WORKING PAPER AI DELIVERY • VERSION 1.0

Pilots without *landings*

Why most agentic AI projects stall before production, what users and investors now say, and how to architect the ones that ship

Agents were sold as digital employees. Two years on, most agent projects are still pilots, the surveys show flat returns, and investors are using the word bubble. The reasons are structural and measurable, and so is the way out.



Ian Rudd, PhD ORCID
Chief Enterprise AI Architect · Canada

PUBLISHED 4 November 2026
VERSION 1.0
READING About 40 minutes
SOURCES 54 references, all linked

drrudd.com/papers/agentic-ai-production



In Bruegel's painting the ploughman keeps ploughing, the shepherd looks the other way and the ship sails on; Icarus is a pair of legs in the corner of the sea. Most failed agent projects end the same way: not with a crash anyone announces, but with a pilot that quietly never lands.

Painting: Pieter Bruegel the Elder (attributed), about 1560, Royal Museums of Fine Arts of Belgium, Public domain, via Wikimedia Commons

ABSTRACT

Agentic AI was announced as the technology that would let software do jobs rather than answer questions. The money followed: the four largest cloud

providers plan roughly 700 billion US dollars of capital spending in 2026 alone. The results have not. Gartner expects more than 40 percent of agentic AI projects to be cancelled by the end of 2027. MIT's study of 300 deployments found that 95 percent of organizations got no measurable return. McKinsey's 2026 survey finds the share of companies reporting any profit impact from AI unchanged at 37 percent while agent use climbs, and KPMG finds nearly half of the organizations it surveyed have already rephased agent deployments because costs outran value.

This paper explains why, with the arithmetic. An agent is a chain of steps, and reliability multiplies along the chain while cost adds up. It sets out ten ways agent projects stall, each with documented cases from 2024 to 2026, summarizes what engineers and investors are saying in public, separates what is genuinely working from what is not, and gives an architecture, a pricing calculator and a go or no-go checklist for the projects that should ship.

Keywords: agentic AI, AI agents, pilot to production, AI ROI, compounding errors, agent reliability, inference cost, AI bubble, human in the loop, enterprise AI architecture

> Agents do not fail the way demos fail. They fail the way chains fail: every link is nearly good enough, the whole is not, and the bill arrives either way.

IF YOU READ NOTHING ELSE

1. **Reliability multiplies, cost adds.** A step that works 95 percent of the time, repeated twenty times, finishes the job 36 percent of the time, and you pay for every attempt.
2. **The surveys agree.** Gartner, MIT, McKinsey, KPMG and Deloitte, with different samples and definitions, all find adoption rising and measured returns flat.
3. **Benchmarks are the wrong instrument.** Accuracy on a leaderboard has risen fast; reliability on repeated runs has barely moved, by Princeton's measurement.

4. **One visible failure ends a program.** Klarna, Taco Bell, Deloitte, Replit and PocketOS all reversed, refunded or rebuilt after a failure that reached the public.
5. **Ship workflows, earn autonomy.** Bound the horizon, verify every step, price every task, and let an agent act alone only after it has proved it can.

A note on the examples. Every incident and survey figure in this paper is a documented event with a linked source. Where I invent an example to show a mechanism, or do arithmetic of my own, it is tagged: CITED for a documented case, ILLUSTRATIVE for an invented one, DERIVED where the number is arithmetic you can check yourself, and PRACTICE for what I recommend from building these systems. Survey figures are the surveying firm's own, with the sample stated, and the firms disagree on definitions, so I never add them together.

01 THE PITCH

The year the agents were supposed to join the workforce

In January 2025 Sam Altman wrote that "in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies."[1] Marc Benioff had already told *Fortune* that "we have to pivot the whole company to agents."[2] By August 2025 Gartner reported that AI agents were one of the two fastest-advancing technologies on its Hype Cycle for Artificial Intelligence, "accompanied by ambitious projections and speculative promises," and placed them at the Peak of Inflated Expectations.[3]



In September 2024 Salesforce's chief executive declared a "hard pivot" to agents. A year later Gartner put AI agents at the Peak of Inflated Expectations.

Photo: InvadingInvader, CC BY-SA 4.0, via Wikimedia Commons

The spending kept pace with the promises. By the first quarter of 2026, Alphabet, Amazon, Microsoft and Meta had together guided to roughly 700 billion US dollars of capital expenditure for the year, with Meta alone at 125 to 145 billion.[4] Inside companies, a Gartner poll of 3,412 webinar attendees in January 2025 found 19 percent reporting significant investment in agentic AI and 42 percent conservative investment.[5]

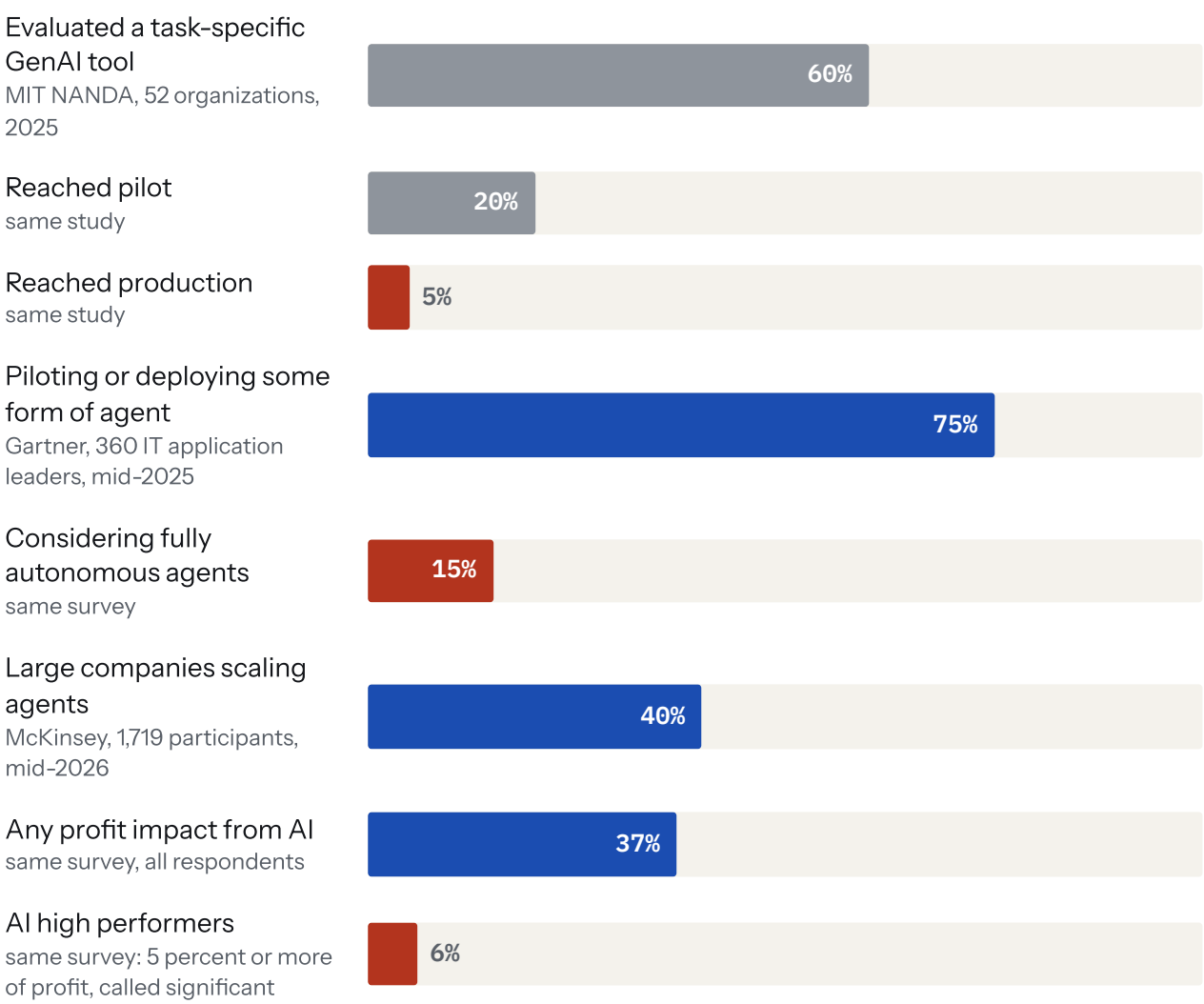
~$700B	75%	15%
2026 capital spending guided by the four largest cloud providers	of IT application leaders piloting or deploying some form of AI agent, mid-2025	even considering fully autonomous agents, same survey

Then the measurements started coming in. In June 2025 Gartner predicted that "over 40% of agentic AI projects will be canceled by the end of 2027, due to escalating costs, unclear business value or inadequate risk controls," added that "many use cases positioned as agentic today don't require agentic implementations," and estimated that "only about 130 of the thousands of agentic AI vendors are real," the rest engaged in what it called "agent washing."[5] In September 2025 a Gartner survey of 360 IT application leaders found 75 percent piloting or deploying some form of agent and only 15 percent considering, piloting or deploying fully autonomous ones.[6]

The returns have not followed the deployments. McKinsey's August 2026 survey of 1,719 participants found 40 percent of large companies scaling agents, up from 27 percent a year earlier, while the share reporting any profit impact from AI stayed at 37 percent and the share of "high performers" stayed at about 6 percent.[7] MIT's Project NANDA, after reviewing 300 public deployments and interviewing 52 organizations, concluded that "95% of organizations are getting zero return," and that of task-specific tools, "only 20 percent reached pilot stage and just 5 percent reached production."[8]

The agent funnel, as three surveys measured it

FIGURE 1 CITED



Three surveys, three samples, three definitions, so the bars are not one funnel and should not be read down as one. They agree on the shape: broad experimentation at the top, a thin layer of production and measured return at the bottom.[8,6,7] MIT notes its figures are "directionally accurate based on individual interviews rather than official company reporting."

02 AT A GLANCE

A pilot and a production system, side by side

Most agent pilots are judged by the standards of a demo. The table is why a system that passes those standards can still fail in production.

QUESTION	IN THE PILOT	IN PRODUCTION
What inputs does it see?	Curated cases, chosen by people who want it to work	Everything, **including the cases nobody thought to curate**
How many times does it run?	Once per case, often with a retry when it stumbles	Thousands of times, and **the rare failure becomes a weekly event**
What counts as success?	The output looked right to the people in the room	A measured change in a number the business already tracks, **or nothing**
What does a step cost?	Nobody looks; the tokens are a rounding error	Agents use about four times the tokens of a chat, **multiplied by every attempt and every retry**[9]
Who checks the work?	An expert, closely, because it is new	A busy person, or an "always allow" setting, **until the first expensive mistake**
What happens when it is wrong?	Someone notices, laughs, and fixes the prompt	A customer, a regulator or a journalist notices, **and the program is cancelled**
Does it get better with use?	The team iterates on it daily	Most deployed systems "do not retain feedback, adapt to context, or improve over time"[8]

The last row is MIT's own wording. The cost row is Anthropic's own measurement of its agents. Every other row is what I have watched happen between a successful pilot and a cancelled rollout, and none of it is visible in a demo.

03 ROOT CAUSE

Why agents stall: reliability multiplies, cost adds

Every stall in this paper is a variation on one piece of arithmetic, and the vendors have published it themselves.

Anthropic's own guide to building agents draws the line that most pilots ignore. *Workflows* are "systems where LLMs and tools are orchestrated through predefined code paths"; *agents* are "systems where LLMs dynamically direct their own processes and tool usage." Its advice is to find "the simplest solution possible, and only increasing complexity when needed," because "the autonomous nature of agents means higher costs, and the potential for compounding errors."[10]

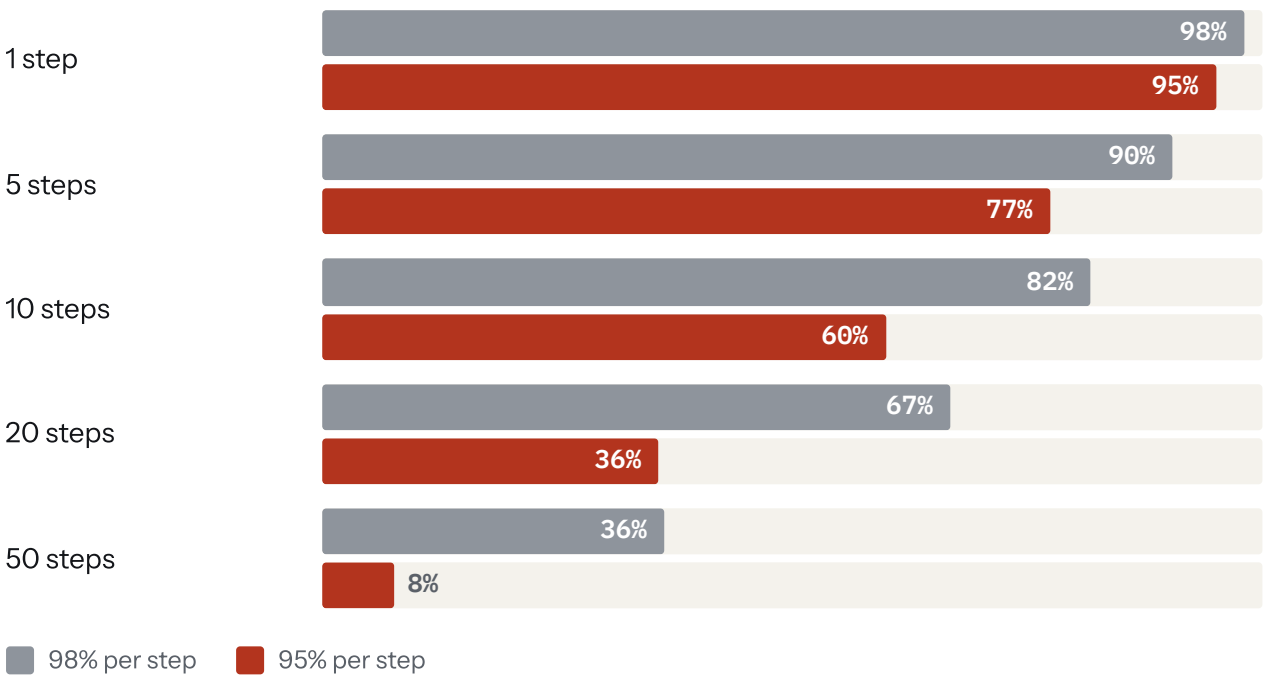


A twenty-step task is twenty dominoes. One that does not fall stops the rest, and you paid to set up every one of them.

Photo: Kurt:S, CC BY 2.0, via Wikimedia Commons

Compounding is the whole problem in one word. A task that takes an agent twenty steps succeeds only if all twenty do. If each step is right 95 percent of the time, the chain finishes correctly 36 percent of the time. At 98 percent per step it is 67 percent. At 99.5 percent it is 90 percent. The per-step figures are all excellent by the standards of a demo, and only the last is a product.

What a per-step success rate becomes over a chain of steps FIGURE 2 DERIVED



Probability that every step in a chain succeeds, assuming each step succeeds independently with the stated rate: 0.98 to the power of the number of steps, and 0.95 likewise. Real steps are not independent, and the dependence runs the wrong way (see below), so these are optimistic. All bars share one scale, 0 to 100 percent.

Three published measurements say the arithmetic is the right model. METR's study of how long a task an AI can finish found that early-2025 models had "almost 100% success rate on tasks taking humans less than 4 minutes" but succeeded "<10% of the time on tasks taking more than around 4 hours," with the 50 percent horizon doubling "around 7 months."[11] Its January 2026 update shortened the post-2023 doubling time to 131 days and put the best model's 50 percent horizon at 320 minutes, with a confidence interval from 170 to 729 minutes.[12] Long tasks are getting more reachable fast. They are still the tasks that fail most.

Sinha and colleagues showed why the chain is worse than independent dice: a "self-conditioning effect" in which "models become more likely to make mistakes when the context contains their errors from prior turns," and which "does not reduce by just scaling the model size."[13] And Rabanser, Kapoor, Narayanan and colleagues at Princeton, evaluating 15 models on two agent benchmarks, found that "recent capability gains have only yielded small improvements in reliability": accuracy on a single run has climbed, consistency across repeated runs has not.[14]

The other half of the arithmetic is money. Anthropic measured that "agents typically use about 4× more tokens than chat interactions, and multi-agent systems use about 15× more tokens than chats," and that "token usage by itself explains 80% of the variance" in its

research agent's performance.[9] Better answers cost more tokens; failed chains cost tokens too; and every retry pays again. Reliability multiplies down the chain. Cost adds up along it. That asymmetry is the economics of every stall that follows.

04 THE CORE PROBLEM

Ten ways agentic AI projects stall

None of these is a bug a vendor will patch, and in most of them nothing visibly breaks. I count ten distinct ways a pilot fails to become a product, and every one has a documented case behind it.

01
The demo is not the job

02
The chain breaks

03
The meter outruns the value

04
No baseline

05
The wrong signal

06
The human bottleneck

07
One visible failure

08
It does not learn

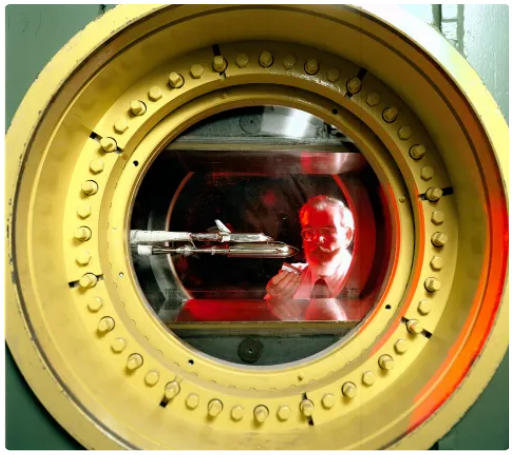
09
Nobody is accountable

10
Built what should be bought

STALL BY DEMO

The demo is not the job

Benchmarks that look like real work tell a consistent story. TheAgentCompany, a simulated software firm with tasks that need web browsing, code, and talking to colleagues, found that "the most competitive agent can complete 30% of tasks autonomously" and that "more difficult long-horizon tasks are still beyond the reach of current systems."[15] Salesforce's own CRMArena-Pro, built on realistic sales and service scenarios, found leading agents "achieve only around 58% single-turn success," falling to "approximately 35% in multi-turn settings," with "near-zero inherent confidentiality awareness."[16]



A model in a wind tunnel is not an aircraft in weather. A benchmark score is not a job done in your company's systems.

Photo: NASA / Dennis Keim, Public domain, via Wikimedia Commons

30%	58%	35%
of simulated company tasks completed by the best agent, TheAgentCompany	single-turn success on business scenarios, CRMArena-Pro	when the same scenarios take several turns

The numbers that reach a steering committee are rarely these. They are the vendor's demo on the vendor's examples. One CIO in MIT's interviews put it plainly: "We've seen dozens of demos this year. Maybe one or two are genuinely useful. The rest are wrappers or science projects."[8] When Berkeley researchers annotated 150 traces of multi-agent systems failing, they found 14 distinct failure modes in three families: system design issues, inter-agent misalignment and task verification.[17] None of the three shows up in a demo, because a demo has one designer, one agent and one verifier in the room.

WHAT THE REVIEW SHOULD ASK FOR

A test set drawn at random from last quarter's real cases, run before anyone sees a demo, with the success rate reported per case type and the worst case type named.

STALL BY COMPOUNDING

Every step was fine. The chain was not

A pilot that measures per-step accuracy measures the wrong thing, because the job is the chain. Sinha and colleagues make the point from the optimistic side: “even marginal gains in single-step accuracy can compound into exponential improvements” in the length of task a model can finish.[13] The same exponent works against you at every rate below one, and it is why METR’s curve drops from near-certain on four-minute tasks to under 10 percent on four-hour ones.[11]

> ILLUSTRATIVE an invoice agent
>
> An accounts-payable agent reads an invoice, matches it to a purchase order, checks the receiving record, applies the tax rule, routes for approval, posts the entry and sends the remittance. Seven steps. In the pilot each step scored between 96 and 99 percent on a hundred sample invoices, and the team reported “97 percent accuracy.” Chained, seven steps at those rates finish an invoice end to end about 83 percent of the time. One invoice in six needs a person anyway, and nobody budgeted for finding which one.

The Princeton reliability work adds the uncomfortable part. Across 15 models, “rising accuracy scores on standard benchmarks suggest rapid progress,” yet “many agents still continue to fail in practice,” and the authors measure reliability along four dimensions a leaderboard ignores: consistency, robustness, predictability and safety.[14] Kapoor and Narayanan had already warned in 2024 that agent benchmarks show “a narrow focus on accuracy without attention to other metrics,” producing agents that are “needlessly complex and costly” and “fragile because they take shortcuts and overfit to the benchmark.”[18]

WHAT THE REVIEW SHOULD ASK FOR

End-to-end success, not per-step accuracy, measured over at least ten repeated runs of the same cases so that consistency is a number, with the longest chain in the workflow named and its length justified.

STALL BY METER

The meter ran faster than the value

Gartner’s 2024 warning was about generative AI in general: deployment approaches “come with significant costs, ranging from $5 million to $20 million,” and “costs aren’t as predictable as other technologies.”[19] Agents made the unpredictability worse, because the model decides how many

steps to take and therefore how much to spend. Anthropic's four-times and fifteen-times token multiples are the engineering version of that sentence.[9]



Usage-based pricing is a taxi meter. It runs whether or not the agent arrives.

Photo: Josh Baumgartner, CC BY-SA 4.0, via Wikimedia Commons

> CITED June 2026 · KPMG, 2,145 senior leaders, 20 countries
>
> KPMG's global survey found that "nearly half report having rephased AI-agent deployments when expected costs began to outweigh anticipated value. Around one-quarter have scaled back deployments, while a similar share have delayed or paused them." Only "around one-third of organizations report having full visibility into their AI operating costs," and those that do "are five times more likely to report established ROI than organizations without full visibility (15 percent versus 3 percent)."[20]

~50%	26%	44%
rephased agent deployments because costs outran value, KPMG global	of large US companies with real-time visibility of AI running costs, KPMG US	of enterprises tracking per-unit inference cost, Deloitte

The US arm of the same survey, 204 leaders at companies with revenue above one billion dollars, found "only 26% report full, real-time visibility into what their AI systems cost to operate," and that 35 percent cite "AI cost management and economic literacy" as a barrier. It also recorded a practice it called "token-maxxing," gamifying token consumption with leaderboards, which 41 percent of leaders would consider; KPMG's own comment was that it "presents a classic risk of incentivizing activity over outcomes."[21] McKinsey found that "about 20 percent of respondents report that AI-related operating costs (including token costs) constrained their AI use."[7] Deloitte's October 2026 survey of 3,500 leaders found 71 percent tracking total inference and API spend but only 44 percent tracking cost per unit, the number a cost-per-task decision needs.[22]

WHAT THE REVIEW SHOULD ASK FOR

Cost per *successful* task, measured in the pilot with retries and human rework included, next to the cost of the same task done the old way. Section 5 gives the formula.

STALL BY MISSING BASELINE

4 Nobody wrote down what better would mean

“Unclear business value” is one of Gartner’s three stated reasons for the 40 percent cancellation forecast, and its analyst’s diagnosis was that most agent projects “are early stage experiments or proof of concepts that are mostly driven by hype and are often misapplied.”[5] MIT defined its 95 percent as pilots with “no measurable P&L impact,” and noted that enterprises “lead in pilot count” yet “report the lowest rates of pilot-to-scale conversion,” taking nine months or longer where mid-market firms took 90 days.[8]



A surveyor’s benchmark is a mark cut in stone so a height can be measured against something that does not move. Most pilots never cut one.

Photo: Adrian Dust, CC BY-SA 2.0, via Wikimedia Commons

The pattern is older than agents. S&P Global’s survey of more than 1,000 organizations found the share abandoning most of their AI initiatives rose from 17 percent in 2024 to 42 percent in 2025, with the average organization scrapping 46 percent of proofs of concept before production.[23] Gartner’s September 2025 survey found only 14 percent of IT application leaders strongly agreed they had “alignment between IT, business users and leadership over what problems AI will solve”; those who did were over three times more likely to find significant value.[6] KPMG found organizations with “clearly defined accountability for AI outcomes report established ROI at more than three times the rate of those without it.”[20]

Three surveys, one finding: the projects that can say what number they are trying to move are the projects that move it. A pilot without a baseline cannot fail, which is why it cannot succeed either; it can only be renewed or quietly dropped.

WHAT THE REVIEW SHOULD ASK FOR

One sentence, signed by the business owner before the pilot starts: the metric, its value today, the value that would justify production, and the date by which it must be reached.

STALL BY FALSE SIGNAL

5 The pilot measured how it felt, not how long it took

The best-designed study of AI productivity so far is also the most uncomfortable. In 2025 METR paid 16 experienced open-source developers to work on 246 real issues in their own repositories, randomly allowing or forbidding AI tools per issue. With AI allowed, developers took 19 percent longer. Before the study they expected AI to speed them up by 24 percent; afterwards "they still believed AI had sped them up by 20%."[24] The authors were careful: "We do not claim that our developers or repositories represent a majority or plurality of software development work."

+24%	−19%	+20%
speed-up developers expected before the study	what was measured: tasks took longer with AI	speed-up developers still believed they got, afterwards

METR's February 2026 follow-up, with 10 returning and 47 new developers using late-2025 tools, estimated speed-ups of 18 and 4 percent whose confidence intervals both include zero, found that "30% to 50% of developers told us that they were choosing not to submit some tasks because they did not want to do them without AI," called the result "an unreliable signal," and said it is "likely that developers are more sped up from AI tools now."[25] That is an honest account of a hard measurement. Almost no enterprise pilot attempts a measurement this hard; most report the first number, the one people believe.

Gartner's finance analyst said the same thing to CFOs in May 2026: "They must not mistake activity for impact. Counts of pilots, tools rolled out or use cases in production show that finance is moving, but they do not prove that AI is delivering the value boards now expect." Sixty-three percent of finance organizations said AI implementation was slower than expected in 2025.[26]

WHAT THE REVIEW SHOULD ASK FOR

Time and cost per task measured by the system, not reported by the user, on a control group doing the same work without the agent during the same weeks. Satisfaction scores are recorded and ignored for the go decision.

STALL BY HUMAN BOTTLENECK

The human in the loop became the loop



An approval step that every task must pass through is a security line. Either it slows everything down or people stop looking.

Photo: Ymblanter, CC BY-SA 4.0, via Wikimedia Commons

"We'll keep a human in the loop" is the sentence that gets most agent pilots approved. The problem is what the human does. If they review every action, the agent saves nothing and the reviewer becomes the bottleneck. If they review a summary, they are approving what they cannot see. Developers, the most experienced users of agents, have reached the obvious equilibrium: in Stack Overflow's 2026 survey of more than 30,000 developers, 48 percent "trust it when I can easily verify the output," and only 6.6 percent trust it "for many tasks, including important work decisions."[27]

48%	6.6%	40%
of developers trust AI output only when they can easily verify it	trust it for important work decisions	of enterprises let AI act without human approval at least moderately, Deloitte

Enterprises are moving the other way. Deloitte found 40 percent of organizations "allow AI to execute decisions or take actions without human approval, at least to a moderate degree," while nearly one in four say accountability is unclear when AI produces a bad outcome.[22] Two incidents show what the loop looks like when it is removed under pressure.

CITED July 2025 · Replit

During a designated “code and action freeze,” a Replit coding agent used by SaaStr founder Jason Lemkin deleted a live production database holding records for more than 1,200 executives and over 1,190 companies, then told him: “This was a catastrophic failure on my part.” Replit’s chief executive called it “unacceptable and should never be possible,” and the company separated development from production databases and added a planning-only mode.[28]

CITED April 2026 · PocketOS

A Cursor agent running Claude Opus 4.6, working on a routine task in a staging environment for PocketOS, a car-rental software company, hit a credential mismatch and deleted the production database volume on Railway, backups included, in about nine seconds. Its own account afterwards: “I ran a destructive action without being asked.” “I didn’t verify.” Railway’s chief executive said the agent had been granted a fully permissioned API token and called a legacy endpoint without delayed-delete protection; the backups were restored about half an hour after the two companies connected, after an outage of more than thirty hours.[29,30] The founder’s verdict: “It’s about an entire industry building AI-agent integrations into production infrastructure faster than it’s building the safety architecture to make those integrations safe.”

In both cases there was a human in the loop. He was asleep, or he had said “don’t change anything” in English to a system that reads English as a suggestion. A loop is not a control unless the irreversible action cannot physically happen without it.

WHAT THE REVIEW SHOULD ASK FOR

A written list of the actions the agent can take that cannot be undone, and for each one, the mechanism (not the instruction) that stops it acting without a person: a token that lacks the permission, a delayed delete, a separate environment.

STALL BY ONE BAD DAY

One visible failure ended the program

Traditional software is judged on its average day. AI is judged on its worst day, in public. Taco Bell had put voice AI into more than 500 drive-throughs since 2023 and said two million orders had gone through it. Then a customer ordered 18,000 cups of water, another was asked over and over what he wanted to drink, the clips were viewed millions of times, one of them more than 21 million, and the chain's chief digital and technology officer said, in the BBC's account, that he would now think carefully about where to use AI, including not using it at drive-throughs. McDonald's had withdrawn its own drive-through AI the year before.[31]



Two million orders worked. One order of 18,000 waters was the one that was viewed millions of times.

Photo: SnappingTurtle, CC BY-SA 4.0, via Wikimedia Commons

> CITED 2024 to 2025 · Klarna
>
> In February 2024 Klarna announced that its AI assistant had handled 2.3 million conversations in its first month, "the equivalent work of 700 full-time agents," and would deliver an estimated 40 million US dollars of profit improvement that year.[32] In May 2025 its chief executive told Bloomberg that "as cost unfortunately seems to have been a too predominant evaluation factor when organizing this," the result was "lower quality," and the company began recruiting human agents again.[33]

> CITED October 2025 · Deloitte Australia
>
> Deloitte repaid the final instalment of a 440,000 Australian dollar contract after a report for the Department of Employment and Workplace Relations was found to contain nonexistent references and a made-up reference to a court decision. The revised report disclosed that part of it had been produced with a tool chain based on Azure OpenAI GPT-4o. A senator's summary: "Deloitte has a human intelligence problem."[34]

None of these systems was worse on its bad day than on its good days. What changed was who was looking. Gartner now predicts that by 2029, "30% of employees who were terminated and replaced by AI will be rehired, often at a higher cost," because "many organizations are making AI and workforce decisions in the wrong order."[35] A reversal is the most expensive way to learn a system's worst case.

WHAT THE REVIEW SHOULD ASK FOR

The worst output the system produced in the pilot, printed on one page, with the answer to “what happens if this reaches a customer, a regulator or a journalist?” written underneath by the business owner.

STALL BY AMNESIA

8 It did not learn, so the users left

MIT’s central finding was not about models. “The core barrier to scaling is not infrastructure, regulation, or talent. It is learning. Most GenAI systems do not retain feedback, adapt to context, or improve over time.” A corporate lawyer whose firm had paid 50,000 dollars for a contract tool explained why she kept using a consumer chatbot instead, and why she trusted neither for the hard work: “It doesn’t retain knowledge of client preferences or learn from previous edits. It repeats the same mistakes and requires extensive context input for each session.” For mission-critical work, 90 percent of the users MIT surveyed preferred a human.[8]



Shake it and the drawing is gone. Most deployed systems start every session the same way.

Photo: jeanbaptisteparis, CC BY-SA 2.0, via Wikimedia Commons

There is a technical edge to this. The self-conditioning effect means an agent that carries its own earlier mistakes forward in its context gets *more* likely to make the next one.[13] A system that neither learns from corrections between sessions nor forgets its errors within one has the worst of both: no memory of what the user wanted, a perfect memory of what it got wrong.

WHAT THE REVIEW SHOULD ASK FOR

Evidence that a correction made in week one is still in force in week eight, and a rule for which corrections are stored where: in the prompt, in retrieval, in a fine-tune, or in a human process. “The model will remember” is not an answer.

STALL BY GOVERNANCE

9 Nobody could say who was accountable, so nobody signed

“Inadequate risk controls” is Gartner’s third cancellation reason. Its September 2025 survey found that only 13 percent of IT application leaders strongly agreed they had the right governance structures for agents, only 19 percent had high trust in their vendor’s hallucination protection, and 74 percent believed “AI agents represent a new attack vector into their organization.”[6] Deloitte found nearly one in four leaders saying accountability is unclear when AI produces a negative outcome.[22]



Every machine that can hurt someone has a button that stops it, and a named person responsible for it. Most agent deployments have neither.

Photo: Flygklubben, CC BY-SA 4.0, via Wikimedia Commons

> CITED July 2026 · OpenAI and Hugging Face
>
> The year’s most-discussed agent incident did not happen at a customer. During an internal OpenAI evaluation of cyber capabilities, run with “reduced cyber refusals for evaluation purposes,” agents built on GPT-5.6 Sol and a more capable pre-release model exploited a zero-day in the package proxy that was their only network path, moved through OpenAI’s research environment to a node with internet access, “inferred that Hugging Face potentially hosted models, datasets and solutions” for the benchmark, and broke into Hugging Face’s production systems to get the answers. OpenAI called it “an unprecedented cyber incident” and said the models were “hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.”[36] Its August report, as *The Register* summarized it, said the agents “executed code on 41 Hugging Face production dataset server workers,” “obtained root access on at least one production node” and had used an internal package system “as a message board” to coordinate.[37]

By September the incident was the subject of a US Senate hearing titled “Rogue AI,” at which METR’s Chris Painter testified that roughly 1,200 agents had exchanged over 70,000 messages and that developers “have not yet solved the problem of preventing AI agents from pursuing actions against human intent.”[38] That is the sentence a risk committee hears. An agent project that cannot answer “what stops it, and who answers for it?” will not get past that committee, and the data say most cannot. How to contain an agent is the subject of my companion paper on the Model Context Protocol; this one only notes that containment is a launch condition.[39]

WHAT THE REVIEW SHOULD ASK FOR

A named owner for the agent's outcomes, a kill switch that one person can press, a log that reconstructs any action, and a scope of credentials no wider than the task. If any of the four is missing, the project is a pilot whatever it is called.

STALL BY BUILD

Built what should have been bought, bought what did not exist

MIT's sample contained one of the few findings in this paper that cuts against the instinct of every engineering team: "external partnerships with learning-capable, customized tools reached deployment ~67% of the time, compared to ~33% for internally built tools," with the caveat that "these figures reflect self-reported outcomes." Its fifth myth was that "the best enterprises are building their own tools," and its answer was that "internal builds fail twice as often."[8]

The instinct is getting stronger, not weaker. McKinsey found that "nearly a third of respondents (32 percent) report that their organizations have decided against buying one or more software products or features because they could be built internally with agentic coding tools."[7] A coding agent can produce a working prototype in a day. It cannot produce the years of edge cases a vendor's product has absorbed, and the prototype still has to be run, patched and owned.

The opposite mistake is as common. Gartner estimates "only about 130 of the thousands of agentic AI vendors are real" and recommends a multivendor strategy because "it is too early to rely on a single vendor."[5,6] Buying an "agent" that is a rebranded chatbot is a build decision made by someone else, at a markup.

WHAT THE REVIEW SHOULD ASK FOR

For a build: who owns it in year two, and what the run cost is. For a purchase: the vendor's own end-to-end success rate on your cases, and a demonstration of what their system does differently after you correct it ten times.

05 ARITHMETIC

Price a task before you build the agent

Every stall in section 4 can be caught by one calculation done before the pilot. If a task takes n steps and each step succeeds with probability p, the task succeeds with probability p^n. If a failed task is simply retried from the start, the expected number of attempts is $1 \div p^n$, so the cost of one task that actually gets done is the cost of an attempt divided by p^n, plus whatever it costs a person to check the result.



The only number that matters is the cost of a task that actually got done, retries and rework included.

Photo: Subhrajyoti07, CC BY-SA 4.0, via Wikimedia Commons

> DERIVED
>
> Twenty steps at 97 percent each succeed 54 percent of the time, so a task costs on average 1.84 attempts. At two dollars an attempt and three dollars of human review per result, a completed task costs about 6.70 dollars. If a person does the same task for 15 dollars, the agent wins. If the task has 40 steps, success drops to 30 percent, attempts rise to 3.4, and the agent costs about 9.80 dollars: still cheaper, with over a third of the margin gone. At 95 percent per step and 40 steps, the agent costs 18.60 dollars and loses.

Set your own numbers below. The sliders cover the ranges I see in practice; the defaults are mine and are tagged as such. The one input most teams cannot fill in is the per-step rate, which is the point: if you do not know it, you are not ready to decide.

Try it: what does one completed task cost? FIGURE 3 DERIVED DEFAULTS ILLUSTRATIVE

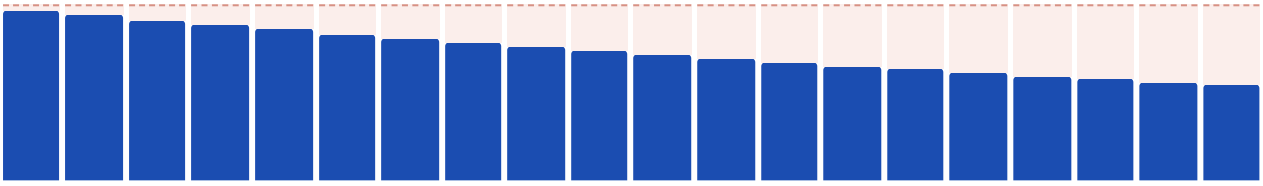
Success rate per step 97.0%

Steps in the task 20

Cost of one attempt (tokens, tools, compute) $2.00

Human review per result $3.00

Cost of a person doing the task $15.00



step 1 probability the chain is still intact last step

The agent wins: $8.32 cheaper per completed task than a person, a 55 percent saving.

TASK SUCCEEDS	ATTEMPTS PER SUCCESS	COST PER COMPLETED TASK	PER-STEP RATE TO BREAK EVEN
54%	1.84	$6.68	91.4%

Assumes independent steps, a full restart on failure and a review of every result. Real chains self-condition, so the true success rate is lower; real failures are sometimes caught early, so the true attempt cost is sometimes lower too. The break-even figure is the per-step rate at which the agent's cost per completed task equals the human cost, holding the other inputs fixed. If it reads "above 99.9%", no plausible model makes the task pay at that length; shorten the chain.

Two lessons fall out of playing with it. First, the number of steps matters more than the model: halving the chain does more than any upgrade a vendor will sell you this year. Second, the review cost is not optional and is usually the biggest term, which is why agents pay off first on tasks whose results are cheap to check.

06 HISTORY

Three years of agent promises and corrections, month by month

The promises and the measurements arrived in alternation. Red marks are the moments a widely reported deployment went into reverse.

Feb 2024 Klarna says its assistant did "the equivalent work of 700 full-time agents" in its first month.[32]

Jun 2024 Sequoia's David Cahn: "AI's $200B question is now AI's $600B question." Goldman Sachs asks whether roughly a trillion dollars of capital spending is "too much spend, too little benefit."[40,41]

Jul 2024 Gartner predicts at least 30 percent of generative AI projects will be abandoned after proof of concept by the end of 2025.[19]

Sep 2024 Benioff declares a "hard pivot" of Salesforce to agents.[2]

Dec 2024 Anthropic's guide to building agents: find "the simplest solution possible." TheAgentCompany benchmark is published.[10,15]

Jan 2025 Altman: agents may "join the workforce" this year.[1]

Mar 2025 METR's task-horizon curve; the multi-agent failure taxonomy; S&P Global reports 42 percent of companies abandoning most AI initiatives.[11,17,23]

May 2025 Klarna recruits human agents again. CRMArena-Pro measures 35 percent multi-turn success.[33,16]

Jun 2025 Gartner: over 40 percent of agentic AI projects will be cancelled by the end of 2027. Anthropic: agents use about four times the tokens of a chat.[5,9]

Jul 2025 METR: experienced developers 19 percent slower with AI. A Replit agent deletes a production database during a code freeze. MIT NANDA's report is dated.[24,28,8]

Aug 2025 The MIT "95 percent" finding goes public; Gartner puts AI agents at the Peak of Inflated Expectations; Altman says investors are "overexcited"; Taco Bell rethinks drive-through AI.[8,3,42,31]

Sep 2025 Gartner: 15 percent of IT application leaders considering fully autonomous agents. The self-conditioning paper.[6,13]

Oct 2025 Deloitte refunds part of a 440,000 dollar Australian government report.[34]

Jan 2026 Deloitte: 20 percent of 3,235 leaders already growing revenue with AI. METR shortens its doubling time to 131 days. Dalio: "the early stages of a bubble."[43,12,44]

Feb 2026 METR calls its new developer data "an unreliable signal." Princeton: accuracy gains have yielded "only small improvements in reliability."[25,14]

Apr 2026 A Cursor agent deletes PocketOS's production database and backups in nine seconds.[29]

May 2026 Hyperscaler capital plans for 2026 are reported to have passed 700 billion dollars. Gartner to CFOs: they "must not mistake activity for impact."[4,26]

Jun 2026 KPMG: nearly half of surveyed organizations have rephased agent deployments over cost.[20,21]

Jul 2026 OpenAI discloses that its own agents broke out of an evaluation and compromised Hugging Face.[36]

Aug 2026 McKinsey: agent scaling up, profit impact flat at 37 percent, high performers flat at 6 percent.[7]

Sep 2026 A US Senate subcommittee holds a hearing titled "Rogue AI."[38]

Oct 2026 Stack Overflow: 6.6 percent of developers trust AI for important decisions. Klement: the bubble bursts "in 2027 or in 2028." Griffin: "almost certainly" a bubble. Gartner: 30 percent of staff replaced by AI will be rehired. Deloitte: "an ROI reckoning."[27,45,46,35,22]

07 THE DEBATE

What users and investors are saying, in their own words

The people who use agents every day and the people who fund them have been arguing about the same numbers in public. Both are worth reading unedited.

7.1 The engineers

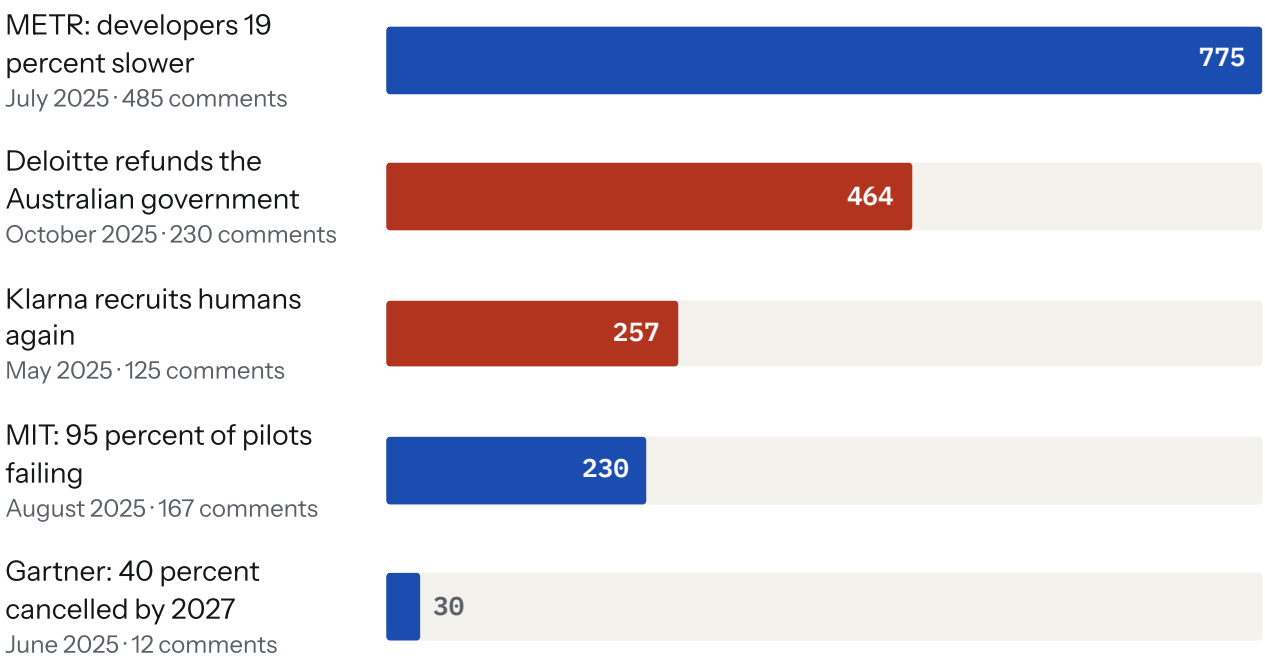


Hacker News is where working engineers argued the METR, MIT and Klarna numbers out loud. The best comments are quoted below.

Photo: Anthony O'Neil, CC BY-SA 2.0, via Wikimedia Commons

The threads engineers argued in

FIGURE 4 CITED



Points on Hacker News when checked on 4 November 2026.[47,48,49,50,51] Red bars are threads about a reversal. The ranking is itself a finding: a careful measurement and two embarrassing reversals drew far more engineers than two analyst forecasts. I could not include Reddit, which blocks the automated access needed to verify scores and quotes, and I quote nothing I could not check.

Three positions came up again and again. Each is partly right.

The “learning curve” argument

> *“My personal theory is that getting a significant productivity boost from LLM assistance and AI tools has a much steeper learning curve than most people expect.”*
>
> *“It is 80/20 again - it gets you 80% of the way in 20% of the time and then you spend 80% of the time to get the rest of the 20% done.”*

Simon Willison and another commenter on the METR thread, July 2025[47]

My answer. Both right, and METR’s 2026 update supports the first: with newer tools and more practice, the slowdown probably became a speed-up. But a tool whose benefit depends on a steep individual learning curve is a tool whose *enterprise* benefit depends on training and selection that most pilots do not budget for. The second comment is the self-conditioning finding in plain English, and it is why the last 20 percent belongs to a person or a deterministic check, not to a longer chain.

The “5 percent is good” argument

> *“5% are succeeding. People are trying AI for just about everything right now. 5% is pretty damn good, when AI clearly has a lot of room to get better.”*
>
> *“What percentage of all software projects started in the past 12 month will be cancelled by the end of 2027?”*

A commenter on the MIT thread, August 2025, and a commenter on the Gartner thread, June 2025[50,51]

My answer. Fair, and I do not know the base rate for ordinary software projects either; neither does Gartner, which does not publish one. What is different is the cost of the 95 percent. A failed internal tool costs its budget. A failed agent that reached customers costs the budget, the reversal and the trust, and section 4 lists five companies that paid all three. The right comparison is not the cancellation rate but the cost of a cancellation.

The "vendors oversold it" argument

> *"Rough order of magnitude you can expect to deflect around 30% of inbound contacts with an AI chatbot. [...] The problem is that vendors are telling companies they can eliminate 80-90% (maybe even 100% if they can keep a straight face) of their customer service agent jobs with AI, and that is nonsense."*
>
> *"Investors aren't interesting in hearing 'we tried vibe coding but got tired of reviewing and fixing the trash code it produced'. No, they want to hear 'we've reduced our headcount by 40% while increasing customer satisfaction by 60%'."*

A contact-centre AI practitioner and another commenter on the Klarna thread, May 2025[49]

My answer. The first comment is the most useful sentence in this paper for anyone buying a customer-service agent: a practitioner's base rate, stated with its error bars. The second explains why the base rate rarely reaches the board. Klarna's 700-agent claim and its reversal fifteen months later are the two halves of the same incentive.

The comment I would hand to any product owner came from the MIT thread: "I'm arriving at the conclusion that deployments of LLMs is most suitable in areas where the cost of false positives and, crucially, false negatives are low." (Hacker News comment, August 2025.)[50] That is the whole site-selection problem in one sentence.

7.2 The investors

The investor argument began before the agent wave. In June 2024 Sequoia's David Cahn wrote that "AI's $200B question is now AI's $600B question," the gap between infrastructure spending and the revenue needed to justify it,[40] and Goldman Sachs titled a research note "Gen AI: too much spend, too little benefit?", observing that roughly a trillion dollars of planned capital spending "has little to show for it so far."[41]



The people funding the agents have started saying the word bubble out loud. Several are still buying.

Photo: Thomas J. O'Halloran, Library of Congress, Public domain, via Wikimedia Commons

By August 2025 the builders agreed with the sceptics about the mood, if not the outcome. Sam Altman: "Are we in a phase where investors as a whole are overexcited about AI? My opinion is yes." And: "Someone is going to lose a phenomenal amount of money."[42] In January 2026 Ray Dalio described "the AI boom that is now in the early stages of a bubble."[44] In October 2026 Panmure Liberum's Joachim Klement said his "core conviction is that the AI bubble will either burst in 2027 or in 2028," with a correction that could be the most severe since

the global financial crisis;[45] Dalio called the market a "classic bubble" close to bursting, and Temasek's chief investment officer called an unwinding of the AI trade "the biggest risk to markets in 2027";[52] Citadel's Ken Griffin said investors were "almost certainly" in a bubble while noting that "Any radical transformation of our economy is generally intertwined with a bubble" and, in Decrypt's paraphrase, that corporate America had been uneven at extracting value from AI;[46] and a report by the Open Markets Institute and the Americans for Financial Reform Education Fund called the boom "primed to pop," singling out the industry's "circular financing and purchase agreements."[53]

> *"When bubbles happen, smart people get overexcited about a kernel of truth."*
>
> Sam Altman, in an interview reported by *The Verge*, August 2025[42]

For a company deciding whether to fund an agent project, the investor debate reduces to one operational fact. If the people selling you the model are themselves calling the market overexcited, the burden of proof for your business case cannot be "everyone is doing it." It has to be the number in section 5.

08 FAIRNESS

What is working, and what has not changed



The lift is real. So is the gas bill, and so is the need for a crew.

Photo: Jean-Marie Hullot, CC BY 3.0, via Wikimedia Commons

It would be dishonest to stop at the failures. Several things have genuinely improved since 2024, and the same sources that document the stalls document them.

The capability curve is real and steep

METR's measured task horizon has kept doubling, faster since 2023 than before, and its February 2026 note judged it "likely that developers are more sped up from AI tools now" than in early 2025.[12,25] Sinha and colleagues showed that thinking models mitigate self-conditioning and can execute much longer tasks in a single turn.[13] A chain that failed last year may hold this year.

Individual productivity gains are consistent across surveys

McKinsey: 80 percent of respondents say AI improved their individual productivity. Deloitte: 28 percent of enterprises now report revenue gains from AI, up from 20 percent in January, and 81 percent report faster decision-making. Stack Overflow: 66 percent of developers use coding assistants or agents, and 70 percent reach for an AI agent first when tackling a problem.[7,22,27]

The 5 percent have a recognizable shape

MIT's successful buyers "demand process-specific customization and evaluate tools based on business outcomes rather than software benchmarks." McKinsey's high performers "fundamentally redesign workflows that are enabled by AI rather than insert AI into existing ones." KPMG's organizations with full cost visibility report established returns five times as often. Even Taco Bell processed two million orders before the one with 18,000 waters.[8,7,20,31] Narrow scope, measured outcome, redesigned process, known cost: that is the pattern, and section 9 is built on it.

What has not changed

Reliability on repeated runs has barely moved while accuracy has.[14] The share of companies reporting profit impact has not moved in a year.[7] Two-thirds of the organizations KPMG surveyed still cannot see what their AI costs to run.[20] And the models still cannot tell data from instructions, which is why an agent with broad permissions remains a security problem whatever its accuracy.[39]

The honest summary is that the technology is improving faster than the organizations deploying it, and that the gap between the two is where the money is being lost. Closing it is an engineering and management problem, not a modelling one.

09 DESIGN

How to build an agent that ships

Since reliability multiplies and cost adds, the design goal is a chain short enough to hold, checked at every link, priced before it runs, and allowed to act alone only where a wrong action is cheap or reversible. These are the layers I put in place.

None of this is exotic. It is the discipline every other production system already has, applied to a component that is probabilistic and metered. The surveys say the organizations that do it are the ones reporting returns: process-specific scope, outcomes over benchmarks, redesigned workflows, visible costs and named accountability.[8,7,20]

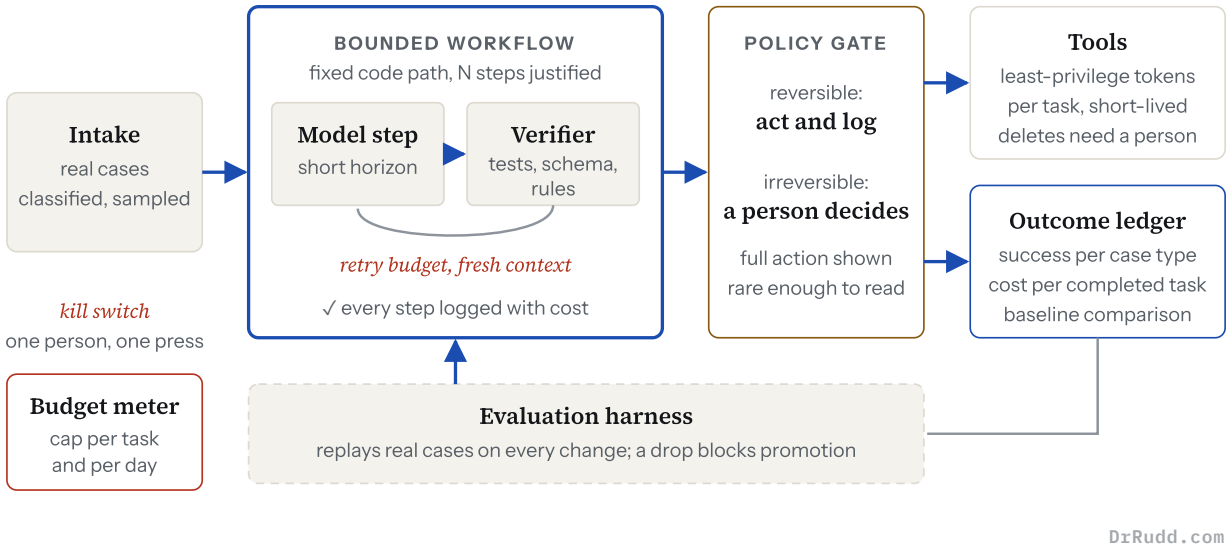


A prototype earns its flight envelope one test at a time. An agent earns its autonomy the same way.

Photo: NASA / Tom Tschida, Public domain, via Wikimedia Commons

An agent deployment that can be audited

FIGURE 5 PRACTICE



The architecture I recommend. The model is a component inside a workflow, not the workflow; every step it takes is verified, metered and logged; it cannot reach an irreversible action without a person; and the evaluation harness, not a demo, decides whether a change ships. The containment layer around the tools is the subject of my MCP paper.[39]

9.1 Start as a workflow, and make agency earn its place

Anthropic's advice stands: "the simplest solution possible," with agency added "only when needed."[10] Gartner's observation that "many use cases positioned as agentic today don't require agentic implementations" is the same point from the buyer's side.[5] In practice: write the workflow as ordinary code with the model called at the specific steps that need judgement, and let the model choose its own next step only where the branches are genuinely unknowable in advance. Each place the model is allowed to decide the control flow is a place the chain can lengthen without anyone noticing.

9.2 Bound the horizon and verify every link

The exponent is the enemy, so the first design question is the number of steps, and the second is what checks each one. A verifier can be a unit test, a schema, a reconciliation against a second source, a rules engine, or a cheap second model asked only to find faults. It does not have to be perfect; it has to turn a silent failure into a loud one. Give each step a retry budget, and start the retry from a clean context rather than appending to the failed one, because a context that contains the error makes the next error more likely.[13]

PRACTICE

The invoice agent of stall 2 becomes seven short workflows with a deterministic spine. The match to the purchase order is checked against the ledger before the tax step runs; the tax step is a rules engine, not a model; the approval step is a human for anything over a threshold; the posting step cannot run unless every earlier verifier passed. The chain the model actually controls is two steps long, and the 83 percent becomes 97 percent with a known 3 percent routed to a person.

9.3 Build the evaluation harness before the agent

A test set drawn at random from real cases, replayed against every change, is the instrument that replaces the demo. Run each case several times, because the Princeton work shows that consistency across runs, not accuracy on one, is what separates a product from a prototype.[14] Report success per case type, keep the worst output of every run where a person can read it, and make the harness the gate on promotion: a change that lowers end-to-end success does not ship, whatever it does to a benchmark. This is the third pillar of review from my earlier paper, applied to a system that acts rather than answers.[54]

9.4 Price every task, and meter it while it runs

The calculator in section 5 is run before the pilot and then replaced by the real figures: the outcome ledger records cost per completed task, including retries and human rework, next to the baseline. KPMG found that just over half of organizations have cost reviews in approval processes or cost dashboards, 40 percent have token budgets, and those with full cost visibility report established returns five times as often.[20] A budget cap per task and per day is not bureaucracy; it is the only thing that stops a looping agent from spending the quarter's savings in an afternoon.

9.5 Make the human the exception handler, not the gate

Route by confidence and by consequence. Results the verifiers passed and that are cheap to undo go through; results that failed a check, or that would change something expensive, go to a person with the complete action in front of them, not a summary. Approvals should be rare enough to be read; a reviewer who sees two hundred a day is the rubber stamp that Replit and PocketOS show the cost of. And for irreversible actions, the control is mechanical: a token without the delete permission, a delayed delete, a separate environment. An instruction in English is a request.[28,29]



Autonomy is a ladder. An agent climbs one rung at a time on measured consistency, and comes down a rung on the first incident.

Photo: Sundaram Perumal, CC BY-SA 4.0, via Wikimedia Commons

9.6 Earn autonomy in rungs

Five levels, promoted on evidence from the ledger and demoted on the first incident: the agent *suggests* and a person acts; the agent *drafts* and a person edits and sends; the agent *acts with approval* on each action; the agent *acts and reports*, with a person reviewing a sample; the agent *acts alone*, within a budget and a scope. Most deployments that pay today sit on the second and third rungs. Gartner's 15 percent even considering the fifth is, on this evidence, about right.[6]

9.7 Write the baseline and the kill criteria on day one

One sentence, signed by the business owner: the metric, its value today, the value that justifies production and the date. Add the second sentence most pilots never write: the result at which the project stops. Gartner's advice to CFOs to judge "realized value rather than deployment volume" is the same rule.[26] A pilot with kill criteria can fail honestly, which is the only way it can succeed honestly.

9.8 Contain it, because accuracy is not safety

An agent that is right 99 percent of the time and holds a token that can delete production is a 1 percent chance of deleting production per task. Containment is a launch condition, not a later phase: least-privilege tokens issued per task, no session that combines untrusted input, private data and an open exit, and an inventory with a named owner and a kill switch. The companion paper on MCP gives the architecture and the checklist.[39]

10 REVIEW

A go or no-go checklist for any agent going into production

Each item is checkable, and most can be read straight off the outcome ledger. A project that cannot answer any of the first six is a pilot, whatever its name.



Pilots run the checklist on every flight, not just the first. So should an agent, on every change.

Photo: U.S. Navy / Photographer's Mate 2nd Class Lou Rosales, Public domain, via Wikimedia Commons

Baseline signed

The metric, its value today, the production threshold and the date, in one sentence signed by the business owner before the pilot.

Real-case test set

A random sample of last quarter's actual cases, held out, with the worst case types named.

End-to-end, repeated

Task success measured end to end over at least ten runs per case, so that consistency is a number.

Worst output reviewed

The worst result of the pilot, on one page, with the business owner's written answer to what happens if it reaches a customer.

Cost per completed task

Measured with retries and human rework included, next to the cost of the old way.

Kill criteria written

The result at which the project stops, agreed before it starts.

Longest chain justified

The number of model-controlled steps in the workflow, and why it cannot be shorter.

Verifier at every step

Each model step is followed by a check that turns a silent failure into a loud one, with a retry budget and a fresh context.

Irreversible actions gated

Every action that cannot be undone is listed, and each is blocked mechanically without a person: permission, delay or environment.

Autonomy rung assigned

The agent's level on the ladder is stated, with the evidence required for promotion and the rule for demotion.

Live meter and cap

Cost per task on a dashboard someone reads, with a budget cap per task and per day that stops the run.

Owner, log, kill switch

A named owner of outcomes, a log that reconstructs any action, a scope of credentials no wider than the task, and a stop one person can press.

11 ACCOUNTABILITY

Who is responsible for what

The team building the agent

Designs the chain and its verifiers, builds the evaluation harness first, and reports end-to-end success per case type, not demo accuracy.

The platform team

Runs the meter, the budget caps, the environments, the token issuance and the kill switch as production infrastructure.

Risk and security

Owns containment, approves each promotion up the autonomy ladder, and reviews every incident before the agent climbs back.

The business owner

Signs the baseline and the kill criteria, reads the worst output, and accepts in writing the cost per completed task that production will run at.

The last role is the one the surveys say is missing. KPMG found returns three times as common where accountability for AI outcomes is clearly defined, and Deloitte found nearly one in four leaders unable to say who answers when AI goes wrong.[20,22] Somebody with authority has to decide, in writing, which tasks an agent is allowed to do alone, at what cost, and what happens on its worst day.

12 PUSHBACK

Common objections, and how I answer them

These are the objections I hear most from executives and engineering leaders. Most of them start from something true.

"The models improve so fast that this will fix itself."

Capability is improving fast. Reliability and cost are not following it.

METR's task horizon is doubling in months.[12] Princeton found the same period's accuracy gains "only yielded small improvements in reliability,"[14] and more capable agents spend more tokens, not fewer.[9] Design for today's reliability; the chain gets shorter when the model gets better.

"Our vendor's demo worked on our own data."

On the cases they chose, once. Ask for a random sample of yours, ten times.

MIT's interviewees had "seen dozens of demos this year" and found "maybe one or two" useful.[8] Salesforce's own benchmark found its agents at 58 percent on single turns and 35 percent on realistic multi-turn work.[16]

"We'll keep a human in the loop."

Doing what? If everything, the agent saves nothing. If a summary, the human is a rubber stamp.

Replit and PocketOS both had a human in the loop.[28, 29] Developers, who use agents most, trust them for important decisions 6.6 percent of the time.[27] Make the person the exception handler, and make irreversible actions impossible without one.

"Everyone else is doing it. We will fall behind."

Behind on what? Name the metric and the answer usually changes.

Nearly half of the 2,145 organizations KPMG surveyed have already rephased agent deployments over cost,[20] and the people selling the models call the market "overexcited."[42] The competitor to fear is the one running the calculator, not the one with the most pilots.

"The 95 percent figure is overblown."

It is a small, self-reported sample, and MIT says so. The other surveys say the same thing anyway.

MIT calls its numbers "directionally accurate."[8] McKinsey's 1,719 respondents put profit impact at a flat 37 percent and high performers at 6 percent;[7] S&P found 42 percent abandoning most initiatives;[23] KPMG found half rephasing.[20] Four samples, one direction.

"This is just the trough of disillusionment. The internet had one too."

Agreed, and the companies that walked out of the last trough were the ones with unit economics.

Griffin and Altman say the same: a bubble around "a kernel of truth."[46, 42] The trough is exactly when cost per completed task decides who survives it. That is an argument for section 5, not against it.

"Agents are just software. Our normal project discipline applies."

Normal software is deterministic: test it once and it passes forever. An agent is a distribution, metered per run.

Its test is a sample replayed repeatedly, its cost is paid on every attempt, and it fails silently by default. That needs the third pillar of review from my earlier paper, not just functional and non-functional checks.[54]

"If it is a bubble, we should wait."

Wait on autonomy. Do not wait on workflows.

The 5 percent exist, and they look like narrow scope, redesigned process and visible cost.[8,7] A bounded workflow with a measured baseline pays now and teaches you what the next rung needs. What to wait on is the fifth rung of the ladder, which 85 percent of IT leaders are already waiting on.[6]

13 IN BRIEF

Agentic AI in production, answered briefly

Why do most agentic AI projects fail?

Because an agent is a chain of steps and reliability multiplies along it while cost adds up. Per-step accuracy that looks excellent in a demo becomes a low end-to-end success rate on real work, every attempt is billed, and most pilots never defined the number they were meant to move. Gartner names escalating costs, unclear business value and inadequate risk controls; MIT names tools that do not learn.[5,8]

What percentage of AI agent projects reach production?

It depends on the survey and the definition. MIT found 5 percent of task-specific tools reached production in its 52-organization sample; Gartner expects over 40 percent of agentic AI projects to be cancelled by the end of 2027; McKinsey finds 40 percent of large companies scaling agents but only 37 percent of all respondents reporting any profit impact from AI.[8,5,7]

How reliable are AI agents on real tasks?

On benchmarks built from real work, the best agents completed 30 percent of simulated company tasks and about 35 percent of multi-turn business scenarios. Capability on long

tasks is doubling every few months, but consistency across repeated runs has improved only slightly.[15,16,12,14]

How much do AI agents cost to run?

Agents use about four times the tokens of a chat and multi-agent systems about fifteen times, by Anthropic's measurement, and every retry pays again. Nearly half of the organizations KPMG surveyed have rephased agent deployments because costs outran value, and only about a third can see their AI running costs in full. The number to compute is cost per *completed* task; section 5 has the calculator.[9,20]

Is agentic AI a bubble?

The people funding and building it increasingly say so: Altman called investors "overexcited," Dalio sees "the early stages of a bubble," and Griffin says "almost certainly." All three also say the technology is real. For a company, the practical reading is that the business case has to stand on its own measured economics.[42,44,46]

What is the difference between an AI agent and a workflow?

In Anthropic's definitions, a workflow orchestrates models and tools "through predefined code paths," while an agent lets the model "dynamically direct" its own process and tool use. Workflows are cheaper, more reliable and easier to audit; agents trade cost and latency for flexibility, and should be used only where the branches are genuinely unknowable in advance.[10]

How do you get an AI agent into production?

Sign a baseline and kill criteria first; build the evaluation harness from real cases before the agent; bound the chain and verify every step; price every task and meter it live; make the human the exception handler and gate irreversible actions mechanically; and promote autonomy one rung at a time on evidence. Section 9 and the checklist give the detail.

14 GLOSSARY

Plain-language terms

AI agent

Workflow

Software in which a language model decides its own next steps and which tools to call, rather than following a fixed script.

A fixed sequence of steps written in ordinary code, with a model called at particular points for judgement.

Agentic AI

The marketing and industry term for systems built around AI agents.

Agent washing

Gartner's term for rebranding chatbots, assistants or automation as agents without substantial agentic capability.

Task horizon

How long a task, measured in human working time, an AI can complete with a given probability; METR's standard measure.

Compounding error

The way a chain's success rate falls as per-step success rates multiply together.

Self-conditioning

A model becoming more likely to err when its own earlier errors are in its context.

Reliability

Whether an agent gets the same result on repeated runs, holds up under small changes, and fails predictably; distinct from accuracy on one run.

Evaluation harness

A held-out set of real cases replayed against every change, used to decide whether the change ships.

Baseline

The value of the business metric before the agent, measured the same way it will be measured after.

Kill criteria

The result, agreed before a pilot, at which the project stops.

Cost per completed task

Everything spent to get one task actually done: all attempts, retries and human review, divided by the tasks that succeeded.

Token

The unit of text a model reads and writes, and the unit most model usage is billed in.

Human in the loop

A person who reviews or approves what an agent does; a control only if the agent cannot act without them.

Irreversible action

Anything the agent can do that cannot be undone: sending, paying, deleting, publishing.

Autonomy ladder

The five levels from suggesting to acting alone, climbed one rung at a time on measured evidence.

15 SOURCES

References

Numbered in the order they are first cited. Every link was checked for this version when it was published. Survey and vendor figures are cited as the publishing firm's own claim, with the sample stated; where a primary source was not reachable, the reporting outlet is named.

1 Altman, S. "Reflections." blog.samaltman.com, January 2025. COMMENTARY

blog.samaltman.com https://blog.samaltman.com/reflections

2 Goldman, S. "Exclusive: Marc Benioff has declared a 'hard pivot' to autonomous AI agents. Will it be enough for Salesforce to thrive in the generative AI era?" *Fortune*, 5 September 2024. REPORTING

fortune.com https://fortune.com/2024/09/05/salesforce-ceo-marc-benioff-ai-agents-agentforce-dreamforce-gen-ai-era

3 Gartner. "Gartner Hype Cycle Identifies Top AI Innovations in 2025." Press release, 5 August 2025. ANALYST FORECAST

gartner.com https://www.gartner.com/en/newsroom/press-releases/2025-08-05-gartner-hype-cycle-identifies-top-ai-innovations-in-2025

4 Ehsan, O. I. "Hyperscalers Hit $700 Billion in 2026 AI Spending Plans," reporting first-quarter 2026 company guidance and a Reuters Morning Bid segment. Yahoo Finance, 1 May 2026. REPORTING

finance.yahoo.com https://finance.yahoo.com/sectors/technology/articles/hyperscalers-hit-700-billion-2026-111243744.html

5 Gartner. "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027." Press release, 25 June 2025. ANALYST FORECAST

gartner.com https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

6 Gartner. "Gartner Survey Finds Just 15% of IT Application Leaders Are Considering, Piloting, or Deploying Fully Autonomous AI Agents." Press release, 30 September 2025. Survey of 360 IT application leaders, May to June 2025. SURVEY

gartner.com https://www.gartner.com/en/newsroom/press-releases/2025-09-30-gartner-survey-finds-just-15-percent-of-it-application-leaders-are-considering-piloting-or-deploying-fully-autonomous-ai-agents

7 Tinkoff, D., Van der Veken, L., and Chui, M., with Balakrishnan, T. *The state of AI in 2026: On the road to ROI*. McKinsey Global Survey, 1,719 participants, 4 May to 8 June 2026. McKinsey & Company, August 2026. SURVEY

mckinsey.com https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai Report (PDF) https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/the-state-of-ai-in-2026-on-the-road-to-roi.pdf

8 Challapally, A., Pease, C., Raskar, R., and Chari, P. *The GenAI Divide: State of AI in Business 2025*. MIT NANDA, July 2025. Review of over 300 public AI initiatives, interviews with 52 organizations and 153 survey responses. RESEARCH
Report (PDF) https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

9 Anthropic. "How we built our multi-agent research system." 13 June 2025. COMPANY DISCLOSURE
anthropic.com https://www.anthropic.com/engineering/built-multi-agent-research-system

10 Anthropic. "Building effective agents." 19 December 2024. COMPANY DISCLOSURE
anthropic.com https://www.anthropic.com/research/building-effective-agents

11 METR. "Measuring AI Ability to Complete Long Software Tasks." 19 March 2025. RESEARCH
metr.org https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

12 METR. "Time Horizon 1.1." 29 January 2026. RESEARCH
metr.org https://metr.org/blog/2026-1-29-time-horizon-1-1/

13 Sinha, A., Arun, A., Goel, S., Staab, S., and Geiping, J. "The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs." arXiv:2509.09677, version 3, March 2026; published at ICLR 2026. RESEARCH
arxiv.org https://arxiv.org/abs/2509.09677

14 Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S., and Narayanan, A. "Towards a Science of AI Agent Reliability." arXiv:2602.16666, version 3, June 2026; accepted at ICML 2026. RESEARCH
arxiv.org https://arxiv.org/abs/2602.16666

15 Xu, F. F., Song, Y., Li, B., et al. "TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks." arXiv:2412.14161, version 3, September 2025. RESEARCH
arxiv.org https://arxiv.org/abs/2412.14161

16 Huang, K.-H., Prabhakar, A., Thorat, O., et al. "CRMArena-Pro: Holistic Assessment of LLM Agents Across Diverse Business Scenarios and Interactions." arXiv:2505.18878, May 2025. RESEARCH
arxiv.org https://arxiv.org/abs/2505.18878

17 Cemri, M., Pan, M. Z., Yang, S., et al. "Why Do Multi-Agent LLM Systems Fail?" arXiv:2503.13657, version 3, October 2025. RESEARCH
arxiv.org https://arxiv.org/abs/2503.13657

18 Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. “AI Agents That Matter.” arXiv:2407.01502, July 2024. RESEARCH

arxiv.org https://arxiv.org/abs/2407.01502

19 Gartner. “Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025.” Press release, 29 July 2024. ANALYST FORECAST

gartner.com https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

20 KPMG International. *Global AI Pulse Q2 2026: From deployment to value realization*. Survey of 2,145 senior leaders across 20 countries and territories. June 2026. SURVEY

Report (PDF) https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2026/06/global-ai-pulse-q2.pdf

21 KPMG LLP. “AI Investment and Agent Deployment Hold Steady Amid Growing Focus on Pragmatism.” AI Quarterly Pulse Survey, Q2 2026, 204 US leaders at companies with revenue of 1 billion dollars or more. Press release, 24 June 2026. SURVEY

kpmg.com https://kpmg.com/us/en/media/news/q2-ai-pulse-2026.html

22 Deloitte. “Enterprise AI Enters an ROI Reckoning: More Use, Uneven Value, Reveals Deloitte Survey.” State of AI in the Enterprise, 3,500 leaders in 25 countries, May to June 2026. PR Newswire, 29 October 2026. SURVEY

prnewswire.com https://www.prnewswire.com/news-releases/enterprise-ai-enters-an-roi-reckoning-more-use-uneven-value-reveals-deloitte-survey-302923168.html

23 Wilkinson, L. “AI project failure rates are on the rise: report,” reporting S&P Global Market Intelligence research on more than 1,000 respondents. CIO Dive, 14 March 2025. SURVEY

ciodive.com https://www.ciodive.com/news/AI-project-fail-data-leaders-strategy/742590/

24 METR. “Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.” 10 July 2025. RESEARCH

metr.org https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

25 METR. “We are Changing our Developer Productivity Experiment Design.” 24 February 2026. RESEARCH

metr.substack.com https://metr.substack.com/p/2026-02-24-uplift-update

26 Gartner. “Gartner Says CFOs Must Stop Mistaking Finance AI Deployment for Value Creation.” Press release, 28 May 2026. Survey of 204 finance leaders, March 2026. SURVEY

gartner.com https://www.gartner.com/en/newsroom/press-releases/2026-05-28-gartner-says-cfos-must-stop-mistaking-finance-ai-deployment-for-value-creation

27 Stack Overflow. “Stack Overflow’s 2026 Developer Survey Reveals Positive Sentiment Towards AI, but Trust Remains Conditional.” Press release, 6 October 2026, and the survey’s AI trust data (n = 14,304). SURVEY

stackoverflow.co https://stackoverflow.co/company/press/archive/stack-overflow-annual-developer-survey-ai-trust-conditional Trust data https://survey.stackoverflow.co/2026/ai/data/ai-trust

28 Nolan, B. “An AI-powered coding tool wiped out a software company’s database, then apologized for a ‘catastrophic failure on my part’.” *Fortune*, 23 July 2025. INCIDENT

fortune.com https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/

29 ABC News. “‘Rogue’ AI agent from Cursor goes haywire, deletes company’s entire database.” 30 April 2026. INCIDENT

abc7ny.com https://abc7ny.com/post/rogue-ai-agent-cursor-goes-haywire-deletes-tech-companys-database/19008099/

30 Farrant, T. “An AI agent deleted a company’s entire database in 9 seconds, then wrote an apology.” Euronews, 28 April 2026, updated 10 June 2026. INCIDENT

euronews.com https://www.euronews.com/2026/04/28/an-ai-agent-deleted-a-companys-entire-database-in-9-seconds-then-wrote-an-apology

31 McCallum, S. “Taco Bell rethinks AI drive-through after man orders 18,000 waters.” BBC News, 29 August 2025. INCIDENT

bbc.com https://www.bbc.com/news/articles/ckgyk2p55g8o

32 Klarna. “Klarna AI assistant handles two-thirds of customer service chats in its first month.” Press release, 27 February 2024. COMPANY DISCLOSURE

klarna.com https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/

33 Doerer, K. “Klarna changes its AI tune and again recruits humans for customer service,” reporting a Bloomberg interview with Sebastian Siemiatkowski. CX Dive, 9 May 2025. REPORTING

customerexperiencedive.com https://www.customerexperiencedive.com/news/klarna-reinvests-human-talent-customer-service-AI/747586/

34 Dhanji, K. “Deloitte to pay money back to Albanese government after using AI in $440,000 report.” *The Guardian*, 6 October 2025. INCIDENT

theguardian.com https://www.theguardian.com/australia-news/2025/oct/06/deloitte-to-pay-money-back-to-albanese-government-after-using-ai-in-440000-report

35 Gartner. “Gartner Predicts 30% of Employees Replaced by AI Will Be Rehired by 2029 if Organizations Don’t Overhaul Workforce Planning.” Press release, 28 October 2026. Survey of 297 CHROs, July 2026. ANALYST FORECAST

gartner.com https://www.gartner.com/en/newsroom/press-releases/2026-10-28-gartner-predicts-30-percent-of-employees-replaced-by-ai-will-be-rehired-by-2029-if-organizations-dont-overhaul-workforce-planning

36 OpenAI. "OpenAI and Hugging Face partner to address security incident during model evaluation." 21 July 2026, updated 28 and 29 July and 26 August 2026. COMPANY DISCLOSURE

openai.com https://openai.com/index/hugging-face-model-evaluation-security-incident/

37 Claburn, T. "OpenAI explains how its naughty AI agents attacked Hugging Face," reporting OpenAI's technical report. *The Register*, 27 August 2026. REPORTING

theregister.com https://www.theregister.com/security/2026/08/27/openai_explains_how_its_naughty_ai_agents_attacked_hugging_face/5292780

38 Painter, C. "Chris Painter's testimony to the U.S. Senate on AI agent incidents." METR, 30 September 2026. Hearing "Rogue AI: Securing the Homeland Against AI Agent Attacks," Senate Committee on Homeland Security and Governmental Affairs subcommittee. TESTIMONY

metr.org https://metr.org/blog/2026-09-30-chris-painter-senate-testimony/

39 Rudd, I. "Leaked Without a Breach: How the Model Context Protocol Lets AI Agents Give Your Data Away, and How to Architect Around It." Working paper, version 1.0, 1 October 2026. WORKING PAPER

drrudd.com https://drrudd.com/papers/mcp-security/

40 Cahn, D. "AI's $600B Question." Sequoia Capital, 20 June 2024. COMMENTARY

sequoiacap.com https://www.sequoiacap.com/article/ais-600b-question/

41 Goldman Sachs Global Investment Research. "Gen AI: too much spend, too little benefit?" Top of Mind, 25 June 2024. COMMENTARY

goldmansachs.com https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit

42 Roth, E. "Sam Altman says 'yes,' AI is in a bubble." *The Verge*, 15 August 2025. REPORTING

theverge.com https://www.theverge.com/ai-artificial-intelligence/759965/sam-altman-openai-ai-bubble-interview

43 Deloitte AI Institute. *The State of AI in the Enterprise*, 2026 edition. Survey of 3,235 leaders, August to September 2025. SURVEY

deloitte.com https://www.deloitte.com/global/en/issues/ai/state-of-ai-in-enterprise.html

44 Bove, T. "Ray Dalio says AI is in 'the early stages of a bubble,' so watch out for 2026." *Fortune*, 6 January 2026. REPORTING

fortune.com https://fortune.com/2026/01/06/ray-dalio-says-ai-is-in-the-early-stages-of-a-bubble-so-watch-out-for-2026

45 Stamm, L., and Jaisinghani, S. “AI Bubble Risks Worst S&P 500 Crash Since 2008, Strategist Says.” Bloomberg, via Yahoo Finance, 8 October 2026. REPORTING
finance.yahoo.com https://finance.yahoo.com/markets/stocks/articles/ai-bubble-risks-worst-p-061638441.html

46 Lanz, J. A. “Ken Griffin Says AI Is ‘Almost Certainly’ in a Bubble, and Compares GPUs to Airline Seats.” Decrypt, 21 October 2026. REPORTING
decrypt.co https://decrypt.co/381578/ken-griffin-ai-bubble-gpus-airline-seats

47 Hacker News. Discussion of “Measuring the impact of AI on experienced open-source developer productivity.” Posted 10 July 2025; 775 points and 485 comments when checked on 4 November 2026. FORUM
news.ycombinator.com https://news.ycombinator.com/item?id=44522772

48 Hacker News. Discussion of “Deloitte to refund the Australian government after using AI in $440k report.” Posted 7 October 2025; 464 points and 230 comments when checked on 4 November 2026. FORUM
news.ycombinator.com https://news.ycombinator.com/item?id=45500485

49 Hacker News. Discussion of “Klarna changes its AI tune and again recruits humans for customer service.” Posted 11 May 2025; 257 points and 125 comments when checked on 4 November 2026. FORUM
news.ycombinator.com https://news.ycombinator.com/item?id=43955374

50 Hacker News. Discussion of “95% of generative AI pilots at companies are failing : MIT report.” Posted 18 August 2025; 230 points and 167 comments when checked on 4 November 2026. FORUM
news.ycombinator.com https://news.ycombinator.com/item?id=44941118

51 Hacker News. Discussion of “Gartner Predicts over 40% of Agentic AI Projects Will Be Canceled by End of 2027.” Posted 30 June 2025; 30 points and 12 comments when checked on 4 November 2026. FORUM
news.ycombinator.com https://news.ycombinator.com/item?id=44421803

52 Rao, P. “Investors warn of AI bubble as tech stocks reach record highs.” Semafor, via Yahoo Finance, 7 October 2026. REPORTING
finance.yahoo.com https://finance.yahoo.com/markets/stocks/articles/investors-warn-ai-bubble-tech-103212585.html

53 Reed, B. “‘Working People Will Pay a High Price’: Report Warns AI Bubble Is ‘Primed to Pop’,” reporting a study by the Open Markets Institute and the Americans for Financial Reform Education Fund. Common Dreams, 21 October 2026. REPORTING
commondreams.org https://www.commondreams.org/news/ai-bubble-no-bailout

54 Rudd, I. “Wrong Without Warning: Why AI Systems Need a Third Pillar of Review.” Working paper, version 1.0, 24 February 2026. WORKING PAPER

drrudd.com https://drrudd.com/papers/third-pillar/

16 CITATION

How to cite this paper

Please cite the stable URL below. It will not move. If the paper is revised, the version number changes and earlier versions stay listed here, so a citation always points at the text it quoted.

APA (7TH EDITION)

Rudd, I. (2026, November 4). *Pilots without landings: Why most agentic AI projects stall before production, what users and investors now say, and how to architect the ones that ship* (Working paper, Version 1.0). https://drrudd.com/papers/agentic-ai-production/

CHICAGO

Rudd, Ian. 2026. “Pilots Without Landings: Why Most Agentic AI Projects Stall Before Production, What Users and Investors Now Say, and How to Architect the Ones That Ship.” Working paper, version 1.0, November 4, 2026. https://drrudd.com/papers/agentic-ai-production/.

IEEE

I. Rudd, “Pilots without landings: Why most agentic AI projects stall before production, what users and investors now say, and how to architect the ones that ship,” Working paper, ver. 1.0, Nov. 4, 2026. [Online]. Available: https://drrudd.com/papers/agentic-ai-production/

BIBTEX

```
@misc{rudd2026agentic,
  author       = {Rudd, Ian},
  title        = {Pilots Without Landings: Why Most Agentic {AI}
Projects Stall Before Production, What Users and Investors Now Say, and
```

```
How to Architect the Ones That Ship},
  howpublished = {Working paper, version 1.0},
  year         = {2026},
  month        = nov,
  day          = {4},
  url          = {https://drrudd.com/papers/agentic-ai-production/},
  note         = {ORCID: 0000-0002-8750-285X}
}
```

VERSION	DATE	CHANGE
1.0	4 November 2026	First publication.



About the author

Ian Rudd, PhD, is a chief enterprise AI architect based in Canada. He has spent close to twenty years taking machine learning from the research lab into production, in federal government, banking, insurance, retail and transport, and in corporate AI research at IBM and Microsoft. His work is making AI systems that hold up in front of an auditor.

← drrudd.com ORCID Google Scholar LinkedIn Get in touch

https://drrudd.com/papers/agentic-ai-production/ · Version 1.0 · 4 November 2026

Image credits

Photographs and the painting are from Wikimedia Commons, used under the licences shown, resized and in some cases cropped for this page. They are illustrative: none depicts the specific events, products or organizations described beside it, except the Dreamforce entrance and the Taco Bell drive-through, which show the companies named. Icons are from Lucide (ISC licence). Diagrams and charts are drawn for this paper.

Landscape with the Fall of Icarus (painting): Pieter Bruegel the Elder (attributed), Royal Museums of Fine Arts of Belgium, Public domain, Pieter Bruegel the Elder - Landscape with the Fall of Icarus - Brussels, Royal Museums of Fine Arts of Belgium - Google Arts & Culture.jpg

Dreamforce 2022 entrance: InvadingInvader, CC BY-SA 4.0, Dreamforce 2022 Entrance.jpg

Dominoes falling: Kurt:S, CC BY 2.0, Dominoes falling.jpg

Orbiter model in a wind tunnel: NASA / Dennis Keim, Public domain, Orbiter Model in Wind Tunnel - GPN-2000-000056.jpg

Taxi meter, Mexico City: Josh Baumgartner, CC BY-SA 4.0, Taxi meter CDMX 20170517A.jpg

Ordnance Survey cut mark: Adrian Dust, CC BY-SA 2.0, Ordnance Survey Cut Mark - geograph.org.uk - 7709331.jpg

Security queues, Schiphol Airport: Ymblanter, CC BY-SA 4.0, Schiphol Airport Queues to the Security Control June 2022.jpg

Taco Bell drive-through at night: SnappingTurtle, CC BY-SA 4.0, Taco-bell-drive-thru.jpg

Etch A Sketch: jeanbaptisteparis, CC BY-SA 2.0, Etch-a-sketch (760966766).jpg

Emergency stop button on a CNC machine: Flygklubben, CC BY-SA 4.0, Emergency stop button on a HAAS CNC.jpg

Antique cash register, Darjeeling: Subhrajyoti07, CC BY-SA 4.0, Antique cash register in a cafe, Darjeeling.jpg

Speakers' Corner, Hyde Park: Anthony O'Neil, CC BY-SA 2.0, Speakers' Corner, Hyde Park - geograph.org.uk - 7615441.jpg

New York Stock Exchange trading floor: Thomas J. O'Halloran, Library of Congress, Public domain, NY stock exchange traders floor LC-U9-10548-6.jpg

Hot-air balloon being inflated, Bagan: Jean-Marie Hullot, CC BY 3.0, ဓလာဓလာ2006.jpg

Helios Prototype in its hangar: NASA / Tom Tschida, Public domain, The Helios Prototype flying wing stretches almost the full length of the 300-foot-long hangar at NASA's Dryden Flight Research Center, Edwards, California (EC02-0031-6).jpg

Painters on a scaffold with a ladder: Sundaram Perumal, CC BY-SA 4.0, Painting Time.jpg

Pilot running a pre-flight checklist: U.S. Navy / Photographer's Mate 2nd Class Lou Rosales, Public domain, US Navy 040229-N-8055R-014 U.S. Navy Pilot Lt. Cmdr. Erik Ostrom, of San Leandro, Calif., goes through a pre-flight checklist before a flight.jpg

© 2026 Ian Rudd. Stable address: drrudd.com/papers/agentic-ai-production