

Agent Adoption by the Numbers: 71% in Production, Up From 40%, and the 76% That Still Breaks

Seventy-one percent of large US enterprises now run AI agents in production, up from 40% a year earlier. The share still piloting fell from 32% to 17%, so the growth came almost entirely out of the pilot cohort.

That is the finding from Zapier's 2026 AI Agents Survey, published 11 November 2026, based on 513 responses from managers through C-level at US organisations with 1,000+ employees, fielded 16 to 21 September. The same week, VB Pulse's October reliability wave reported that 42% of organisations running pre-release tests on AI had a system pass those tests and then fail in production anyway. Among the subset letting agents push changes without human review, that number was 76%.

Two surveys, two different populations, one week apart. Read together they describe something I find more interesting than either number alone: adoption is not outrunning

reliability work. It is outrunning the pilot phase that was supposed to produce the reliability work.

71%
enterprises with AI agents in production
Zapier AI Agents Survey, n=513, fielded 16 to 21 Sep 2026

17%
still in pilot/testing, down from 32%
Zapier, 2026 wave vs 2025 wave (n=514)

42%
had an eval-passing system fail in production
VB Pulse October wave, 106 qualified responses

76%
same failure rate where agents ship without human review
VB Pulse October wave, 6 Nov 2026

The pilot cohort did not grow, it drained

The headline most people will take from Zapier is 88%: that share of US enterprise leaders say their organisation is using or testing AI agents, up from 72% in 2025. Sixteen points of new entrants. Fine, but unremarkable at this stage of a hype cycle.

The number that actually tells you something is 17%. If the piloting cohort had stayed flat at 32% while production climbed, you could tell a comfortable story: new organisations enter, pilot, graduate, and a healthy queue keeps forming behind them. That is not what happened. Production rose 31 points and piloting fell 15. Arithmetic says most of the production growth came from organisations that were piloting a year ago and are no longer piloting, plus a slice from the 12% who were not using agents at all in 2025.

My read: a pilot that converts is supposed to leave something behind. Evaluation harnesses, failure taxonomies, a sense of which tasks the agent is reliably bad at. A pilot that gets skipped or compressed leaves nothing behind except a deployment date. The Zapier data cannot tell us which of those two happened, because it does not ask. The VB Pulse data gives us a hint about the outcome either way.

Why do AI agents that pass evals still fail in production?

Because the evals measure the wrong surface, and the gap gets wider the less a human stands between the agent and the customer.

The VB Pulse October wave is small: 117 responses, 106 qualified, from people who build AI or consult on it at organisations with 100+ employees. No published margin of error. Treat the figures as directional. What they show is a clean gradient.

42% of all eval-running organisations had a system pass pre-release testing and fail in production. A quarter (25%) say it happened more than once. Of those who had a failure, 49% say the most significant one reached customers. And among organisations that let agents push some changes without human review, the failure rate climbs to 76%.

The September wave, 115 responses, found the same shape on a slightly different question: 80% of enterprises allowing agents to skip human review reported a customer-facing failure after the AI passed tests, against 36% where review was required. Two waves, two phrasings, roughly the same two-to-one ratio.

One figure in there reads backwards at first. Organisations using automated checks or monitoring alerts reported failures at 63%; those relying on people to notice reported 34%. VB Pulse describes the automated group as nearly twice as likely to report a failure. I do not read that as monitoring causing failures. I read it as monitoring causing *detection*. The 34% shops are not safer, they are blinder. If your only failure detector is a human noticing something odd, your reported failure rate measures attention span rather than system health.

The two surveys sample different populations and were not designed to be joined. Zapier covers US organisations with 1,000+ employees (±4% at 95% confidence, unweighted); VB Pulse covers AI practitioners at organisations with 100+ employees, 106 qualified responses, no margin of error published. Every cross-reading in this piece, mine included, is inference rather than a measured correlation.

Thirty percent say agents run the majority of tasks unsupervised

That is the Zapier figure I would put on a slide for a board asking why reliability spend went up. Among organisations using agents, 30% of leaders say the technology handles a majority of tasks without human intervention. Set that beside the VB Pulse 76% and the shape of the exposure is clear enough: the configuration associated with the worst failure rate is close to a third of this market.

I want to be careful about what that does and does not establish. “Handles a majority of tasks without human intervention” (Zapier) and “lets agents push some changes without human review” (VB Pulse) are not the same claim. The first is about task execution, the second about change deployment. An agent can execute a thousand customer conversations unsupervised while every configuration change still goes through a pull request. The overlap between the two groups is unmeasured. What I will say is that both numbers point the same direction, and neither population looks like a cautious one.

The comparison worth sitting with is year over year. In the 2025 wave, 40% had agents in production and 32% were piloting, roughly a 1.25-to-1 ratio of shipped to testing. In 2026 it is 71% to 17%, better than four to one. Whatever ratio of production systems to active experimentation your organisation thinks is prudent, the market median moved by a factor of three in twelve months.

The 42% baseline matters more than the 76%

My take: even with a human in the review loop and a working eval suite, four in ten organisations shipped something that passed and then broke. That tells me the standard eval setup, prompt-level test cases run before release, does not cover the failure modes that matter in agentic systems: state that drifts across a long run, tool calls that succeed with wrong arguments, retries that compound. Removing human review does not create those failures. It removes the last layer that was catching them.

Three decisions this data should change on Monday

Only one of them costs much.

First, instrument before you widen autonomy. The 63% versus 34% split says most organisations do not know their own failure rate; they know their noticing rate. If you cannot produce a number for how many agent runs ended in a bad outcome last month, your autonomy decision is being made without the input that matters. This is cheap. It is logging and alerting, not research.

Second, treat the review gate as a measured control rather than a cultural preference. Both VB Pulse waves put the gap between review and no-review at roughly two to one on customer-facing failure. If you are going to remove the gate, remove it for a defined task

class, keep the failure counter running, and compare. The surveys measure a binary; your system does not have to be one.

Third, be honest about what your pilot produced. If an organisation moved from the piloting 32% to the production 71% in the past year, the useful question is what artefacts came out of that pilot: an eval suite that catches tool-call errors, a rollback path, a list of inputs that reliably break the agent. If the answer is a demo and a positive steering-committee note, the pilot was a procurement exercise. That is not a crime, but it means the reliability work is still ahead of you, and the 42% suggests it is harder than the eval vendors' marketing implies.

On tooling, the consolidation is real and fast-moving. Fifty-seven percent of VB Pulse respondents use Anthropic's Claude Console to evaluate AI, 43% use OpenAI's developer platform. More telling: OpenAI's native eval tooling fell as primary evaluation platform from 59% in August to 32% in September 2026, a 27-point drop in a single month on a sample of around 115. That is a small sample and a short window, so I would not call it a trend yet. But if you are standardising on a single vendor's eval stack, note that the practitioner population moved off one of them faster than most procurement cycles complete.

The durability question sits underneath all of this. An agent that passes evals and then fails in production often fails because the run did not survive something: a crash, a timeout, a restart in the middle of a multi-step task. I wrote about which agent frameworks actually survive a crash because that failure mode is invisible to prompt-level evaluation by construction. And on the question of how many of these production deployments are agents in any meaningful sense, the vendor-side numbers are less flattering than the adoption surveys.

What I am confident about: production adoption tripled against the pilot cohort in twelve months, and the configuration with the highest measured failure rate is the one that removes human review. What I am not confident about: whether the 42% baseline failure rate is getting better or worse, because VB Pulse has two waves and no year-ago comparison. If your own numbers say something different from these surveys, I would genuinely like to hear what they say, because the gap between what organisations report and what their logs show is the part nobody has measured yet.



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 →