AI agents in 2027: task cost will not fall like token cost

Forecast: coding agents stay in production, computer use stays on runbooks, and a cheaper token does not buy a cheaper successful run.

Last updated November 10, 2026

Executive summary

The useful question for 2027 is not whether another model demo will clear a benchmark. It is whether a successful agent run gets cheaper at anything like the rate a benchmark question has, and whether the firms that already have a grader turn that into profit. This page forecasts that the run does not get cheaper at the question's rate, and that profit attribution stays a minority outcome. Both sentences are forecasts. The kill-criteria are in the last chapter before the methodology.

The base, measured in 2026, is uncomfortable for anyone selling a transformation. McKinsey's survey, fielded May 4 to June 8, found large companies pulling ahead on agent scaling (40 percent of respondents at firms above $1 billion in revenue, up from 27 percent) while the share attributing any EBIT impact to AI stayed at 37 percent and the high-performer share stayed at about 6 percent. Eighty percent said their own productivity improved. BCG's CEOs, surveyed for a January 15 report, were already expecting the payoff: about 90 percent said agents would produce measurable returns in 2026, and they had committed more than 30 percent of that year's AI investment to agentic AI. The belief and the EBIT line are both dated. They do not describe the same object. [1][4]

Epoch AI estimates that the price of reaching a given benchmark score has fallen about 47 percent a quarter since 2023, roughly 13 times a year, with their own list of reasons not to treat the figure as exact. That is the price of a question. On Terminal-Bench 4.0, two Codex runs at max effort both score 58.18 percent. GPT-6.1 Sol's listed run cost is $634.24. GPT-6 Astra's is $3,267.18, 5.2 times as much. A buyer who banks the 13-times question-price decline and ignores the loop will not recognize the invoice. [3][5]

Regulation does not need a forecast to be on the 2027 calendar. The Commission timeline dates December 2, 2026, August 2, 2027, and December 2, 2027. The MCP specification dated July 28, 2026 makes Roots, Sampling, Logging, and Dynamic Client Registration eligible for removal in a revision released on or after July 28, 2027. Eligibility is not deletion. The twelve-month report is the grid of those dates next to the company filings. This page does not add product launches to it. [6][7]

A forecast is not a launch calendar

Two pages on this site already list what is dated between November 2026 and November 2027. The six-month report stops at May 2027. The twelve-month report carries the EU sandbox date, the MCP eligibility window, and the company filings that actually name a day. Those pages sort claims into announced, likely, and watch-for, and they leave a month blank when nobody has dated it. This page does not refill the blanks.

What this page does is say what we think the measured 2026 pattern implies for how agents get bought and built through 2027, and what evidence would make that wrong. The implications are marked as forecasts in the chapter that states them. A sentence that cites McKinsey, Menlo, Epoch, BCG, a leaderboard, the Commission, or the MCP spec is a measurement or a document, with the date on the source. If a sentence has no source number and talks about 2027, it is a forecast.

BCG's January 15, 2026 AI Radar is the clean example of an expectation that should not be promoted. The survey covered 2,360 executives in 16 markets and nine industries, including 640 CEOs. About 90 percent of those CEOs said agents would produce measurable returns in 2026. CEOs had committed more than 30 percent of their organization's 2026 AI investment to agentic AI. Ninety-four percent of companies planned to keep investing even if the near-term return did not show up. Seventy-two percent of CEOs said they were the

main decision maker on AI, twice the prior year's share. Four in five said they were more optimistic about ROI than a year earlier. Half said their job depended on getting AI right. Every one of those is a belief held in January about the year ahead, by people whose job, on their own account, is now tied to the outcome. [4]
McKinsey's August survey is the check that was available before this page was written. Agent scaling had risen at large companies. The EBIT line had not. A belief surveyed in January and a profit line surveyed in the summer are not a contradiction until you pretend they measured the same thing. They did not. The forecast in this page is about whether 2027 repeats that split.

- Buyers: When a 2027 roadmap arrives, mark each row as a date someone has published or as a forecast, including ours. The twelve-month report is the list of published dates. A row that is only a quarter and a verb is a forecast, and it owes you the same kill-criteria this page writes down for itself.

The profit gap is the thing 2027 has to close

McKinsey's high performer is a respondent who attributes at least 5 percent of EBIT to AI and calls the impact significant. About 6 percent of respondents qualify, the same share as the year before. Any EBIT impact, a weaker bar, is 37 percent, also unchanged. Self-reported productivity is 80 percent. The chart puts the three shares in one frame so the distance is obvious. They are different questions. Plotting them on one axis is a comparison of magnitudes, not a claim that they partition the sample. [1]
The mechanism people reach for is 'adoption has not turned into redesign.' That sentence is too soft to use. The survey already separates the respondents who say they are scaling agents from the respondents who attribute profit, and the second group did not grow in the year the first group did, at least among large companies. High performers are twice as likely as others to say they are scaling coding agents, and about three times as likely to say coding-agent costs constrained them. The firms closest to a real deployment are the firms reporting that the loop is expensive. A 2027 plan that assumes EBIT appears because more teams 'have agents' is the plan this survey just ran, with a flat result.
Menlo's model is the demand side of the same year: deployment spend estimated at $176.8 billion by September 2026, coding and chatbots the bulk of the application market, frontier labs taking 69 percent of deployment dollars. Spend can scale while the share of companies attributing profit does not. A category can be large and still be a cost center for most of the buyers in it. The 6 percent is not a forecast that the market is small. It is a forecast-relevant fact that the market being large has not, so far, meant the typical buyer can point at EBIT. [2]
Forecast, not a percentage. We do not publish a 2027 figure for the high-performer share. The directional call is that it stays a small minority through 2027 unless buyers instrument cost per successful task and keep a person on the steps a grader cannot see. 'Small minority' means we expect another survey with a comparable question to print a number a reader would still call a minority. We are not hiding a private point estimate.

- Buyers: A 2027 budget that cites BCG's 90 percent and skips McKinsey's 6 percent is citing the hope and skipping the measurement. Fund the workflows where you can already name the grader. Treat a company-wide agent rollout as a cost, and review it on cost per successful task, until the EBIT attribution is something your own finance system can see.

Three different questions from the same survey, plotted together so the gap is visible. They are not slices of one pie.

	Share
Any EBIT impact	37%
High performers	about 6%
Own productivity up	80%

A cheaper token is not a cheaper task

Epoch AI's estimate is the best public account of how fast a benchmark question has been getting cheaper, and it comes with the caveats in the authors' own words. Since 2023, the cost of reaching a given level of performance has fallen about 47 percent a quarter, or about 13 times a year. The drop is slower on game-like puzzles, 39 to 43 percent a quarter, and faster on math, 50 to 52 percent. Fresh state-of-the-art performance falls faster still, about 66 percent a quarter, and a level that has been around for two years falls about half as fast, 32 percent a quarter. Their worked example: on January 31, 2025, o3 could reach their adjusted 75 percent bar on GPQA Diamond for an average of about 30 cents a question. Under 18 months later, GPT-5.6 Luna matched that bar for four hundredths of a cent, $0.0004, a 725-fold drop. The page carries a September 23, 2026 update stamp. [3]

Epoch lists the reasons the 13-times figure is not a law. Labs may train to the tests. A benchmark score is not useful work. Their frontier is the cheapest model that hits the score, which assumes a buyer who switches constantly. Real buyers do not. The series is about three years long, the sample of model-benchmark pairs is incomplete, and several reasonable averages exist. They say the bottom-line numbers are representative and should not be read as exact. This page takes them as stated, including that warning. It does not tighten them.

An agent run is not a question. It is a loop: a plan, a tool call, an observation, a retry, and a stop. Each step writes tokens back into the window. Context engineering is the choice of which of those tokens survive to the next step. Compact the tool output and the next call is cheap. Resend the trace and the token-price decline is spent on history. The Terminal-Bench rows are the public demonstration that the loop dominates the model card. Codex, effort max, 58.18 percent, 192 of 330: GPT-6.1 Sol is listed at $634.24, GPT-6 Astra at $3,267.18. Same agent, same score, 5.2 times the listed bill. Per recorded success that is $3.30 against $17.02. Opus 5.5 with Claude Code, a different harness and a higher score (64.85 percent, 214 of 330), lists $4,718.65, or $22.05 per recorded success. The third bar is there so nobody ranks the Codex pair against Claude Code as if the harness were held still. [5]

Forecast: through 2027, the price of a successful agent run falls, and it falls by less than Epoch's question-price rate. We do not name the rate. The reasoning is the one the rows already force. A 13-times decline in the price of a scored question is diluted by retries, by tool output left in the window, and by buyers who do not switch models every time the frontier price moves. Epoch already warns that buyers do not switch. An agent harness switches even less often, because the harness is code, evals, and permissions, not a drop-down. The teams that compact context and cache the runbook keep more of the saving. The teams that resend the trace do not. McKinsey's high performers, who are the ones reporting coding-agent costs as a constraint, are the buyers for whom this distinction is already the budget.

- Builders: Put cost per successful task on the same row as the eval, and split it into tokens, retries, and cache hits. A 2027 model-card price cut is not your discount until that row moves. Buyers should renegotiate the agent contract on that unit. A vendor who will only quote the per-million-token price is quoting

Epoch's object, not the invoice.

Percent decline per quarter in the price of reaching a given score. These are Epoch's estimates for benchmark questions. The lower end of each published range is plotted where they gave a range.

	Decline per quarter
All five benchmarks	about 47%
Game puzzles, low end	39%
Math, low end	50%
Just-achieved SOTA	66%
Two years after SOTA	32%

Four dates are already written down

Most of what will be called '2027 agent news' is a product post with no day on it. Four dates are different, because a public document names the day and says what happens. They were read on November 7, 2026 for the twelve-month report. This page repeats them so a reader who never opens that grid still has the constraints. It does not add interpretations the timeline does not support. [6][7]

December 2, 2026 is close enough that it is an operating fact, not a 2027 strategy item. The timeline's new prohibitions and the Article 50(2) transition land then. August 2, 2027 is a duty on member states to have a sandbox, not a launch window for vendors. December 2, 2027 is the Annex III high-risk date. An agent that takes a decision in one of those high-risk categories needs a design that assumes the rule is in force on that day. This page does not forecast an enforcement wave, a delay, or a court outcome. The date is the date on the timeline.

The MCP date is the one harness engineers will trip on, because it looks like a breaking-change deadline and it is not. The July 28, 2026 specification deprecates Roots, Sampling, Logging, and Dynamic Client Registration. Under the feature-lifecycle policy, a deprecated feature becomes eligible for removal in the first revision released on or after a twelve-month mark, here July 28, 2027. The registry, read November 7, said no feature had been removed under that policy yet. The policy also has an expedited-removal exception. Eligibility plus an exception is not 'these four features disappear on July 28.' A harness that depends on Sampling, which is how a server asks the client's model to do work, needs a migration plan before the eligibility date. It does not need a war room on the morning of it.

- Builders: Put July 28, 2027 on the MCP dependency you actually call, and December 2, 2027 on any agent that makes a high-risk decision under Annex III. Those are engineering dates. They are not a reason to rewrite a coding agent that never touches Sampling and never takes a regulated decision. The twelve-month report has the surrounding filings, including Anthropic's compute language and Apple's risk-factor language. This page does not restate them.

Dates a public document already names. None of these is a product launch, and none is a forecast from this page.

Date	What the document says	Source
Dec 2, 2026	New prohibitions, and an Article 50(2) transition for some providers already on the market before Aug 2, 2026.	[EU AI Act timeline](https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act)
Jul 28, 2027	Roots, Sampling, Logging, and Dynamic Client Registration become eligible for removal. Removal is a later maintainer decision.	[MCP deprecated features](https://modelcontextprotocol.io/specification/2026-07-28/deprecated)
Aug 2, 2027	Member states should have at least one national AI regulatory sandbox operating.	[EU AI Act timeline](https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act)
Dec 2, 2027	Annex III high-risk rules. Rules for high-risk AI embedded in Annex I products are dated Aug 2, 2028, outside this note.	[EU AI Act timeline](https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act)

The forecast

Three calls, each with the evidence that would have to stay true and the limit that keeps it from being a number.

Coding agents remain the default production deployment

Forecast. Through 2027, the agent a company can defend in a review is still a coding agent, or a close cousin whose output a test or a diff can reject. The evidence is the concentration already measured: Menlo's model puts $34.0 billion of application spend in coding and describes that category as 96 percent AI-native; McKinsey finds coding-agent scaling at 31 percent of large-enterprise respondents and finds those same respondents using the tools to refuse a software purchase; the public boards that can be graded, SWE-bench Verified and Terminal-Bench, are software tasks. The limit: Menlo's dollars are an estimate, McKinsey's shares are what respondents said, and a public board is not your repository. A private monorepo with no tests does not inherit the 76.8 percent. [1][2]

Computer use stays on runbooks

Forecast. Through 2027, computer-use deployments that survive contact with a bill are cached, repeatable UI paths with a check at the end. We do not forecast that OSWorld 2.0 binary completion joins the SWE-bench 70s. The best binary row copied on November 9 is 20.6 percent for Claude Opus 4.8 on the 500-step primary metric, with a 54.8 percent partial score beside it. Andreessen Horowitz's August note cites 85 percent on OSWorld-Verified, a different board, and in the same piece says agents break when the work leaves the runbook. The limit: a new harness, or a narrower official metric, could move the binary rate without producing a general desktop employee. A higher score on a benchmark whose tasks are themselves runbooks would confirm this forecast, not refute it. [8]

The headcount expectation misses again

Forecast. The 39 percent of McKinsey respondents who expect an AI-related decline in total employment over the coming year will turn out high relative to the share who later report that a decline happened. The evidence is the same question's last outcome: 32 percent expected a decline, 14 percent of respondents at organizations using AI now report that AI contributed to one. Forty-three percent this year expect no change. We do not forecast a percentage, a sector, or a company's layoff. One miss is not a long series. A recession, or a single large employer's disclosure, can break a

survey pattern without breaking the technology story. The call is only that this instrument's expectation is a poor predictor that planning headcount off the 39 percent repeats the error. [1]

What we are not forecasting: a 2027 market size, a token-price path of our own, a date for a named product, a winner among Anthropic, OpenAI, and Google, or a change in the 6 percent high-performer share precise enough to print. Menlo's lab-share figures (Anthropic 55 percent, OpenAI 31 percent, Google 9 percent of enterprise API spend in their estimate) describe 2026. Extending them another year would be a story about distribution and coding quality, and this page does not have the 2027 invoices.

- Builders and buyers: Staff 2027 for graded coding work and for a small number of computer-use runbooks. Keep a person on the approval the grader cannot see. Do not staff a general agent employee, and do not staff a headcount reduction because 39 percent of survey respondents expect one. Builders who sell into this should lead with the eval and the cost per success. The category they are selling into is the one Menlo's dollars and McKinsey's scaling already picked.

The first bar is the share who expected an AI-related workforce decline in the prior survey. The second is the share now reporting that one happened. The third is this year's expectation for the coming year.

	Share
Expected last year	32%
Reported this year	14%
Expect next year	39%

What would make this forecast wrong

A forecast without a kill-criteria is a slogan. These are the observations that would retire the calls above. They are specific enough to check, and they do not depend on this site being the one that checks them.

Computer use. A public eval with a named harness, a binary grader, and a task count in the hundreds, printing a completion rate in the same range as the SWE-bench Verified lead (the mid-70s on the November 5 copy), on a setting a vendor documents as the production default rather than a contest configuration. A higher score on OSWorld-Verified alone does not count, because that board is already the one quoted at 85 percent while OSWorld 2.0's binary row sits at 20.6 percent. The two would have to be the same setting, or the vendor would have to say which one they run.

Profit. A successor to the McKinsey survey, with the high-performer question recognizable as the same question, in which that share is no longer a small minority while respondents still report agent scaling. We are not writing down the number that would count. If a reader looks at the new printout and would still call it a minority, the call stands. If the printout has clearly left that range, the call is wrong. A vendor case study does not substitute for the survey.

Task price. Agent-loop invoices, not single-question prices, falling on the order of Epoch's 13-times-a-year figure for a fixed success rate on a fixed eval. The invoice has to be the loop: harness, retries, and tool output included. A blog post about the price of a million tokens does not count, including Epoch's, which we are already using for the question and not for the loop. If a published series of agent-run prices on something like Terminal-Bench or an internal eval with a stable grader shows that kind of decline through 2027, the cost forecast is wrong and the builders who did nothing but wait for the model card were right.

The calendar. MCP maintainers removing Roots, Sampling, Logging, or Dynamic Client Registration before July 28, 2027 would falsify the reading that the date is eligibility rather than deletion. The lifecycle policy's expedited-removal exception is the path that would do it. We are not forecasting that they use the exception. An EU timeline change that moves the August 2, 2027 sandbox date or the December 2, 2027 Annex III date would falsify the 'already written' chapter, which would then need to be rewritten against the new timeline rather than defended.

Headcount. A follow-up in which the share reporting an AI-related workforce decline lands near the 39 percent who expected one, on a comparable question, would falsify the claim that this instrument overshoots. A single company's layoff does not. The claim is about the survey, not about every employer.

- Buyers: Put these four checks in the quarterly review: the computer-use setting you actually run, cost per success on your eval, whether finance can see EBIT, and whether the MCP features you call still exist. If the checks move, update the plan. If they do not, the 2027 forecast on this page is still the one the evidence supports.

Methodology

This page is a forecast built on other people's measurements. It did not field a survey, convene a panel, or fit a model of the agent market. There is no unpublished dataset behind the directional calls. Where this page divides two printed numbers, the division is labeled.

What was used

McKinsey's 37 percent, about 6 percent, 80 percent, 40 percent, 27 percent, 22 percent, 31 percent, 32 percent build-versus-buy context, the threefold cost constraint on coding agents among high performers, and the workforce trio of 32, 14, and 39 percent, plus the 43 percent who expect no change. The field dates, the sample of 1,719, the 97 countries, the GDP weights, and the $1 billion cut are theirs. [1]

Menlo's $176.8 billion, $34.0 billion, $25.0 billion, 96 percent AI-native coding share, and 69 percent frontier-lab share of deployment dollars. These are estimates from their November 3, 2026 report, not recounts. BCG's 90 percent, 30 percent, 94 percent, 72 percent, four in five, and half, from the January 15, 2026 release, n = 2,360 including 640 CEOs. Epoch's quarterly decline rates, the 13-times yearly figure, and the GPQA price example, from the article carrying a September 23, 2026 update stamp, with their caveats kept. [2][3][4]

Terminal-Bench costs and accuracies from the November 5, 2026 copy. The ratio 5.15 is $3,267.18 divided by $634.24. Per-success figures divide listed cost by the successes field. Opus 5.5 is included in that chart and excluded from the 'same harness' claim. EU and MCP dates are those recorded on November 7, 2026 from the Commission timeline and the MCP deprecated-feature registry. OSWorld 2.0's 20.6 percent and 54.8 percent are the November 9 copy. The 85 percent and about 72 percent human figure are Andreessen Horowitz's, on OSWorld-Verified, August 10, 2026. [5][6][7][8]

What was not done

No 2027 revenue number was estimated. No probability was attached to the three calls. No expert was invented, and no credential is claimed for the author line beyond the site's research byline. Deloitte's January 2026 survey was not used as an agent statistic, because its fieldwork is August and September 2025 and its questions are about AI programs. The Stanford Index is used in the production report for the SWE-bench wording and is not a source of the forecasts here. Product rumors without a date are out. They live, if they live anywhere, as blanks on the twelve-month grid.

The production cross-section, which has the charts this forecast compresses, is AI agents in production, 2026. Cite that page for the 2026 measurements if you want the full table. Cite this page only for the forecast and the kill-criteria.

- Builders and buyers: If you repeat a sentence from the forecast chapter, repeat the word forecast. If you repeat a percentage, repeat the source number. The two are not interchangeable.

Sources

[1] McKinsey, The state of AI in 2026: On the road to ROI. August 25, 2026. Survey fielded May 4 to June 8, 2026, n = 1,719, 97 countries, weighted by national GDP. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] Menlo Ventures, 2026: The State of Enterprise AI. November 3, 2026. Deployment and application figures are Menlo's estimates, annualized run rate. https://menlovc.com/perspective/2026-the-state-of-enterprise-ai/

[3] Epoch AI, The plunging price of thought. Luke Emberson and David Roodman. The page carries a September 23, 2026 update stamp. Figures below are their estimates for benchmark-question prices, not agent loops. https://epoch.ai/publications/the-plunging-price-of-thought

[4] BCG, AI Radar 2026. Press release January 15, 2026, and the article at https://www.bcg.com/publications/2026/as-ai-investments-surge-ceos-take-the-lead. Survey of 2,360 executives, including 640 CEOs. https://www.bcg.com/press/15january2026-as-ai-investments-surge-ceos-take-lead

[5] Terminal-Bench 4.0, as copied on this site. Board updated-at stamp 2026-10-23. Copy read November 5, 2026 for the harness benchmark. https://hub.harborframework.com/datasets/terminal-bench/terminal-bench/4?leaderboard=4-0-0&tab=leaderboard
[6] European Commission, AI Act implementation timeline. Dates below are the ones recorded from this timeline, including the Digital Omnibus, for the November 7, 2026 twelve-month report on this site. https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act
[7] Model Context Protocol, deprecated features and lifecycle. July 28, 2026 specification. Lifecycle policy: https://modelcontextprotocol.io/community/feature-lifecycle. Read November 7, 2026. https://modelcontextprotocol.io/specification/2026-07-28/deprecated
[8] OSWorld 2.0 and the Andreessen Horowitz computer-use note. OSWorld 2.0 rows copied November 9, 2026. a16z note, August 10, 2026: https://a16z.com/can-agents-use-a-computer-yet-weve-got-the-data/ https://osworld-v2.xlang.ai/
[9] AI agents in production, 2026, on this site. The cross-section this forecast starts from. Survey figures and board rows are cited to their publishers above, not to that page. /reports/ai-agents-in-production-2026