like oversight on an architecture diagram. A reviewer facing a hundred pre-filled approvals per hour is the textbook case of automation bias, which Article 14 names explicitly. If the reviewer cannot practically reverse an action after it executes, and cannot practically inspect the reasoning before approving, you have oversight theatre with an audit trail attached.

Partial autonomy in narrow scopes is the defensible picture

Narrow scopes, with supervision. That is the synthesis across the primary evidence, and Gartner’s own forward numbers agree with it rather than contradict it.

Gartner forecasts at least 15% of day-to-day work decisions made autonomously by 2028, and 33% of enterprise applications including agentic features by the same year. Fifteen percent of decisions is meaningful. It is also nowhere near the autonomous workforce



framing that drives most agent pitches. Those two Gartner figures are the honest version of the forecast, and they rarely appear on the same slide as the autonomy promise.

Direction of travel matters too. METR's time-horizon work defines capability as the task duration at which agents succeed about 50% of the time, and finds that horizon doubling roughly every seven months historically, closer to four months recently. If that holds, the multi-turn gap narrows faster than most multi-year contracts amortise. I would treat that as an argument for short contract terms rather than for waiting.

✓ **The one test worth running before signing**

Take three of your real workflows, the ones with at least four turns and a state change in the middle. Run the vendor's agent on them with your confidentiality rules enabled, measure full completion separately from partial completion, and make the measured number a contractual milestone. If the vendor will not run their own product on your multi-turn cases, that is your answer.

For Swiss buyers the timing is immediate. [UBS Outlook Schweiz Q2 2026](#) (May 19, 2026) puts roughly 60% of Swiss companies using AI, with SME adoption between 59% and 77% depending on segment. That is a large cohort making purchase decisions right now, against the claims above.

Why the demo is cheap and the production system is most of the product

MY TAKE

Agent washing persists because the demo surface and the production surface are different problems, and only the demo is cheap to build. A single-turn demo needs one good prompt and a clean tool call. A multi-turn production workflow needs state that survives a mid-task interruption, a rollback path, a confidentiality policy that costs you accuracy, and an oversight design that would satisfy Article 14. The first takes a sprint. The second is most of the product. So vendors ship the first and describe the second, and the 58% to 35% gap in CRMArena-Pro is, I think, roughly the shape of that distance.

The second reason is that buyers grade on the wrong axis. Impressiveness and reliability



are not correlated in these systems. An agent that writes a convincing CPQ quote and an agent that writes a correct one look identical until someone checks, and the checking is exactly the labour you were trying to remove. Unconfirmed, but my expectation: a meaningful share of the 40% Gartner expects cancelled by 2027 will go quietly, because verification overhead ate the savings, not because the agent failed loudly.

Three practical consequences I would hold to. First, buy for a 35% baseline and treat anything above it as upside. If the business case only works at 80% autonomous completion, you do not have a business case, you have a bet on METR's doubling curve holding for three more cycles.

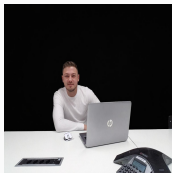
Second, separate the agent from the orchestration. The value in most workflow automation sits in reliable routing, retries, logging and state, conventional engineering that works today at close to 100%. The model call is one node. I went through the trade-offs between the main platforms in the [honest automation stack comparison](#), and the structural point stands whichever you pick: a deterministic spine with probabilistic nodes fails in ways you can debug, while an all-agent architecture fails in ways you can only re-prompt.

Third, design oversight for the case where the agent is confidently wrong, because that is the case Article 14 is written about and the one CRMArena-Pro's confidentiality finding points at. If a human cannot reverse the action after the fact, approval before the fact is not enough.

One soft spot in my own position: METR's numbers are fast, and a four-month doubling in task horizon, if it is real and not an artefact of benchmark selection, makes much of what I just wrote stale inside two years. I would not sign a five-year contract on today's capability. I also would not skip the multi-turn test, because the vendors who can pass it in 2028 will be a different 130. If you are testing agents now and want a second opinion before signing, [I am happy to look at the shortlist and the benchmark you were shown](#), since those two things rarely match.



Workflow Bots: Myth vs Reality, Only 130 Real Agent Vendors and 58% That Drops to 35%



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 →](#)