

WHITE PAPER NO. 1

Cost per Accepted Outcome

The unit economics of AI workforces



Manelink White Paper No. 1 · October 2026

Content type: White Paper · **Primary topic:** Agent economics and AI workforce unit economics · **Audience:** Executives, operators, finance and technology leaders · **Reading time:** about 25 minutes

In one sentence: Token prices are collapsing, yet agent bills and project cancellations are rising, because the cost of AI work is set less by the price of a token than by how many tokens an attempt burns, how often attempts fail, and how much human checking each result needs.

How to read the evidence in this paper. Every claim carries a label so you can see how much weight it bears.

- **Established evidence:** peer-reviewed or large, well-documented studies with consistent findings.
- **Emerging evidence:** recent preprints, new benchmarks or single studies; directionally useful, not settled.
- **Industry observation:** analyst forecasts, consulting analysis and practitioner surveys.
- **Manelink analysis:** our own synthesis, frameworks and illustrative models. These are proposals, not established academic models.

Contents

Executive summary	3	5. Business and organizational implications	15
1. The emerging shift: cheap tokens, expensive work	4	6. Recommendations	16
2. What the research says: three multipliers between the token and the bill	6	7. Future outlook	18
3. Analysis: why cheaper tokens do not reach the P&L	9	8. Conclusion	19
4. The Manelink framework: Cost per Accepted Outcome	10	Checklist: is this agent workflow economically ready?	20
		References	21

Executive summary

For three years the economics of AI have been told as a story about falling prices. That story is true and incomplete. The price of reaching a fixed level of model capability has fallen by between 9× and 900× a year depending on the task ^[1], and analysts expect the cost of serving a trillion-parameter model to fall more than 90% between 2025 and 2030 ^[5]. Yet one in five organizations now says AI operating costs are constraining their use of AI ^[9], more than 40% of agentic AI projects are forecast to be cancelled by the end of 2027 ^[6], and Gartner now expects the inference cost of each agentic workflow to rise more than fivefold through 2028 ^[7].

These facts are not in conflict. They describe a change in the unit of work. A chatbot answers a question; an agent completes a task. Completing a task means planning, calling tools, re-reading context, checking its own work, trying again and, often, waiting for a person to approve the result. Each of those steps multiplies the token count, and several of them add costs that never appear on an AI invoice at all.

This paper argues that executives should stop managing AI by the price of a token and start managing it by the **cost per accepted outcome**: the fully loaded cost of one unit of finished, approved work, including compute, retries, human review, errors that slip through and the fixed cost of running the system. We propose a simple model for that number, show what the research says about each of its parts, and work through an illustrative example.

20–25%

of an agent's variable run cost was tokens

70–75%

was human oversight

A banking customer-service workflow analysed by McKinsey ^[10]

Key findings

- **Token deflation is real, but agentic work consumes tokens faster than prices fall.** Agentic coding tasks use about 1,000× more tokens than code chat or single-turn reasoning ^[12], and the cost of running frontier-level models on hard benchmarks is rising 3–18× a year even as price per unit of performance falls ^[2].

● Established evidence
● Emerging evidence

- **Listed prices are a poor guide to real cost.** In 32% of model comparisons, the model with the lower list price cost more to complete the same work, by up to 28× ^[13].
- **Reliability, not capability, is the binding constraint.** Twenty-four months of rapid capability gains produced only small improvements in agent reliability ^[14], and production teams respond by keeping agents on a short leash: 68% of deployed agents run 10 steps or fewer before a human steps in ^[17].
- **Human verification is usually the largest controllable cost.** Redesigning where and how people approve agent work moved our illustrative cost per accepted outcome four times more than halving token prices.
- **Memory, skills and routines pay off when they are reused, not when they are rebuilt.** Online memory modules that rebuild knowledge on every task often fail to beat a plain agent given the same token budget ^[21], while reuse of stable context through caching can cut input cost by up to 90% ^[23].

● Emerging evidence

● Established evidence

● Emerging evidence

● Manelink analysis

● Emerging evidence

● Industry observation

What to do: make cost per accepted outcome the unit of account, measure real cost distributions rather than list prices, scope agent work to the horizon where it is reliable, route by default and gate frontier models, reuse stable context, design the approval layer by risk, and review agent economics every quarter.

01

The emerging shift: cheap tokens, expensive work

The price of intelligence is collapsing

On almost any measure, a given level of AI capability is far cheaper than it was. Epoch AI tracked the cheapest model able to reach fixed scores on six benchmarks and found prices falling between 9× and 900× a year, with a median of about 50× ^[1]. A separate study, built on the largest dataset of benchmark prices assembled to date, found the price of a given level of performance falling about 5–10× a year for frontier models, and estimated that about 3× a year of that comes from algorithmic efficiency alone ^[2].

● Established evidence

Analysts expect the trend to continue. Gartner forecasts that by 2030, inference on a trillion-parameter model will cost providers more than 90% less than in 2025 ^[5].

● Industry observation

The bill is not following

The same researchers who documented falling prices also found the opposite trend at the frontier: the cost of running state-of-the-art models on demanding benchmarks has been rising 3–18× a year, because each new increment of capability demands disproportionately more inference ^[2].

● Emerging evidence

Agents sit squarely on that frontier. In Anthropic's account of its production research system, a single agent used about 4× the tokens of a chat interaction and a multi-agent system about 15×; token usage alone explained 80% of the performance differences in its evaluation ^[4]. Gartner's estimate for enterprise use is a range of 5–30× more tokens per task than a standard chatbot ^[5].

● Industry observation

The most detailed measurement to date comes from a 2026 study of eight frontier models on real software-engineering tasks. Agentic coding consumed roughly 1,000× more tokens than code chat or single-turn code reasoning, and input tokens, not output, drove the cost, because agents keep re-reading the context they accumulate ^[12].

● Emerging evidence

1,000× more tokens for agentic coding than for code chat or single-turn reasoning ^[12]

The business results are flat

Adoption has not been the problem. In McKinsey's 2026 global survey of 1,719 respondents, 40% of large organizations reported scaling AI agents, up from 27% a year earlier ^[9]. But the share reporting any enterprise-level EBIT impact from AI was essentially unchanged at 37%, and the share of "AI high performers" stayed at about 6% ^[9]. About one in five respondents said AI operating costs, including tokens, had constrained their AI use ^[9]. Stanford's 2026 AI Index likewise finds agent deployment still in single digits across most business functions, even though 88% of organizations use AI in some form ^[3].

● Industry observation

Finance teams have noticed. In the FinOps Foundation's 2026 survey of 1,192 practitioners managing more than US\$83 billion in annual spend, 98% now manage AI spend, up from 31% two years earlier; the top challenges they cite are seeing AI costs, allocating them to business units and determining ROI [8]. Gartner has named the pattern the "Inference Paradox": better unit economics escalating the overall cost of AI without a clear path to commensurate value [7].

● Industry observation

● **Manelink view.** The paradox resolves once you notice that the unit has changed. Organizations are buying cheaper tokens and then spending them on a more expensive product: completed, multi-step work. Tracking token prices to manage agent costs is like tracking the price of steel to manage the cost of a building.

02

What the research says: three multipliers between the token and the bill

If the token is the wrong unit, what sits between it and the bill? The research points to three multipliers, plus a fourth factor that only scale can fix.

2.1 The volume multiplier: tokens per attempt

Agents spend unpredictably. On the same software task, repeated runs of the same agent differed by up to 30× in total tokens, and higher spending did not buy higher accuracy: accuracy peaked at intermediate cost and saturated beyond it, with expensive runs showing repetitive re-reading and re-editing of the same files [12]. When asked to predict their own token usage before starting, frontier models systematically underestimated it, with correlations no better than 0.39 [12].

● Emerging evidence

List price misleads. A 2026 study of eight reasoning models across 12 tasks found that in 32% of model pairs, the model with the lower listed price cost more in practice, by as much as 28× [13]. One model whose list price was 80% lower than a competitor's cost 38% more across all tasks, because it used far more "thinking" tokens and more rounds of tool use. Repeated runs of the same query varied by up to 9.7× in thinking tokens [13].

● Emerging evidence

What drives volume. Three things dominate: how much context an agent accumulates and re-reads, how many steps it takes, and how much it "thinks" per step. Each is a design choice as much as a model property.

2.2 The reliability multiplier: attempts per accepted result

Capability is rising quickly. METR's measure of the length of task an AI agent can complete with 50% success has doubled roughly every seven months since 2019, and faster since 2023, roughly every 4.3 months on its updated methodology [16]. Agents on OSWorld, a test of real computer tasks, went from about 12% to about 66% success in a year [18].

● Established evidence

Reliability is rising slowly. Princeton researchers decomposed reliability into consistency, robustness, predictability and safety, evaluated 15 models released over 24 months, and found that rapid capability gains had produced only small improvements in reliability [14]. Agents that could solve a task often failed to solve it consistently. Their conclusion is directly economic: an agent that succeeds on 90% of tasks but fails unpredictably on the rest may be a useful assistant yet an unacceptable autonomous system [14].

● Established evidence

The gap shows up wherever researchers repeat tasks. On τ -bench, a customer-service benchmark, a leading model succeeded on about 61% of retail tasks in a single try but on only about 25% when it had to succeed on all eight repeated attempts [15]. On TheAgentCompany, a simulated software company, the best agent fully completed 30.3% of tasks, taking an average of 27 model calls and US\$4.20 per task [19].

● Established evidence

Newer results show how fast the frontier is moving, and why price per task can mislead. A 2026 re-run of the WorkBench workplace benchmark reported that the best agent's task completion rose from 43% in 2024 to 97.7% in 2026 [20]. In the same evaluation, a small model cost about US\$0.03 per task but completed 58.8% of tasks and caused harmful side effects, such as emails sent to the wrong person, on 30.3% of them; a frontier model cost about US\$0.21 per task, completed 94.9% and caused side effects on 3.9% [20].

● Emerging evidence

The cheaper model in that comparison was about **8× cheaper per attempt** and about **8× more likely to cause a side effect** [20]. Per attempt it wins; per accepted outcome it may not.

2.3 The verification multiplier: human time per result

Oversight is the dominant variable cost. McKinsey's analysis of agentic workflows found that for a customer-service task in banking, tokens represented 20–25% of an agent's variable run cost, and human oversight by risk and functional experts 70–75% ^[10]. In banking customer onboarding, it expects 10–20% of agent runs to be reviewed by experts ^[10].

● Industry observation

Production teams keep humans close. The first systematic study of production agents, based on 20 case studies and 86 practitioners across 26 domains, found that 68% of agents execute at most 10 steps before human intervention, 70% use off-the-shelf models with prompting rather than fine-tuning, 74% rely primarily on human evaluation, and reliability is the top development challenge ^[17].

● Established evidence

68% of production agents run 10 steps or fewer before a person steps in ^[17]

Unchecked output moves cost downstream. In a survey of 1,150 US desk workers, about four in ten said they had received low-quality AI-generated work ("workslop") in the previous month; each incident took an average of 1 hour 56 minutes to resolve, an estimated US\$186 per worker per month ^[24]. The figures are self-reported, but the mechanism matters: when verification is skipped upstream, it is paid for downstream at a higher rate.

● Industry observation

Review can erase the gain. In a randomized trial of experienced open-source developers using early-2025 AI tools, tasks took 19% longer with AI, even though developers believed they were 20% faster; time shifted from writing code to prompting, waiting for and reviewing AI output ^[25]. METR has since noted that these historical results no longer reflect current tools ^[25], but the lesson about review overhead stands. By contrast, where AI suggestions slot into a well-scoped workflow, gains are large: access to an AI assistant raised customer-support productivity by 14% on average and 34% for novice workers ^[26].

● Established evidence

2.4 The amortization factor: fixed cost per outcome

Agents carry fixed costs that per-task thinking ignores: infrastructure, orchestration, monitoring, evaluation and the people who maintain them. McKinsey reports that production agents often need adjustment every

● Industry observation

couple of days as models, protocols and business requirements change ^[10].

Volume changes the arithmetic. In one McKinsey example, a conversational agent onboarding 2,500 customers a year cost US\$10,000–15,000; doubling the volume raised the cost only to US\$15,000–20,000 ^[10]. In a fully loaded bank account-opening workflow involving five to seven agents and two to four oversight teams, the total cost to complete an onboarding could fall from about US\$50–150 per customer to about US\$10–30 ^[10].

● Industry observation

03

Analysis: why cheaper tokens do not reach the P&L

This section is Manelink analysis built on the evidence above.

● Manelink analysis

The bill moves rather than shrinks. When organizations deploy agents, costs leave one budget and land in three others. Some appear on the AI invoice as tokens. Some appear in payroll as review time. Some appear later, as rework, customer remediation or incidents. A finance team that watches only the AI invoice will see the smallest of the three and conclude, wrongly, that the economics are either wonderful or out of control, depending on whether token prices happened to fall that quarter.

Token price is a small lever on a large number. Because tokens are often a minority of variable cost ^[10], even dramatic price declines move the total modestly. Halving the cost of every token in our illustrative workflow below reduces cost per accepted outcome by less than 9%. Raising first-pass acceptance from 70% to 85% does more. Redesigning the review layer does far more.

Under 9%

saved per accepted outcome by halving every token price in the worked example. Redesigning the approval layer saves four times as much. Manelink analysis

Reliability compounds through everything. A failed attempt does not just waste its own tokens. It wastes the reviewer's time, triggers a retry that burns more tokens and more review, and, if the failure is not caught, creates

an escaped error whose cost can dwarf all of the above. This is why the reliability research matters to executives: consistency is not an academic property; it is the denominator in your unit cost.

The cheapest model is often not the cheapest option. Price reversal [13], unpredictable token spend [12] and side-effect rates that differ by an order of magnitude [20] all point the same way: compare models on cost per accepted outcome in your own workflow, never on list price.

Scaffolding is not free either. Memory, skills and multi-agent orchestration can raise success rates, but they consume tokens, and those tokens are rarely counted. A 2026 study gave a plain web agent the same token budget that memory- and skill-augmented agents consumed and let it simply take more steps. Across three models and three domains, the plain agent matched or beat all three augmentation methods while often using fewer tokens [21]. The augmentations paid twice: once for their own model calls, and again because every retrieved memory or skill inflated the agent's prompt on every step. Worse, on shopping tasks roughly half or more of the lessons one memory system had labelled as successes came from tasks that had actually failed [21].

● Emerging evidence

Where the evidence conflicts. Earlier work found that agents which induce and reuse workflows from past tasks improved success by 24.6% and 51.1% relative on two web benchmarks while taking fewer steps [22]. The 2026 budget-matched study does not refute this; it shows the gains can vanish once total token cost is held constant and stronger base models are used [21]. The reconciliation, in our reading: reuse pays when knowledge is curated once and used many times, as with stable routines and cached context, and is weakest when it is rebuilt on every task without quality control. ●

04

The Manelink framework: Cost per Accepted Outcome

Proposed framework. Cost per Accepted Outcome (CPAO) is a Manelink management model derived from the research above. It is not an established academic standard.

● Manelink analysis

4.1 Definition

An accepted outcome is one unit of finished work that meets acceptance criteria defined before the agent is deployed: a reconciled invoice, a resolved deduction dispute, a closed support case, a merged code change.

"Accepted" means it passed whatever check the business requires, automated or human.

Cost per accepted outcome is the fully loaded cost of producing one such unit:

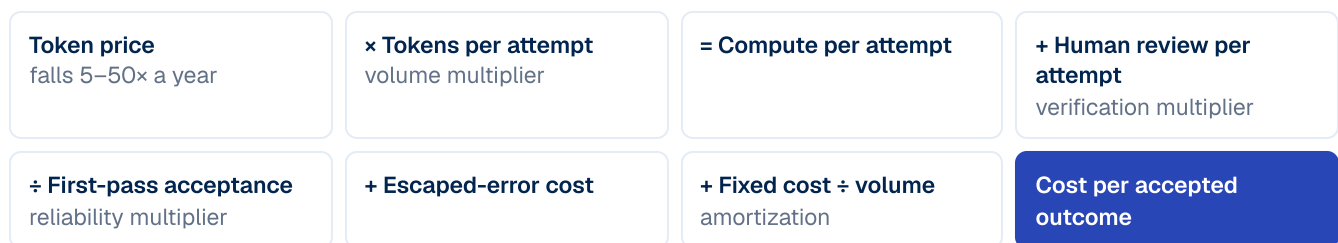
$$\text{CPAO} = \frac{C_{run} + r \cdot t_{rev} \cdot w}{a} + e \cdot L_{esc} + \frac{F}{V}$$

Where:

- C_{run} is compute and tool cost per attempt (tokens × price, plus tool and API calls)
- r is the share of attempts a person reviews, t_{rev} the minutes per review, and w the loaded cost per minute
- a is the first-pass acceptance rate, so $1/a$ is the expected number of attempts per accepted outcome
- e is the escape rate (unacceptable results that pass review) and L_{esc} the average cost of one escaped error
- F is annual fixed cost and V annual volume of accepted outcomes

The model makes a simplifying assumption: failed attempts are detected and retried. Where failures go undetected, they show up in the escape term instead.

4.2 The CPAO stack



4.3 What drives each term, and which lever moves it

Term	What drives it	Main levers	Through the AI-workforce lens
Compute per attempt	Context size and re-reading, steps, thinking tokens, model choice	Routing and tiering, caching, step caps, context hygiene	Orchestration that routes each task to the right model; workspaces that hold stable context once
Human review	Share of work reviewed, time per review, reviewer cost	Risk-tiered approval, evidence packs, automated checks	Human approval placed only where it buys down risk
Acceptance rate	Task scope versus reliable horizon, consistency, quality of instructions	Narrow scoping, proven routines, evaluation before deployment	Skills and routines that encode how good work is done
Escaped errors	Review coverage, calibration, severity of failure modes	Hard gates on irreversible actions, sampling audits, incident learning	Approval at the point of commitment; memory of past failures, curated before reuse
Fixed cost ÷ volume	Build and maintenance effort, number of workflows served	Reusing agents across workflows, shared platforms, quarterly review	One workforce serving many workflows rather than one agent per pilot

4.4 Worked example: a retail deduction dispute

Illustrative only. The figures below are assumptions chosen to be realistic for a mid-market consumer-goods brand, not measured results. Token prices reflect published frontier and lightweight rates as of July 2026, cited by McKinsey ^[10].

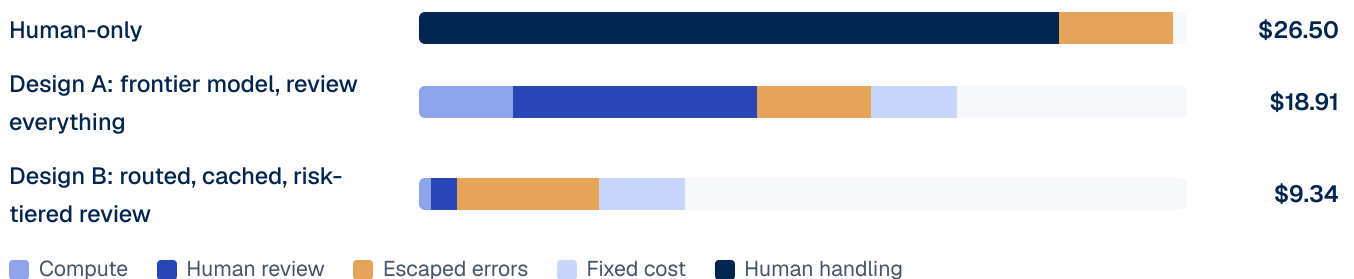
● Manelink analysis

Consumer-goods suppliers receive deductions from retailers' payments for shortages, pricing disputes and compliance fines, and each one must be researched, matched to documents and either accepted or disputed. It is a good test case for agents: high volume, document-heavy, rule-bound, with real money at stake when an error escapes.

Assumption	Human-only	Design A: frontier model, review everything	Design B: routed, cached, risk-tiered review
Volume	20,000 disputes a year	20,000	20,000
Model mix and price (per million tokens, input / output)	None	100% frontier at \$5 / \$30	30% frontier, 70% lightweight at \$0.20 / \$1.25
Tokens per attempt	None	400K input, 8K output, no caching	400K input (300K served from cache at 10% of price), 8K output
Tools and data calls per attempt	None	\$0.10	\$0.10

Assumption	Human-only	Design A: frontier model, review everything	Design B: routed, cached, risk-tiered review
Human time	30 min handling at \$45/hour loaded	100% reviewed, 8 min each	25% reviewed (high-value or low-confidence), 4 min each with an evidence pack
First-pass acceptance	Not applicable	70%	85% (narrow scope, proven routine)
Escape rate × cost per escaped error	2% × \$200	2% × \$200	2.5% × \$200
Fixed cost (orchestration, monitoring, maintenance)	None counted	\$60,000 a year	\$60,000 a year
Cost per accepted outcome	\$26.50	\$18.91	\$9.34

Chart 1. Same work, three operating designs: tokens are the smallest line. Cost per accepted outcome, US\$.



Four things stand out.

- 1. Tokens are the smallest line.** In Design A, compute is US\$2.34 per attempt and review US\$6.00, so oversight is 72% of variable cost, in line with McKinsey's 70–75% ^[10].
- 2. Retries are a hidden tax.** At 70% acceptance, Design A pays for 1.43 attempts per accepted dispute. Rework adds US\$3.57, more than the compute itself.
- 3. Cutting review raises a different risk.** Design B reviews a quarter of the work, and we assumed its escape rate rises from 2% to 2.5%. Escaped errors become its largest line. The approval layer has to be designed, not just thinned.
- 4. After optimization, reliability and scale dominate.** In Design B, escaped errors and fixed cost make up 86% of the total. The next gains come from fewer failures and more volume, not cheaper tokens.

Chart 2. Where the savings are: redesigning approval beats waiting for cheaper tokens. Saving in cost per accepted outcome from Design A, US\$.



Redesigned approval: 25% of attempts reviewed for 4 minutes with an evidence pack, escape rate up from 2% to 2.5%.
 Recomputed from the worked example's stated assumptions.

4.5 The Autonomy–Economics Matrix

Proposed framework. The worked example shows that where people approve agent work is the most powerful economic lever. But approval is also the main defence against costly errors. The matrix below places each task by two questions: **how cheaply can the result be checked**, and **how costly is an error that gets through?** The answer sets the operating mode, and with it the review share r and acceptable escape rate e .

● Manelink analysis

	Low cost of error (reversible, internal, small amounts)	High cost of error (irreversible, customer-facing, regulated, large amounts)
Cheap to verify (rules, reconciliations, tests, matching)	1. Automate and audit. Agent acts autonomously; automated checks on every result; humans sample a small share. Example: matching remittance lines to invoices.	2. Gate the commitment. Agent prepares everything; automated checks run; a person approves only the irreversible step. Example: submitting a dispute to a retailer portal.
Expensive to verify (judgment, open-ended research, persuasion)	3. Draft and batch-review. Agent drafts; people review in batches and accept imperfection. Example: first-draft summaries for an internal weekly review.	4. Human-led, agent-assisted. Person owns the decision; agent gathers, drafts and checks. Example: settling a disputed contract term with a key account.

The matrix follows the Princeton recommendation that reliability requirements should scale with the degree of autonomy ^[14], and the production pattern of short, supervised agent runs ^[17]. It adds an economic rule: **move a task one quadrant toward the top-left by making it cheaper to verify, not by trusting the agent more**. Structured outputs, evidence packs, deterministic checks and narrower scopes all make verification cheaper. Unverified trust only moves cost into the escape term.

The sampling rate shown in quadrant 1 is an illustrative starting point; set it from your measured escape rate.

05

Business and organizational implications

For finance leaders

- **Budget per outcome, not per seat or per token.** A per-token budget rewards the wrong behaviour; a per-outcome budget ties AI spend to the work it replaces or creates.
- **Expect AI costs in labour lines.** If review is 70% of variable cost ^[10], most agent economics live in payroll. Chargeback models that allocate only the AI invoice will understate true cost.
- **Model the distribution, not the average.** Token use varies by up to 30× across runs of the same task ^[12]. Plan for the tail and set budget alerts on it.

For operating leaders

- **Review capacity becomes the scarce resource.** When agents produce work faster than people can approve it, the approver becomes the bottleneck. Design approval queues, service levels and evidence formats as deliberately as any production line.
- **Workforce change is slower and different than expected.** In McKinsey's survey, 32% of respondents had expected AI-driven workforce reductions in the previous year; only 14% reported them ^[9]. Roles shift toward specifying, approving and improving agent work before headcount moves.
- **Redesign the workflow, not just the task.** Nearly three-quarters of AI high performers report fundamentally redesigning workflows because of AI, against about one-quarter of others ^[9].

For technology leaders

- **Build cost telemetry per request and per outcome.** List prices are unreliable ^[13], and models struggle to predict their own cost ^[12]. Granular monitoring of AI spend is the top tooling request in FinOps practice ^[8].
- **Evaluate with repeats.** Single-run success rates hide large run-to-run variance ^[14, 21]. Three to five runs per test case is a practical minimum for any agent that will act autonomously.

- **Treat AgentOps as a standing capability.** Agents need frequent tuning and periodic replacement ^[10]. Fund them as products, not projects.

For procurement

- **Ask vendors for cost per accepted outcome on your data.** Agent washing remains common: Gartner estimated that only about 130 of thousands of vendors claiming agentic capability offered it ^[6].
- **Price against outcomes where you can.** Gartner expects generative-AI cost per customer-service resolution to exceed US\$3 by 2030, above many offshore human agents, as subsidies end and use cases grow more token-hungry ^[27]. Contracts priced on outcomes, with caps on per-request cost, shift that risk where it belongs.

Through the AI-workforce lens

The research suggests that the organizations that win will run agents less like software licences and more like a workforce: each agent with a defined job, a workspace holding the context it needs, skills and routines that encode how good work is done, memory that is curated rather than accumulated, and human approval placed where it buys down risk. Each of those elements maps to a term in the CPAO model. Workspaces and cached context reduce compute. Skills and routines raise acceptance. Approval design sets review cost and escape rate. Reuse of the same workforce across many workflows amortizes fixed cost. Agent economics is not a separate discipline from workforce design; it is the measurement of it.

● Manelink analysis

06

Recommendations

Each recommendation follows from the evidence above and comes with a metric you can start tracking this quarter.

1. Make cost per accepted outcome your unit of account. Define acceptance criteria before you build. Report CPAO next to the human baseline for every agent workflow. *Metric: CPAO and human baseline per workflow, monthly.*

2. Measure real cost distributions, not list prices. Run candidate models on a sample of your own tasks, several times each, and compare cost per accepted outcome, not price per million tokens [13, 14]. *Metric: median and 90th-percentile cost per attempt, per model.*

3. Scope work to the reliable horizon. Break workflows into steps an agent completes consistently, cap the number of steps, and escalate early rather than letting failing runs burn tokens [12, 17, 19]. *Metric: first-pass acceptance rate; share of runs hitting the step cap.*

4. Route by default; gate frontier models. Send routine work to lightweight or specialised models and reserve frontier reasoning for the tasks that need it, as Gartner advises [5]. Learned routers have cut costs by more than 2× without losing quality on public benchmarks [28], and model cascades by up to 98% on some tasks [29], but verify the saving per outcome, because cheaper models can fail or cause side effects more often [20]. *Metric: share of attempts on frontier models; CPAO by route.*

5. Reuse what is stable; be sceptical of what is rebuilt. Cache stable instructions, policies and reference documents, which can cut input cost by up to 90% on long prompts [23]. Curate skills and routines offline, test them, and reuse them across many tasks. Count the tokens of any memory or skill system against a budget-matched plain agent before adopting it [21]. *Metric: cache-hit rate; tokens per attempt with and without each module.*

6. Design the approval layer by risk. Use the Autonomy–Economics Matrix to set review share by task. Invest in making review fast, with evidence packs, structured outputs and automated pre-checks, and track escaped errors as carefully as review time. *Metric: review minutes per accepted outcome; escape rate; cost of escapes.*

7. Amortize and review quarterly. Build agents once and reuse them across workflows; McKinsey's examples show volume sharply lowering unit cost [10]. Put agent unit economics in quarterly business reviews and re-test workflows that failed the economics test, because model capability and prices change fast [10]. *Metric: workflows served per agent; fixed cost per accepted outcome.*

07

Future outlook

Manelink analysis. Forecasts are uncertain; the signposts are what to watch.

● Manelink analysis

Over the next 18–24 months we expect three shifts. First, per-workflow cost will keep rising for the most ambitious agentic work even as routine work becomes very cheap, as Gartner forecasts [7]. The gap between well-routed and poorly-routed deployments will widen. Second, the binding constraint will move from model capability to verification capacity: capability is compounding fast [16, 18], reliability more slowly [14], and the human approval layer becomes the bottleneck. Third, pricing will drift toward outcomes, because buyers who cannot predict token consumption [12, 13] will push vendors to absorb the variance.

Signpost	What it would mean
Consistency (pass on every repeat, not just one) improves as fast as headline accuracy	Review shares can fall safely; the verification multiplier shrinks; more tasks move to quadrant 1
Models predict their own cost reliably before running	Fixed-price and outcome-based agent contracts become practical at scale
Budget-matched evaluations show memory and skills beating plain agents	Adaptive learning becomes a cost lever, not just a capability lever
Regulation strengthens the right to reach a human	Assisted volume rises (Gartner expects a 30% rise in assisted service volume by 2028 from regulatory change [30]); hybrid designs win over full automation
Share of organizations reporting EBIT impact from AI breaks out of the high-30s	Outcome-based management is spreading beyond the 6% of high performers

What would change our view: if token use per task stopped growing while prices kept falling, token price would again become the main lever. Current evidence points the other way [2, 7, 12].

08

Conclusion

The falling price of AI is one of the great technology stories of the decade, and it is the wrong number for running an AI workforce. Agents do not sell tokens; they deliver work. The cost of that work depends on how many tokens an attempt consumes, how often attempts succeed, how much human checking each result needs, what happens when errors slip through, and how widely the fixed cost of running agents is shared.

The organizations already getting value from agents behave as if they know this. They redesign workflows rather than bolting agents on, keep humans in the loop by design, measure impact, and scale agents across functions [\[9, 17\]](#). The rest are buying cheaper tokens and wondering why the bill keeps rising.

Cost per accepted outcome is a simple discipline that makes the difference visible. Measure it, design for it, and review it every quarter.

Checklist: is this agent workflow economically ready?

- ☐ Acceptance criteria for one "accepted outcome" are written down and agreed
- ☐ Human-only baseline cost per outcome is known
- ☐ Each candidate model has been run several times per test case and compared on cost per accepted outcome
- ☐ The task is scoped to steps the agent completes consistently, with a step cap and early escalation
- ☐ Routine steps are routed to cheaper models; frontier models are gated
- ☐ Stable context is cached; skills and routines are tested before reuse
- ☐ Every task is placed in the Autonomy–Economics Matrix, with a review share and acceptable escape rate
- ☐ Irreversible actions require approval at the point of commitment
- ☐ Cost per request, review time and escaped errors are tracked per workflow
- ☐ Agent unit economics are on the quarterly business review agenda

References

1. Cottier, B., Snodin, B., Owen, D. & Adamczewski, T. (2025). *LLM inference prices have fallen rapidly but unequally across tasks*. Epoch AI. epoch.ai
2. Gundlach, H. et al. (2025, revised March 2026). *The Price of Progress: Price Performance and the Future of AI*. arXiv:2511.23455. [arxiv.org](https://arxiv.org/abs/2511.23455)
3. Stanford Institute for Human-Centered AI (2026). *The 2026 AI Index Report*. hai.stanford.edu
4. Anthropic (2025). *How we built our multi-agent research system*. Anthropic Engineering. anthropic.com
5. Gartner (25 March 2026). *Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025*. Press release, as reported by Alwire. hpcwire.com
6. Gartner (25 June 2025). *Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027*. Press release. gartner.com
7. Gartner (17 August 2026). *Gartner Predicts AI Inference Costs Per Agentic Workflow Will Increase More Than Fivefold Through 2028*. Press release. gartner.com
8. FinOps Foundation (2026). *State of FinOps 2026*. 1,192 respondents. data.finops.org
9. Tinkoff, D., Van der Veken, L., Chui, M. & Balakrishnan, T. (25 August 2026). *The state of AI in 2026: On the road to ROI*. McKinsey & Company. Survey of 1,719 participants, fielded 4 May – 8 June 2026. mckinsey.com
10. Asif, C., Kiewell, D., Olanrewaju, T., Härmäläinen, L. & Kolaja, T. (24 August 2026). *Where AI agents pay off: A practical guide to the economics of agentic workflows*. McKinsey & Company / QuantumBlack. mckinsey.com
11. MIT NANDA (July 2025). *The GenAI Divide: State of AI in Business 2025*. Preliminary report. Background reading only. nanda.media.mit.edu
12. Bai, L., Huang, Z., Wang, X., Sun, J., Mihalcea, R., Brynjolfsson, E., Pentland, A. & Pei, J. (2026). *How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks*. arXiv:2604.22750. [arxiv.org](https://arxiv.org/abs/2604.22750)
13. Chen, L., Zhang, C., He, Y., Stoica, I., Zaharia, M. & Zou, J. (2026). *The Price Reversal Phenomenon: When Cheaper Reasoning Models Cost More*. arXiv:2603.23971. [arxiv.org](https://arxiv.org/abs/2603.23971)
14. Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S. & Narayanan, A. (2026). *Towards a Science of AI Agent Reliability*. Proceedings of ICML 2026. arXiv:2602.16666. [arxiv.org](https://arxiv.org/abs/2602.16666)
15. Yao, S., Shinn, N., Razavi, P. & Narasimhan, K. (2024). *τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains*. arXiv:2406.12045. [arxiv.org](https://arxiv.org/abs/2406.12045)
16. METR (29 January 2026). *Time Horizon 1.1*. metr.org
17. Pan, M. Z., Arabzadeh, N., Cogo, R. et al. (2025, revised June 2026). *Measuring Agents in Production*. ICML 2026 (oral). arXiv:2512.04123. [arxiv.org](https://arxiv.org/abs/2512.04123)
18. Stanford HAI (2026). *The 2026 AI Index Report: Top takeaways* (OSWorld agent success, organizational adoption). hai.stanford.edu
19. Xu, F. F. et al. (2025). *TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks*. NeurIPS 2025 Datasets and Benchmarks Track. arXiv:2412.14161. [arxiv.org](https://arxiv.org/abs/2412.14161)
20. *WorkBench Revisited: Workplace Agents Two Years On* (2026). arXiv:2606.13715. [arxiv.org](https://arxiv.org/abs/2606.13715)
21. Hajimiri, S., Aminbeidokhti, M., Dolz, J., Ben Ayed, I., Laradji, I. H., Gella, S. & Gontier, N. (2026). *Are Online Skill and Memory Modules Always Worth Their Tokens? A Budget-Constrained Study of Web Agents*. ServiceNow AI Research, ÉTS Montréal, UBC, McGill. arXiv:2606.15017. [arxiv.org](https://arxiv.org/abs/2606.15017)
22. Wang, Z. Z., Mao, J., Fried, D. & Neubig, G. (2025). *Agent Workflow Memory*. ICML 2025. arXiv:2409.07429. [arxiv.org](https://arxiv.org/abs/2409.07429)
23. Anthropic (2024). *Prompt caching with Claude*. anthropic.com
24. Niederhoffer, K., Rosen Kellerman, G., Lee, A., Liebscher, A., Rapuano, K. & Hancock, J. T. (22 September 2025). *AI-Generated "Workslop" Is Destroying Productivity*. Harvard Business Review (BetterUp Labs and Stanford Social Media Lab). hbr.org
25. Becker, J., Rush, N., Barnes, E. & Rein, D. (2025). *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. METR. metr.org
26. Brynjolfsson, E., Li, D. & Raymond, L. (2025). *Generative AI at Work*. The Quarterly Journal of Economics, 140(2), 889–942. [doi.org](https://doi.org/10.1093/qje/qjaf001)
27. Gartner (26 January 2026). *Gartner Predicts GenAI Cost Per Resolution for Customer Service Will Exceed Offshore Human Agent Costs by 2030*. Press release. gartner.com
28. Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W. & Stoica, I. (2025). *RouteLLM: Learning to Route LLMs from Preference Data*. ICLR 2025. [arxiv.org](https://arxiv.org/abs/2504.04401)
29. Chen, L., Zaharia, M. & Zou, J. (2024). *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*. Transactions on Machine Learning Research. arXiv:2305.05176. [arxiv.org](https://arxiv.org/abs/2305.05176)
30. Gartner customer service predictions as reported by Outsource Accelerator (2026). *GenAI customer service costs to surpass human agents by 2030*. news.outsourceaccelerator.com

Token prices are collapsing, yet agent bills and project cancellations are rising, because the cost of AI work is set less by the price of a token than by how many tokens an attempt burns, how often attempts fail, and how much human checking each result needs.

For executives, operators, and finance and technology leaders · About 25 minutes · 30 references

Manelink AI makes the Office: AI agents take seats on your org chart, learn your business and run its routines, so results improve at lower AI cost.

manelink.com · hello@manelink.com

