Rebuilding Ratios

AI Distortion in Value Measurement

How AI quietly changes what the measures of cost, productivity, and return actually measure, and the disciplines that rebuild them.

James E. Murphy

ORCID: https://orcid.org/0009-0001-2217-4306

Working draft for discussion · v1.2 · 25 August 2026

Latest version at jamesemurphy.com/ai-economics/measurement

Abstract

The arithmetic survives; what it measures does not. Every ratio is a projection: it renders some dimensions of an economic object in full and collapses the rest, which is what a measure is for, and it misleads only when the collapse goes undisclosed and the dimensions move. Firms are building their first broadly institutionalized generation of AI value measures, return on investment, cost per outcome, output per employee, spend as a share of revenue, utilization, deflection, and consumption based procurement benchmarks, and the formulas are the familiar ones. The objects being divided increasingly are not.

Change alone is not the problem. A ratio whose base moves is not thereby corrupted, and this paper does not argue that it is. Distortion arises when the change violates an assumption of comparability that reading the measure as evidence of value requires, and when that violation goes unrecognized and unadjusted. The resulting number can be accurate in every observation and still mislead in the comparison it invites.

Four mechanisms recur. First, the measure can become coupled to the measured: successful automation changes the headcount or cost base against which AI spend and productivity are judged. Second, populations can sort themselves: automation absorbs tractable work and leaves humans a progressively harder residual, making both machine and human averages move for compositional reasons. Third, units can drift beneath stable names: a token, task, conversation, or resolved case can change materially in capability, complexity, compute, or human involvement while continuing to be counted as the same unit, and where the drift occurs beneath the version label rather than at it, the nominal unit is not merely unstable but misidentified, since the same model name can span a fivefold range in cost per correct answer depending on which endpoint serves it. Fourth, the unit of trade itself can migrate, as seat based software economics give way to credits, actions, conversations, and other consumption units that are not directly comparable across vendors or time.

Two further cases leave the unit alone and move what surrounds it. The rate can vary for identical work, where committed capacity and on demand consumption serve the same model and traffic overflows silently between them, so that cost per outcome is not single valued for identical outcomes and a per request figure derived from a period blend is an allocation rather than an observation. And the production of the unit can vary, where requests are routed to different models by task or by requester, so that the denominator holds still while the numerator is drawn from a mixture the ratio does not disclose.

These mechanisms matter because value measurement requires more than accurate observations; it requires economically valid matching, and the matching fails from two directions. The mechanisms alter the economic objects being measured. Structural features of AI economics, benefits lagging the investment vintages that created them, capital and expense scattered across owners and accounting treatments, determine whether the right objects are joined at all. Together they can break the match between current benefits and prior investment, between reported spend and the full economic capital employed, between treatment and counterfactual populations, and between nominal units across periods. The consequences extend beyond noisy dashboards. Distorted measures can create false scale and false stop decisions, budget whipsaws, incentive and procurement errors, gaming without false reporting, self reinforcing allocation loops, and ultimately loss of confidence in management information.

NPV, IRR, ROIC, payback, and unit economics remain useful when their inputs refer to comparable economic objects. The practical task is therefore to rebuild the ratios: version the reference, and measure it where change is not declared, condition the population, match costs, benefits, capital, and time, normalize to stable outcomes above the quality floor below which an outcome is worthless at any price, and separate ex post performance measurement from ex ante decision measurement. Each of those presumes the terms can be gathered at all, which is a condition on the attribution path that firms rarely report and should. AI does not make ratios obsolete. It makes their construction consequential.

1. Three numbers that lied

A firm baselines its claims operation in January because the guidance said to. The baseline is careful: staffing, ticket mix, unit cost, cycle time, rework, service quality. By September, the process has been rebuilt around an AI agent. The work that once passed through people now enters a routing layer, tasks split differently, steps disappear, cheap work is demanded more often, and staffing adjusts incrementally rather than through one reorganization. Management calculates ROI from correct spend and operating data, and the internally consistent model compares the redesigned September operation with a January operation that no longer describes the relevant counterfactual. Nobody falsified the inputs. The comparator decayed at the speed of the treatment.

A support organization watches its assistant's cost per ticket fall while the human desk's cost per ticket rises. The conclusion appears obvious: the machine is improving, the human channel is not, and capital should shift. But the assistant takes the tractable work, absorbing progressively more of it as capability improves. The human desk holds what remains: edge cases, emotionally difficult interactions, incomplete information, policy exceptions, and failures of prior automation. Both averages are calculated correctly. Neither represents the cost of handling a common population. If the firm compares them without conditioning for complexity, the selection mechanism created by the AI system is misread as an economic productivity difference.

A procurement team renews an enterprise software agreement on the per seat arithmetic it has negotiated for years. The price is attractive against the benchmark. Yet the vendor's new growth is increasingly tied to actions, conversations, credits, or agent executions rather than human seats. Usage expands without seats, and one product can carry multiple billing units that do not translate into one another. The benchmark says the firm bought well because the benchmark still assumes the unit that used to define use. The contract moved faster than the comparator.

Three rooms, one pattern. The numerator and denominator need not be wrong; the economic meaning of the ratio has changed while the arithmetic has remained stable. A repeated label, employee, ticket, task, token, seat, spend, can cease to denote a comparable economic object across periods or populations; the number's apparent continuity hides a discontinuity in what it represents.

One image from practice sets up what follows better than a definition does. Writing about how two adjacent disciplines relate, the FinOps Foundation's chief technology officer describes them as projections of one object taken along different axes, in the way an architect's floor plan and elevation are both accurate renderings of a building and neither is the building. The sentence worth carrying is his: a projection is not a simplification, it is a

choice about which dimensions to render in full and which to collapse, and every projection collapses something (Fuller 2026).

He offers it for the relationship between two practices rather than for the construction of ratios, and it transposes without strain. Cost per resolved case renders spend and volume in full and collapses who was routed to which channel, what a resolved case now contains, and what the resolution was worth. That collapse is not the defect; it is what a measure is for. The defect is that the collapse goes undisclosed, so a reader takes a number as a summary of the whole object when it is a rendering along one axis. Distortion, as this paper uses the term, is what happens when the axis moves and nobody says so.

A note on vocabulary, since four related words appear throughout and each does a distinct job. The base is what a ratio divides by, and it is a quantity. The population is the set of cases the ratio is computed over. The unit is what a count counts, a token, a task, a resolved case. And the reference is what a reading is compared against, a baseline period, a benchmark, or a counterfactual. A ratio can be broken at any of the four, and the mechanisms below are organized by which one moves.

This paper calls that phenomenon AI distortion. AI distortion occurs when AI driven change in the measurement object, its baseline, population, unit, reference point, or economic relationship, violates a comparability assumption required to interpret a measure as evidence of value, without a corresponding adjustment to the measure or its interpretation. Change in a ratio is not itself distortion; revenue per employee rising after substitution is the ratio doing exactly what it is built to do. Distortion is unrecognized change in what the measure represents, and it therefore lives in the inference, not the arithmetic. The four mechanisms developed below are causal routes into the construct: base endogeneity, population endogeneity, unit content drift, and migration of the unit of trade.

The construct can be stated with minimal notation. Let the observed measure be $R_t = Y_t/X_t$. Interpreting movement in R_t as evidence of value requires invariance assumptions the formula cannot test: that the population P_t over which Y and X are gathered is comparable to P_{t-1}; that the unit content U_t in which they are denominated is comparable to U_{t-1}; and that the reference against which R_t is read, a baseline period, a benchmark, a counterfactual, still describes the alternative it stands in for. AI deployment and the AI market move P, U, and the reference while leaving the formula untouched, so R_t remains correctly calculated while the mapping from R_t to economic meaning quietly changes. Same formula, different economic object: that is the construct in one line, and everything that follows is machinery for detecting when the assumptions have moved and repairing the comparison when they have.

The claim is deliberately narrower than technological exceptionalism. None of these mechanisms is wholly new. Campbell and Goodhart warned that measures change behavior once they become targets. Lucas formalized the problem of relying on relationships that shift when regimes change. Selection bias and Simpson's paradox long predate machine learning. Hedonic adjustment exists because the economic content of a product can change while the product name remains. Automation has altered labor denominators for decades, and software has changed licensing models before. The argument is not that AI invented these problems but that it can combine several of them inside the same firm level value measure, generate several of them endogenously through deployment, while the others can arise from the firm's own system changes or be imposed by vendors and the surrounding market, and

do so inside the budget, reporting, and governance cycles that ordinarily recalibrate the measure.

That distinction separates distortion from two neighboring ideas. Economic uncertainty asks what future outcomes will occur: adoption, performance, competition. Measurement uncertainty asks how imprecisely a quantity is estimated: how wide is the confidence interval, how noisy is the forecast. Measurement distortion asks a different question: does the number still refer to the same economic object or relationship? Uncertainty calls for scenarios, probabilities, and sensitivity analysis. Distortion calls for redesign of the measure itself.

The practical question follows. If an intervention can change the terms by which it is evaluated, how should firms measure AI value without allowing formally correct ratios to become economically non comparable?

Scope, stated plainly. This paper builds no price index series; that work exists and is cited. It does no welfare economics, offers no pricing strategy, and recommends no vendor. It is not a valuation method: the ratios here are management instruments upstream of valuation, the inputs a discounted cash flow analysis consumes, not a substitute for one. Its subject is the integrity of the ratios firms govern by, and its claims are graded to their sources throughout, in the convention of its companions.

2. Why the ratios are being built now

2.1 Levels before ratios

Current enterprise practice emphasizes levels rather than economic ratios. Firms can observe AI budget versus actual spend, model calls, token consumption, licenses activated, employees enabled, workflows touched, pilots launched, and adoption rates. They are operationally useful because they answer questions of control and deployment: who is using the tools, where consumption grows, which teams exceed budget.

The era's signature statistic is the seventy three percent of practitioners who report blowing their AI budget, a level failing against a level. The most visible new instruments are consumption views: the individual spend dashboards and division token budgets one hyperscaler turned on for its engineers. And the one ratio in wide circulation is adoption, usage as its own justification, the instrument the year's most discussed internal correction was aimed at, treated with its sources in this paper's companion.

Levels do not, by themselves, answer the economic question. High adoption can coexist with weak value; low adoption can be a failed product or a deliberately narrow system serving high value work. Consumption can increase because the product is creating value, because agents retry inefficiently, because a default model got costlier, or because cheap work induced demand. Spend can exceed plan because the investment is undisciplined or because the original plan underestimated profitable usage. Levels are necessary controls, but they are not self interpreting measures of economic performance.

That leaves a vacuum. The measures management eventually wants, cost per outcome, output per head, productivity, payback, ROI, contribution margin, return on invested capital, require two or more economic objects to be related. Building them is harder because the terms often live in different systems and because the intervention can change the objects being related.

2.2 The split ledger

AI creates a particularly visible version of an old organizational problem: economically connected costs and benefits are recorded under different owners. AI spend sits in the technology budget; labor capacity in an operating unit; vendor displacement in procurement; improved conversion in sales; quality effects in service or risk systems; training and redesign in still other lines. The firm can therefore possess every observation required for a valid economic ratio without assembling them in one management view.

Call this the split ledger. It is not simply an accounting classification issue. It is an information topology issue. The technology organization sees the cost but not the release it creates; the business unit sees the effects but not the platform cost; finance sees aggregate expense without a causal bridge to the program. The numerator and denominator of a meaningful return measure are institutionally separated.

The split ledger also candidates as the explanation for a strange aggregate fact: function level gains reported in the same survey years that more than eight in ten executives report no tangible enterprise level impact. The gains land in lines nobody joins to the spend.

The split ledger helps explain why adoption becomes such a powerful proxy. Usage is visible to both sides of the ledger at once, and it crosses organizational boundaries without requiring agreement about attribution. It therefore fills the vacuum even when it is a weak measure of value. In the absence of internally designed economic measures, vendor dashboards and usage reports can become de facto governance instruments because they are the measures that exist.

2.3 The instrumentation moment

The vacuum will not last. Boards, CFOs, operating leaders, procurement functions, and emerging standards efforts are pushing AI measurement from deployment statistics toward outcomes and returns. The push is documented: by one mid 2026 survey, forty five percent of chief financial officers are directing AI budgets at productivity measures rather than the strategic outcomes their boards expect, which is board pressure arriving as a measured misalignment; the 2027 planning guidance now in circulation instructs budget owners to cut AI pilots that lack governance, clear ownership, and success criteria, which is an instruction to build the ratios that would tell; and the open standards community has organized the substrate, with cost and value definitions shipping on a near monthly cadence through year end. Firms are moving from asking whether AI is used to whether it creates measurable value, at what cost and quality, against what alternative use of capital.

That transition creates a design moment. The first institutionalized generation of AI ratios will not merely describe performance; once institutionalized, they will set budgets, trigger kill or scale decisions, shape compensation, influence procurement, and define internal benchmarks. A weak ratio that becomes a governance convention can persist long after its original assumptions have ceased to hold. The relevant risk is therefore not that dashboards are already universally corrupted. It is that firms are formalizing value measures while the measurement object is unusually unstable, and, unarmored, the first generation arrives distorted: calibrated to baseline periods the substitution has already shifted, denominated in units that have already drifted, populated by samples the sorting has already skewed. A ratio that fails at debut forfeits trust faster than one that decays in service.

The rest of the paper treats that instability as a design problem. Section 3 identifies four mechanisms through which the measurement object changes. Section 4 sets out the matching problem those changes feed, alongside the structural hazards that trouble the match even where units hold still. Section 5 traces the governance consequences of trusting the resulting ratios. Section 6 derives a measurement architecture intended to preserve interpretability as the technology, workflow, and commercial model evolve.

3. Four mechanisms of AI distortion

3.1 Base endogeneity: the measure is coupled to the measured

Start with ratios a board already understands: revenue per employee, output per head, AI spend as a share of operating cost, and return on the AI investment itself. Each divides an output or expenditure by a base that appears external to the intervention. Under AI deployment, that base can become endogenous.

Consider AI spend as a share of operating cost: successful replacement of external services, repetitive labor, or other software shrinks the non AI base. AI spend can therefore rise as a percentage of total operating cost precisely because the AI program is working. A board that interprets the ratio as evidence of excessive dependence on AI can penalize success. The opposite error is also possible. Revenue per employee can increase when headcount falls even if revenue is unchanged. The ratio is a legitimate description of labor productivity, but it does not by itself say whether AI created enterprise value, shifted risk, degraded quality, or merely changed the denominator.

The point is not that these ratios are unusable; it is that their movement becomes jointly determined with the intervention. When AI changes labor, vendor, infrastructure, or workflow costs, the denominator is no longer a passive reference. Each marginal unit of AI enabled output can change the base against which AI is evaluated.

This coupling can also run through demand. When a unit of work becomes cheaper, the firm does more of it; cost per unit falls while total cost rises, because the price of execution induces volume. The same event can therefore produce an apparently favorable unit economics ratio and an unfavorable budget variance. Neither observation is false. The interpretation depends on whether the additional volume itself creates value.

The coupling is already visible where hiring is measured. Postings for the junior tiers that agents reach first fell by roughly a fifth in 2025, and 2026 brought the first reported seat decline at a major collaboration vendor, the purchased unit of human software use shrinking for the first time in its history.

The antecedents are clear. Automation has long changed headcount, and performance measures have long changed behavior once attached to incentives; the theoretical line runs from Campbell through Goodhart to Lucas, relations breaking down when the regime begins to act on them. The AI specific measurement problem is the location and clock of the adjustment. Earlier technology substitutions were mediated by visible management actions, install, redesign, reorganize, reduce, events with dates an analyst could rebaseline on. Agentic systems increasingly alter the process transaction by transaction. Work is routed differently, human steps are removed, and capacity is released continuously. There may be no discrete reorganization date at which the old denominator officially ends and the new one begins, only a ratio whose two terms began moving each other continuously, which is precisely the

condition under which a relation stops being usable for inference.

The warning is simple: a stable formula does not imply a stable reference. Before reading movement in a ratio, ask whether the intervention changed the denominator's economic meaning.

3.2 Population endogeneity: the populations sort themselves

Consider two support teams in the same company, working the same product, measured the same way. One deploys an assistant to handle incoming tickets; the other does not. A quarter later the first team's cost per resolved ticket has fallen and the second team's has risen, and the natural conclusion is that the assistant worked and the other team should adopt it.

Now look at what each team handled. The assistant took the tickets it could resolve, which were the routine ones, and the humans on that team kept what it declined: the ambiguous, the emotional, the policy exceptions, and the cases the assistant had already tried and failed. The second team handled its whole mix, easy and hard together. Both cost figures are computed correctly. Neither team's number is wrong. They are simply averages over different work, and the comparison between them measures the sorting rather than the productivity.

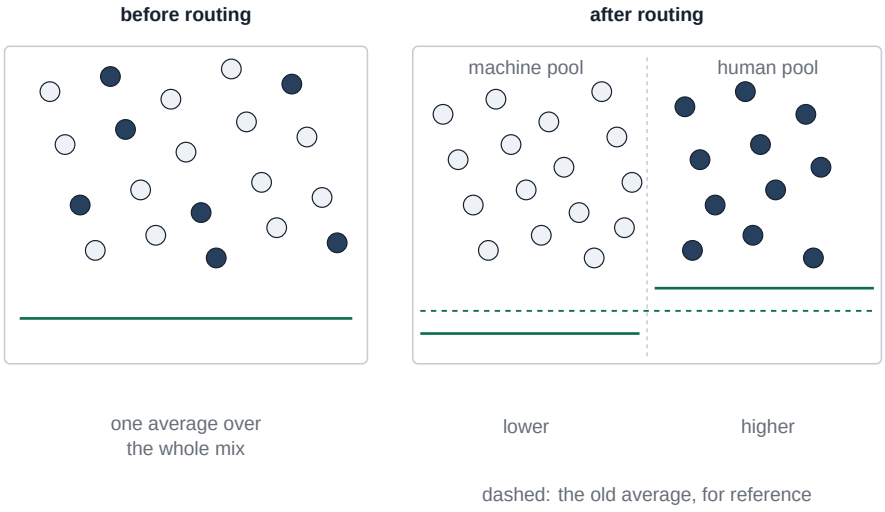

Figure 2. Sorting, not productivity. Light circles are routine cases, dark circles are difficult ones. Before routing, one average is computed over the whole mix. After routing, the machine handles the routine cases and the humans keep what it declined, producing two correct averages that straddle the original. No case became cheaper to resolve and no one became more productive; the population was divided, and the division moved both numbers. Schematic: positions and counts illustrate the mechanism and are not measurements.

The second mechanism is easiest to see in service operations, where automation frequently sits between incoming work and human labor. A naive comparison asks whether the AI channel or the human channel has lower cost, shorter handle time, or higher resolution. But the act of routing work to AI creates the populations being compared.

An assistant is deployed first against work it handles reliably, which is not a random sample: repetitive, well structured, documented, tolerant of standardized resolution. The work left

to people becomes the residual: incomplete, ambiguous, emotional, regulated, exception heavy, or already failed by automation. As AI capability improves, it absorbs additional layers of tractable work, further changing the human residue.

The result is a pair of accurate but non comparable averages. AI cost per ticket falls partly because the machine receives a selected population. Human cost per ticket rises partly because the human population becomes harder by construction. If management compares the two unconditional means, it attributes a compositional change to productivity.

The same problem affects quality. A deflection rate can improve when fewer interactions reach humans, yet customer effort can worsen if some users abandon, recontact, or escalate after failed automation. The practitioner literature states this in its own documents: one vendor's glossary notes that when routine contacts are deflected, the contacts reaching human agents skew toward higher complexity and can lengthen handle time, and one consulting analysis of hybrid programs reports human first contact resolution rising from 58 to 71 percent for the stated reason that the machine absorbs the routine. The same literature has started repairing its headline rate by hand: corrected deflection formulas now subtract the customers who come back within 48 hours, a correction acknowledging that the naive rate can count abandonment and delay as success. That is not merely a better KPI. It is an implicit recognition that the observed population and the outcome definition were selected by the system being evaluated. A further facet crosses the firm's measurement boundary entirely: deflected work does not always disappear, some of it moves onto the customer as self service effort, an unpriced input the firm's ratios cannot see, so cost per contact can fall while the total work of resolution has merely changed hands, and the recontact corrections are partly that effort flowing back. Procurement inherits the mechanism whole, since a pilot scoped to the easiest intents will post the best deflection, and the vendor that sorts hardest can win the contract on the sorting.

Population endogeneity also appears outside support. Coding copilots may be used first on routine tasks, leaving harder architecture and debugging work to senior engineers. AI assisted underwriting may automate straightforward files and leave humans a riskier tail. AI sales tools may be adopted first by motivated teams on better data. In each case, comparing AI handled and human handled averages without a common counterfactual can produce a false estimate of incremental value.

Selection bias, treatment heterogeneity, and Simpson's paradox are established statistical concepts. Outsourcing ran the mechanism at industrial scale: vendors took the commodity work, retained organizations kept what the vendors declined, and benchmarks misread retained cost as inefficiency until the adjustment became standard practice. The relevant departure is operational. The investment being evaluated is often the sorter. More precisely, AI handling is not a treatment but a policy, a routing rule whose eligibility frontier expands continuously with capability and thresholds; the channel comparison is then a clean answer to a question no decision needs, while the objects decisions do need, the effect at the eligibility frontier and the value of the routing policy against alternatives, go unmeasured. As capability, routing, thresholds, and policy change, both populations' composition changes continuously. The old selection effects polluted one pool; this one manufactures two, then invites the firm to allocate capital by comparing them. A stable average can therefore conceal a shifting task mix, while a changing average can reflect composition rather than true performance.

The implication is stronger than control for complexity. Firms must treat workload composition as part of the measurement object. If the population changes materially, the series has changed even when the metric label has not.

One further sorting happens on the production side rather than the case side, and it is now deployed rather than proposed. Firms route requests to different models by task type and by the requester's function, so that a complex engineering request reaches a frontier model while a routine one is served by something cheaper, and a marketing user receives a more prescriptive selection than an engineer does. Two units bearing the same label, a completed task of the same type, may therefore have been produced by different models, at different prices, at different quality, with the choice determined by who asked. The denominator holds still while the numerator is drawn from a mixture the ratio does not disclose. A firm that improves its cost per task by routing more traffic to cheaper models has changed the composition of what it counts, which is a real economic decision, and one that a series reporting only the ratio cannot distinguish from doing the same work for less.

3.3 Unit content drift: the unit changes under its name

The third mechanism concerns units that remain nominally stable while their economic content changes. AI has created a proliferation of such units: tokens, tasks, model calls, conversations, resolved cases, agent runs, tool calls, and actions. They are convenient because they are countable. They are dangerous when countability is mistaken for comparability.

A token is an accounting unit for model input or output, but its economic meaning depends on architecture, tokenizer, context, caching, reasoning behavior, latency, and the useful work produced. The counting authority migrated with it: earlier practitioner units kept the count on the buyer's side, certified counters applying a published standard, while the token's meter is the seller's tokenizer, so the party defining the unit is now the counterparty paid in it. Two token counts can be equal while representing different capability, different compute, and different business value. A task is less stable still: resolve a claim, write a test, review a contract can keep a name while complexity, required quality, human review, and tool path change materially.

Agentic execution makes the issue more pronounced because one nominal task can contain an indeterminate number of intermediate steps. The agent retrieves, calls tools, retries, branches, escalates, or asks another model to verify; two identical requests can consume very different internal work, with agentic execution multiplying the compute inside one nominal task by five to thirty times on current estimates. Cost per task can therefore move because the system changed its reasoning path rather than because the economic task changed. Conversely, the ratio can remain flat while the task itself has become more complex.

The magnitudes remove any temptation to treat this as rounding error. Held at constant capability, inference prices have fallen at rates between ninefold and several hundredfold per year depending on the milestone, while enterprise AI bills rose by triple digits; and over the same period the cost of running frontier level systems rose roughly eighteenfold per year, because marginal capability is bought with more inference. Cheaper tokens, costlier frontiers, and exploding consumption are all true at once, and a ratio denominated in bare tokens cannot see any of it. Nor is the exposure hypothetical at the level of controls: the year's most discussed internal correction denominated division budgets in tokens, a unit whose

content the next default model switch would silently reprice, which makes that correction, among other things, a unit event.

The problem resembles quality change in price measurement. Economists have long adjusted prices for computers and other goods whose capability improves over time. The nearer ancestors are services deflators: the telecommunications pricing dispute, in which alternative treatments of quality and volume weights turned one nominal unit into order of magnitude different measured price histories, and the early cloud computing price measurement work, the same industry one product generation earlier. The analogy is useful because it shows that unit drift is not conceptually unsolved. But the governance setting is different. National statistical systems can apply explicit quality adjustment methods at defined intervals, for goods whose quantity unit stays countable. Firms are making weekly, monthly, and quarterly operating decisions with units whose content may be changed by a vendor model release, a new reasoning mode, a routing configuration, or an internal workflow update, at invoice granularity, under the runtime control of a third party.

A harder case sits beneath the version label rather than at it. The mechanisms above assume that when a unit's content changes, something marks the change: a new version string, a price adjustment, a release note. That assumption fails when the change is not treated as a change by the party who made it. An endpoint's effective identity rests on weights, tokenizer, quantization, inference engine, kernels, caching, routing, and hardware, and any of those can move without a version bump, which is why endpoint monitoring work distinguishes stability from reliability: an endpoint can pass every conventional health check while what it returns has shifted (Leshin et al. 2026).

The distinction matters here for a reason specific to this paper's remedies. Section 6 asks a firm to keep a unit event log, and a log records events. Where the counterparty does not regard a change as an event, there is nothing to record, and the discipline that repairs unit drift has no input.

The variation this produces is measurable and large. Endpoint level benchmarking across 78 endpoints and 12 model families finds that the same nominal model, purchased under the same name from different providers, differs by up to 12.5 points in mean accuracy on mathematics and code, and that the endpoints serving a single open weight model range by a factor of 5.0 in dollars per correct answer, from roughly six tenths of a cent to three cents (Gao, Wang, and Yu 2026). A firm governing by cost per outcome and naming only the model in its series has a fivefold range hiding inside its unit.

That points at something stronger than drift. If the same model name spans a fivefold cost range depending on where it is served, then the model is not the economic unit at all. The unit that holds is the endpoint, the combination of provider, model, stock keeping unit, precision, decoding strategy, and region, and a ratio that identifies only the model has not finished saying what it divides.

It also sharpens the counting authority point above. The party that defines the unit also controls its content at runtime, and standard commercial agreements address availability, latency, and data handling rather than the constancy of what a unit contains or the disclosure of changes to it.

A related case leaves the unit's content entirely alone and moves the rate instead. A firm can hold committed capacity for a model and also consume the same model on demand, and

where the provider allows it, traffic above the reservation overflows to the on demand lane automatically; one provider's documentation describes six inference options for the same model under which the same work can cost half as much or three quarters more, depending which one served the request. Nothing about the request changes. The tokens are identical, the physical work is identical, and no measurement is in error. What varies is the rate applied, and the bill does not report at the granularity that would say which lane caught a given call.

The consequence for a ratio is precise: cost per outcome is not single valued for identical outcomes. Practitioners are already building the remedy, an effective blended rate constructed per model deployment and joined back to per request telemetry, and the remedy is sound. But it should be named for what it is. A per request cost derived from a period level blend is an allocation rather than an observation, and every unit economic built on it inherits that status. The repair is therefore to construct the effective rate explicitly, disclose the period and the mix it was computed over, and stop describing the result as what the request cost.

The same structure appears one layer down, wherever a resource is priced by time rather than by what is drawn from it. A firm provisioning accelerators whose capabilities its workload never exercises pays a unit price for a bundle it does not consume, and no per unit figure it reports will show this, because the unit is the device hour and not the capability used.

Practitioner measurement has arrived at the same phenomenon and given it an operational name. A large software firm reports tracking token to spend drift alongside cache to token ratio among the measures it watches most closely, which is the divergence between physical consumption and its monetary shadow made into a monitored series. That a firm is already watching for it suggests the mechanism is common enough to be noticed without a theory of it.

Tokens and tasks are essential operational measures; the error is reading a constant nominal count as a constant economic quantity. A governance ratio denominated in a drifting unit requires either normalization to a stable outcome or explicit versioning of the unit definition.

3.4 Trade unit migration: the unit of trade is replaced

Unit content drift assumes the unit survives. The fourth mechanism is more disruptive: the old unit of trade ceases to define the commercial relationship.

For much of enterprise software, the seat served as a rough common denominator. Imperfect, it still let buyers compare contracts over time and across vendors: a seat approximated a human user, a recurring unit of organizational access. AI weakens that equivalence. When software performs work rather than merely serving a person, usage no longer scales cleanly with human seats.

Practitioner reporting describes the transitional form and its budgeting consequence: on AI native tools the seat fee has become the floor rather than the price, with a variable component driving total spend, so functions that treated the software as a subscription line item found they had acquired a metered obligation without the visibility or controls a metered obligation usually carries. A ratio anchored to the seat count is measuring the part of the arrangement that stopped moving.

Vendors have responded plurally: credits, tokens, conversations, actions, work units, agent executions, outcome like charges, and hybrids combining platform fees, seats, included consumption, and variable usage. The problem is not merely price volatility. It is loss of a common denominator. A per seat benchmark cannot evaluate a per action contract without assumptions about activity. A per conversation price cannot be compared with a per agent execution price without a crosswalk from both units to a common economic outcome.

The scale of the transition is public. The early 2026 repricing took roughly two trillion dollars out of software valuations as the market marked the seat model down; one flagship vendor now sells the same capability per conversation, per action, and per user simultaneously while investors price whether the new units replace the old ones fast enough; and pricing and packaging now change under live contracts faster than an annual benchmark can be built on them.

This migration affects more than procurement. Budget forecasting, unit economics, internal chargeback, and vendor concentration analysis all depend on a stable unit of trade. If the unit changes during a contract term or if vendors use incompatible units for similar capability, historical series and external benchmarks can fail simultaneously.

Software has migrated between licensing models before; the license to seat transition took roughly fifteen years and moved between two stable units. Price measurement's new goods problem is the deeper ancestor, products entering and exiting handled by chaining across overlap periods; the departure is that chaining needs an overlap with market prices across the join, exactly what mutually incompatible vendor units without a crosswalk refuse to supply. The important distinction is fragmentation. The market is not necessarily moving from one stable unit to another; it may be moving from one common unit to multiple vendor specific units, inside a single budget cycle. That makes comparison a modeling exercise rather than a lookup.

3.5 One construct, four mechanisms

These mechanisms matter as a unified construct because the same ratio can carry several at once. AI ROI can be affected by a moving operating cost base, a self selected workload population, a drifting execution unit, and a contract whose billing denominator changed during the measurement period. Correcting one mechanism does not necessarily repair the others. These problems have generally been addressed in separate literatures and management disciplines, hedonic methods for the drifting unit, corrected rates for the sorted population, crosswalks for the migrated license; AI can make several of them relevant to the same management ratio inside the same governance interval.

The central issue is therefore not simply rapid technological change. It is a mismatch of clocks. Governance typically assumes that measures remain interpretable until the next budget, benchmark refresh, process review, or accounting close. AI can change the measurement object inside that interval. When the relationship decays faster than the process designed to recalibrate it, the ratio can lose economic meaning while continuing to be reported with familiar precision.

The four mechanisms divide into two families by causal route. Base and population endogeneity are deployment driven: the firm's own intervention changes the terms it is judged by. Unit content drift and trade unit migration need not be: the unit's content can move

with the firm's own routing configurations, reasoning modes, and workflow updates, or be repriced by a vendor's model release and the ecosystem's commercial migration, whether or not the firm deployed anything new. Both routes end in the same place, loss of comparability, which is why they belong to one construct; the distinction matters for repair, since a firm can instrument its own deployment but can only version and disclose what it does not control.

The causal chain is:

AI deployment endogenous change in base or population; system or market change drift or replacement of the unit failure of comparability or economic matching distorted inference governance error and feedback.

The chain is also the paper's architecture, and one exhibit carries it:

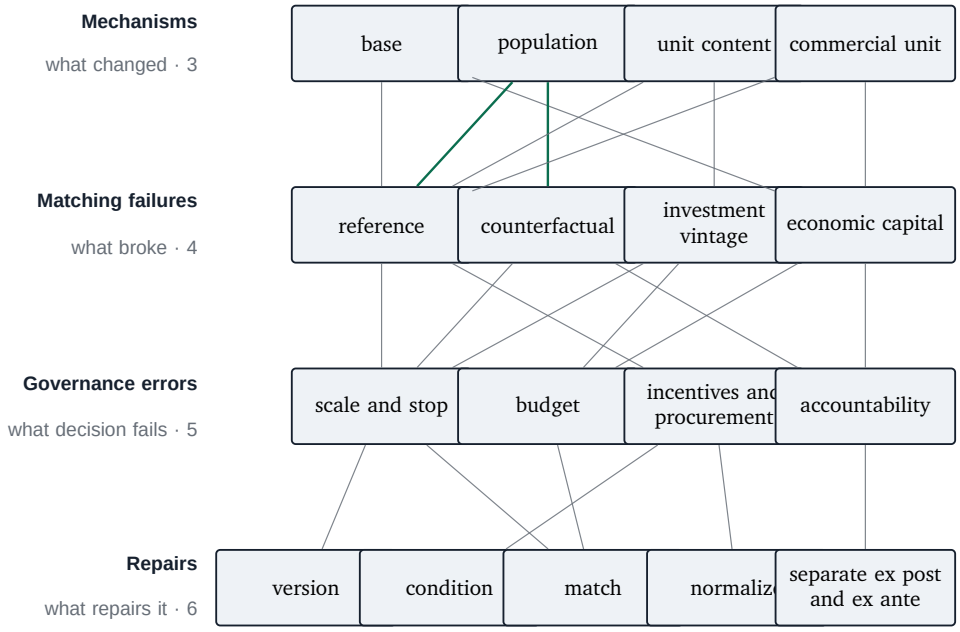

Figure 1. The four layers, and why they are not one to one. Reading down: what changed in the measurement object, what comparison that broke, what decision then fails, and what repair preserves comparability. The connectors are the point. A single mechanism reaches several matching failures, and a single failure arises from several mechanisms; the accent lines trace one mechanism, self sorting populations, through both the counterfactual it invalidates and the reference series it corrupts. No layer maps cleanly onto the one below it, which is why the repairs in section 6 are a discipline rather than a lookup table.

The layers are deliberately not one to one: a single mechanism can break several matching conditions, and a single matching failure can arise from several mechanisms; population endogeneity corrupts both the counterfactual and the reference series, and unit drift reaches reference matching, procurement comparison, and ROI at once. The matching layer also receives a second input the mechanisms do not supply, the structural hazards of vintage and capital with which section 4 opens.

The next section develops the middle of that chain. The mechanisms explain what changes; the matching problem explains how those changes enter value measurement.

4. The matching problem

A useful value measure is not produced merely by dividing two accurate quantities; the quantities must be economically related. Stated once in general form: for a ratio to function as valid evidence of value, accurate observations are necessary but not sufficient; the comparability assumptions linking them must also hold. Conventional controls audit the observations. This paper is about the assumptions, and this section gives them their structure as a matching problem with four recurring failures: reference matching, counterfactual matching, investment vintage matching, and economic capital matching.

The failures arrive from two directions. The distortion mechanisms of section 3 alter the economic objects being measured, and reference and counterfactual matching are where those alterations land. Vintage and capital matching are different in kind: they arise from structural features of AI economics, realization lagging the investment that caused it, capital and expense scattered across owners and accounting boundaries, the split ledger of section 2, and they would trouble AI value measurement even if every unit held still. Both routes converge on the same failure, an economically invalid comparison, which is why the repairs of section 6 treat them together.

4.1 Reference matching: is the comparator still the comparator?

Every ratio requires a reference, whether explicit or implicit. A before and after productivity comparison assumes the process is sufficiently similar across periods. A vendor benchmark assumes the unit of trade represents comparable use. Cost per task assumes a task is economically similar enough across periods that changes in cost are meaningful.

When AI changes workflow design, task boundaries, staffing, model capability, or pricing structure, the reference can become stale. The original baseline remains historically accurate but economically inappropriate for the current system. The response is not always to abandon the series; a crosswalk can sometimes preserve comparability. But the burden should be explicit: either demonstrate that the reference is stable, normalize it, or declare a break in the series.

This is the simplest form of the opening paradox. The January baseline did not become factually wrong in September. It became the wrong comparator for the question management was asking.

A comparator can also carry an assumption about the firm's own work that nobody states. Price comparisons across model endpoints conventionally blend input and output tokens at a fixed ratio, three to one being the common choice, which approximates a chat workload reasonably well and describes almost nothing else. Retrieval augmented generation runs an order of magnitude more input than output; agentic tool use runs heavier still; reasoning workloads invert the ratio entirely. Two endpoints that appear equivalent under the conventional blend can differ several fold in cost on a real traffic mix, and re ranking the same endpoint set under different workload weights produces top ten lists that overlap by only thirty to forty percent (Gao, Wang, and Yu 2026).

The consequence for reference matching is direct. A benchmark comparison is not a neutral fact about the alternatives; it is a claim about the buyer's traffic, made by whoever chose the blend. Matching the reference means stating the mix the comparison assumes and checking it against the mix the firm actually runs, rather than inheriting a convention because it is

the one the leaderboard used.

4.2 Counterfactual matching: are the populations comparable?

Value attribution requires a counterfactual: what would have happened without the intervention? A naive before and after comparison approximates that counterfactual only if relevant conditions remain sufficiently stable. AI often violates that assumption because deployment itself changes workload composition, demand, staffing, and process design.

The repair begins with the estimand, because three different objects answer three different decisions. The marginal effect at the eligibility frontier, what happens to the cases the routing boundary would take next, is what scale the routing decisions need, and it is exactly what the curated pool average is not. The value of the routing policy as a whole, against the alternatives not run, is what routing choices need. Program value against a no AI counterfactual is what ROI needs. Population endogeneity makes the naive substitute concrete: the cost of AI handled work and the cost of human residual work are conditional means for different populations, and neither identifies any of the three. A useful identity disciplines the reading across periods: the change in a policy's measured value decomposes into a composition term, the population moving, and a per case effect term, the performance moving, which turns conditioning from adjustment into accounting.

Reference matching also assumes a reader can say what was bought, and in this market that is not always available. The same nominal token can be purchased from the model owner directly, through two different cloud marketplaces, or bundled into a data platform under a credit system, alongside capacity a firm produces on hardware it owns or leases. The unit's name is stable across all of them while the effective price, the commercial packaging, and in the credit case the unit of account itself are not. A cost per token compared across two such wrappers is comparing purchases whose identity is the same and whose economics are not, and the reference has to name the wrapper as well as the unit.

The same logic applies to productivity. If AI reduces the time required for routine work, employees may use released capacity on harder work, more work, quality, or activity the original metric never captured. A drop in hours per old task is not necessarily the same as a gain in enterprise output. The counterfactual must specify what would have happened to both the task mix and the released capacity.

A harder case has no counterfactual to specify. Where AI enables work that would not otherwise have been done at all, there is no displaced activity to price and no before state to compare against; the standards work now forming has named this as an open problem, since methods built on measuring deflected work are complete only where work was displaced. A working practice has emerged in its place, and it substitutes a design for a measurement: treat the cost as a hypothesis about value, fix the success criteria and the time bound in advance, and let the pilot confirm or disconfirm. That is a defensible move and it has a governance benefit, since it ends pilots that run indefinitely.

It also inherits a hazard this paper's own material names. Criteria set by the sponsor, against which the sponsor later reports, are the structure the performance measurement literature of section 7.1 warns about, and the substitution of an experimental design for a counterfactual only works if the design resists being met by redefinition. The three conditions section 6.6 places on the integrity test apply here without modification: the criteria should be applied by

someone who does not own the number, evidenced rather than attested, and logged before the fact rather than reconstructed after it. Absent those, a value hypothesis is a business case with a date on it.

4.3 Temporal matching: which investment vintage produced the benefit?

Current period AI benefits and current period AI spending often belong to different investment vintages. Infrastructure, data, integration, redesign, and adaptation precede steady state benefits, while current spending expands capacity that matures later.

A contemporaneous ratio of AI benefit in year t to AI spend in year t can therefore mix returns from old capital with costs of new capital. The ratio may rise because earlier investments are maturing while current investment falls. It may fall because a new build cycle is underway while prior benefits remain unchanged. In neither case does the number cleanly represent the return on a specific investment cohort.

The productivity J curve literature provides the economic intuition: general purpose technologies often require complementary intangible investment before measured productivity improves. AI can intensify the problem because complementary investment is fragmented across systems, data, process redesign, training, governance, and experimentation. The first observable financial benefit can arrive after several separate investment decisions. The lag has a formal repair: surrogate indices, short run indicators statistically validated to predict long run value, let a vintage be graded before its cash arrives, with the validation itself testable.

A more coherent internal view tracks investment vintages. For each major AI capability or program, management records when economic capital was committed and then follows realized benefits, incremental operating costs, and remaining economic capital over time. This does not eliminate attribution uncertainty, but it avoids the category error of dividing this year's benefits by this year's spend as if they belonged together.

Temporal matching also clarifies a critical governance distinction. Ex post performance asks how prior capital vintages have performed. Ex ante allocation asks whether the next tranche of capital has a positive expected return from today's starting point. Historical ROI is evidence for the second question, not the decision rule itself.

4.4 Economic capital matching: what capital is actually employed?

The denominator problem is as important as the timing problem. AI economic value may depend on capital and expense that do not appear in a single AI investment line. Infrastructure may be capitalized while model access is expensed; engineering expensed, data preparation elsewhere, redesign embedded in business unit salaries, training dissolved into operating budgets, and common platforms supporting dozens of applications.

The split ledger therefore creates a difference between reported AI spend, accounting AI capital, and economic AI capital employed. Those concepts answer different questions.

Reported AI spend is useful for budget control. Accounting capital is necessary for financial reporting. Economic capital employed is the denominator required for internal return analysis. The repair has a solved ancestor of its own: the intangibles capitalization tradition, the measurement of intangible capital, the case against accounting's treatment of it, and the

valuation practice of capitalizing R&D and leases, exists precisely because reported classifications misstate economic capital; economic capital matching extends that move to AI's scattered complements. It should capture the economically relevant resources committed to create and sustain the capability, whether or not accounting rules capitalize them and whether or not they sit under one organizational owner.

This does not imply that every indirect cost should be allocated mechanically to every AI use case. Fully loaded and marginal economics answer different decisions. A common platform may have unattractive economics if all of its cost is charged to the first application, while the marginal cost of the tenth application may be highly attractive. The firm therefore needs both a platform level economic capital view and use case level marginal economics.

The point is not to find a single true denominator. It is to make the denominator correspond to the decision being made. If the board asks whether the AI platform as a whole is earning an adequate return, shared capital matters. If a product owner asks whether to add one more validated workload to already committed infrastructure, sunk platform capital should not determine the marginal decision.

4.5 Exposure and likely bias

The mechanisms and matching failures do not affect all ratios equally, and their direction is not always predictable. The table below summarizes the main exposures.

Measure	Primary distortion exposure	Typical matching failure	Risk if treated as a value measure
ROI and payback	All four mechanisms	Vintage, capital, counterfactual, reference	Ambiguous; can overstate or understate return
AI spend over operating cost	Base endogeneity; unit drift	Reference	Often rises with successful substitution
Revenue or output per employee	Base and population endogeneity	Counterfactual	Can overstate incremental value
Gross or contribution margin	Coupling; sorting; unit drift	Reference, counterfactual	Mix shift and cost line migration can masquerade as margin change
Cost per ticket or task	Population and unit drift	Counterfactual, reference	Often flatters the automated channel
Deflection and containment	Population endogeneity	Counterfactual, outcome definition	Overstates success without recontact and quality conditioning
Cost per token or model call	Unit content drift	Reference	Direction depends on model, version, and quality change
Utilization and adoption	Base and population effects	Reference	High value is not implied by high usage
Per seat and successor pricing	Trade unit migration; unit drift	Reference	Cross vendor and historical comparisons can fail
Procurement benchmarks	Unit and trade unit effects	Reference	Apparent savings may reflect unit change

The table matters for one reason: a management system should not ask only whether a metric is measured accurately. It should ask what forms of distortion the metric is exposed to and whether the comparison required for the decision remains valid.

Two measures resist the mapping notably, and both were engineered. A risk budget utilization measure resists because its reference is a governed constant, set by decision and changed only by decision, so nothing in the world can move it silently. A burdened cost per outcome built under discipline resists because its outcome definition is conditioned and its baseline is stamped, terms designed against exactly these mechanisms. Both are the companion papers' constructs, cited here as existence proofs. The lesson the table supports is scoped: resistance to these mechanisms was designed in, not inherited.

5. When correct numbers produce wrong decisions

Measurement error matters because firms act on measures. Once a ratio becomes a governance instrument, its distortions can alter budgets, routing, procurement, incentives, and capital allocation. Several failure modes follow.

5.1 False scale and false stop

The most direct consequence is a decision error around continuation and scale. Let the decision relevant return be the return management would estimate under appropriately matched costs, benefits, populations, units, and periods; there is no observable true return, since attribution is itself estimated. Let the measured return be the ratio management actually observes. Two mistakes follow when those diverge.

	Matched return above hurdle	Matched return below hurdle
Measured return above hurdle	Correct scale	False scale
Measured return below hurdle	False stop	Correct stop

False scale occurs when the measure makes a weak investment look attractive. Omitted complementary capital can do this. So can favorable population sorting, a quality decline hidden by throughput, or benefits from older investment vintages being credited to a new spending period.

False stop occurs when an attractive investment looks weak. Benefit realization lags can do this, especially during periods of complementary investment. So can charging a common platform entirely to its earliest applications or comparing a newly difficult residual human population with an outdated pre AI baseline.

The point is not that each distortion has a fixed sign. Many are ambiguous. The useful discipline is to ask what direction the mechanism would push the observed ratio under the current operating design.

5.2 The budget whipsaw

A second failure mode is phase lagged control. AI costs can be visible immediately while benefits arrive later, elsewhere, or in forms not yet joined through the split ledger. Management therefore observes a negative signal before the positive one has matured, cuts or caps usage, then watches prior benefits arrive and make the cuts look premature; spending resumes, and the cycle can repeat.

The year just lived reads as the exhibit rather than the prediction: a budget exhausted in four months and answered with per engineer caps; a deferral wave called in advance as finance chiefs found no numbers they trusted, then a rebound conditioned on governance; record infrastructure spending in the same twelve months as enterprise clampdowns. The sequence is consistent with the mechanism; it does not by itself establish that distorted measurement caused it, and the claim here is consistency, not proof. Splurge, crackdown, and renewed spending can therefore be less a sign of indecision than a predictable response to a lagged and mismatched signal. Traditional variance control works poorly when the cost and benefit clocks differ and when the benefit is recorded under another owner.

The repair is not to suspend discipline. It is to define in advance which indicators are ex-

pected to move at each stage of the investment and to distinguish leading operational evidence from lagging economic realization. The surrogate discipline makes that definition estimable rather than judgmental: the stage indicators are chosen and validated as predictors of the realization they stand in for. A build phase program should not be judged as if it were a mature operating asset, but neither should strategic value excuse indefinite non realization.

5.3 Incentive error and gaming without lying

Metric fixation literature often focuses on false reporting or overt gaming. AI distortion creates subtler channels. A team can improve its reported cost per case by routing difficult work elsewhere. A product owner can narrow the task definition so the measured task becomes easier. A procurement function can select the commercial unit that produces the most favorable benchmark. A budget owner can benefit from a model or unit change that makes reported consumption look better without reducing economic cost.

Every number can survive audit. The distortion resides in the composition, scope, unit, or comparator.

One asymmetry keeps these channels open. A distortion that embarrasses the person reporting it tends to be investigated; a distortion that flatters them tends to be defended, or simply not examined. Nothing in the arithmetic distinguishes the two cases, which means the survival of a distorted measure depends less on how visible it is than on which direction it points.

This is gaming without lying. It matters because conventional controls are better at validating observations than validating the economic equivalence of the objects being compared. Targets inherit the problem: a target calibrated on a distorted baseline carries the distortion forward and then ratchets against it, so the compensation system can institutionalize the error even where every measure survives audit. If compensation or budget authority is tied to an AI metric, the integrity of the population and reference definition becomes part of internal control.

5.4 Procurement and benchmark error

Procurement is especially exposed to trade unit migration. A familiar benchmark can create false confidence even when the buyer and benchmark vendor are no longer selling the same economic unit. A low seat price is no bargain if value migrated to agent executions; a favorable token price says little if one product needs more tokens per outcome.

The appropriate comparison is therefore not always price per vendor unit. It is price per normalized business outcome at an agreed quality level, or, where that is impossible, a transparent model translating each vendor's unit into a common workload. That model introduces assumptions, but those assumptions are preferable to pretending that incompatible units are comparable.

5.5 Accountability error

A distorted ratio can also assign responsibility to the wrong owner. Rising human cost per ticket reads as an operations failure when it is partly AI absorbing easy work; rising AI share

of cost reads as overspend when it is partly successful substitution. A business unit may claim productivity benefits while the technology organization bears all platform expense.

The split ledger therefore creates an accountability problem as well as a measurement problem. When costs and benefits are separated institutionally, each owner can report a locally correct number that does not reconcile to enterprise economics.

5.6 The self confirming loop

The most important dynamic consequence is feedback. Suppose AI appears cheaper because it handles easier cases. Management routes more cases to AI. The human residual becomes harder. Human average cost rises. The relative measured advantage of AI increases. Management responds by routing still more work to AI.

The loop is: distorted measure allocation or routing decision changed base or population more favorable distorted measure stronger allocation decision.

This is more than a one time selection bias. The measure alters the operating design, and the new operating design alters the next measure. The ratio begins to grade a system partly of its own making.

The same feedback can work in the opposite direction. A new platform is charged with heavy upfront costs while benefits are immature. Its measured ROI is weak. Management cuts complementary investment. Because the complements are never completed, the expected benefits fail to materialize. The original weak ROI is then treated as validation of the decision to cut. A temporary J curve becomes a permanent failure through measurement driven underinvestment. The loop is also assessable before it closes: off policy evaluation of logged data can estimate what the routing alternatives not run would have yielded, so the next tightening is a choice made with the counterfactual in hand rather than a drift.

Feedback is where AI distortion becomes a governance problem rather than a technical measurement problem. A distorted number does not merely describe the wrong world; it can help create it.

5.7 Loss of trust in the measurement system

The end state is not necessarily a bad metric. It can be no trusted metric at all. After repeated reversals, productivity claims that fail to reach the P&L, pilots whose ROI disappears at scale, vendor benchmarks that cannot be reconciled with invoices, executives may stop believing AI value measures.

Two substitutes then emerge. One is allocation by advocacy, in which strategic narrative and sponsor influence dominate because no number is trusted. The other is metric freeze, in which firms retreat to inherited ratios because they are at least familiar and comparable, even when the underlying operating model has changed.

Both outcomes are worse than a transparent measurement system that acknowledges breaks, conditions populations, and distinguishes realized from expected value. Measurement credibility is therefore itself an organizational asset. Rebuilding ratios is partly about preserving that credibility before the first generation of AI value measures hardens into convention.

6. Rebuilding the ratios

The response to AI distortion is not to replace financial discipline with strategic storytelling. It is to engineer the terms of the ratio so that changes in the measurement object are visible. Much of the toolkit is the statistical profession's own, vintages, documented methods, stratification, re basing as a logged event, an independent applier, imported to a layer whose private statistical office has been outrun, plus the one discipline that profession never needed, defenses against a measured object endogenous to its own measurement. Five principles follow directly from the failure modes above, with two reporting disciplines reinstated inside them.

6.1 Version the reference

Every material AI value measure should have an explicit reference version. At minimum, the reference should identify the baseline period, workflow definition, model or system version, staffing structure, workload composition, and pricing regime relevant to the comparison.

Versioning assumes the change is announced. Where it is not, as Section 3.3 describes, the reference has to be measured rather than declared, and the instrument for doing so is inexpensive. Periodic fingerprinting samples a fixed prompt set from an endpoint, compares the resulting output distributions against a baseline, and accumulates evidence across successive samples to declare change events and define the stable periods between them (Leshin et al. 2026). Structural changes, to model family, version, inference stack, or quantization, are detected quickly; a small change to a sampling parameter can take many cycles to register, so a firm should expect fast detection of substitution and slow detection of tuning.

Two limits belong with the instrument, and both bear on what a firm can conclude. Comparing an endpoint against the consensus of other providers serving the same nominal model is sensitive to one provider diverging and blind to a change that reaches all of them at once, which is the industry wide case a firm's own series would most need to catch. And comparison against a first party reference is available only where the model owner exposes one; for a closed model served solely by its owner there is no external reference distribution, so change can be detected over time while deviation from an original cannot be established at all. The practical consequence is that the baseline fingerprint should be taken when a system is validated, because for a closed endpoint it is the only reference the firm will ever hold.

A material change in one of those elements should trigger one of three treatments: preserve the series because the economic object remains comparable; restate or crosswalk the historical series using a defensible adjustment; or declare a break in series and start a new reference period. The triad mirrors financial reporting's own machinery for accounting changes, retrospective restatement, prospective application, and disclosed breaks, extended here to the operating ratios no standard reaches.

This is standard practice in other measurement disciplines; the innovation is applying it routinely to AI operating ratios rather than presuming continuity from an unchanged dashboard label. A useful rule is simple: if the process, population, or unit changed materially, continuity must be demonstrated rather than presumed. And where the comparator cannot be measured at all, greenfield agentic work with no human predecessor, it is constructed as a modeled counterfactual with provenance and a grade, not presumed to be zero.

6.2 Condition the population

Raw averages should not be compared across self sorted populations. Measures should be conditioned on the dimensions that materially determine difficulty and value: intent, complexity, risk, customer type, product, workflow stage, or other relevant case attributes.

In support, report cost and resolution in aggregate and within stable complexity bands; in underwriting, compare within risk strata; in engineering, distinguish routine implementation from architecture, debugging, and review.

The identification stack follows the adaptive policy structure. Randomize at the margin: tie breaker randomization in a band around the routing threshold identifies the frontier effect exactly where the next expansion decision lives, cheaply and continuously. Hold out for program value, with two corrections built in: agentic systems interfere across arms, the shared model learns from every ticket including the holdout's and queues spill between them, so cluster and switchback designs replace naive unit randomization; and effects are time indexed under drifting capability, so the contemporaneous effect is the decision relevant one and naive pooling across periods is forbidden. Construct the counterfactual where it vanished, the synthetic control family building the missing comparator from donor units and pre periods, which is the opening vignette's repair. Evaluate policies from logged data, off policy methods assessing the routing alternatives not run. And reweight anything estimated from adaptively collected data. One data discipline underwrites the whole stack: the router's own scores, thresholds, and eligibility versions are the assignment mechanism, so the assignment log is the unit event log's sibling, and keeping it is the single highest value data act in AI measurement. Where none of this is feasible, the firm should at least disclose the extent to which routing or eligibility rules changed the population.

The purpose is not statistical perfection. It is to prevent the investment from creating a favorable comparison merely by changing who or what enters each column.

6.3 Match costs, benefits, capital, and time

The economic relationship inside the ratio should correspond to the decision being made. That requires two linked disciplines.

First, track investment vintages. Major AI programs should have a cohort view showing capital and expense committed by period, the expected realization curve, actual realized benefits, incremental operating costs, and remaining economic capital. Current period benefits should not automatically be divided by current period spend.

Second, reconcile the split ledger. Finance should create an internal economic view that joins technology spend, shared platform cost, business unit implementation, relevant labor or vendor effects, and attributable benefits. The view need not duplicate accounting treatment. Its purpose is to answer the managerial question: what resources were economically committed to produce this capability, and what cash or operating benefits have been realized from them?

Before any of that, a condition holds beneath every ratio in this paper and is almost never reported. Computing a cost per outcome at all requires that consumption be attributable to the work that caused it, which in practice requires that calls pass a point where metadata can be attached. Practitioner guidance is blunt about the architecture, build that point

first, because calls that bypass it are calls the unit economics cannot see. What no one reports is coverage. A firm running several providers, an internal platform, and a hundred or more use cases can see tokens in aggregate while being unable to say which use case or team consumed a given one, and consumption arriving through vendor software or licence priced features may emit no usage telemetry at all. An unmeasured coverage share is the unstated denominator beneath every attribution claim, and a ratio computed over an unknown fraction of the population, presented as though computed over all of it, is the paper's own characteristic failure applied to its own machinery. Coverage should be measured and disclosed alongside the ratio.

The denominator should then change with the decision. Platform level return should include common economic capital. Marginal scale decisions should focus on incremental cost and benefit from today's position. A sunk platform cost may matter greatly for ex post performance and not at all for whether the next validated workload should be added.

A reporting corollary follows for base endogeneity: no single quotient can carry the reading. Report three things together, the level, the share of total cost including the AI itself, and an explicit substitution bridge showing the cost the deployment displaced. The inclusive share is bounded and cannot be flattered by consuming its own base the way a ratio against an external base can, but it still moves under successful displacement; only the bridge says why it moved, which is the interpretive step the naked quotient smuggles.

6.4 Normalize to stable outcomes

Input units such as tokens, tasks, actions, or conversations are useful for operations but weak as ultimate value denominators when their content drifts. Wherever feasible, firms should normalize to a business outcome that remains economically meaningful.

Examples: cost per satisfactorily resolved interaction, per correctly processed claim, per accepted code change at a defined quality bar, per completed workflow within a stated error and rework threshold. The outcome must itself be versioned and quality conditioned; otherwise the firm simply moves the drift one layer up.

That warning bites hardest when the quality condition is assessed by a model. A judged outcome definition inherits the stability properties of its judge, and a judge is an endpoint subject to the same silent change as the system it evaluates. The two can move together, so that real degradation measured by a degraded instrument reads as no change, or apart, so that an unchanged system appears to improve. An outcome definition resting on model judgment therefore needs the judge pinned to a stated endpoint, fingerprinted on the same cadence as the production system, and re anchored periodically to human labeled cases. Without that, the normalization has relocated the drift rather than removed it. The practitioner ancestor is instructive: the function point, software's size and complexity normalized work unit, ran a forty year constant capability program complete with counting standards and certified counters, and its failure modes, standard drift, gaming, cross counter variance, read as design requirements here, which is why the outcome definition carries a version, a quality condition, and an independent applier.

Quality conditioning has a stronger form worth stating, because practitioners state it more sharply than the measurement literature does: every workload has a bar below which the output is worthless at any price, and in some functions the bar is set by a regulator rather

than by the firm. A cost per outcome ratio is therefore defined only above that floor, and a comparison that spans it is not a comparison. An option that is cheaper because it produces work below the bar has not produced a cheaper outcome; it has produced something else.

One corollary bears on the denominator discussion of section 6.3. Because the floor is set per workload rather than per firm, a firm level intensity figure aggregates over denominators that are not commensurable, and the aggregate moves when the mix of workloads moves even if no workload changed.

A stable outcome measure should include enough quality to prevent throughput from masquerading as value. Resolved should account for recontact or escalation. Completed should account for rework. Generated should not be equated with accepted. Time saved should not be treated as realized financial value unless released capacity is removed or redeployed into measurable output.

Where execution is stochastic, the unit cost is a distribution rather than a point. Agent paths vary across nominally identical requests, so the honest object is the curve: budget to the distribution and price the tail, with the open question being which percentile to govern to, not whether to govern the distribution. A point estimate can conceal economically material dispersion and tail cost.

Where a stable outcome cannot be defined, the firm should disclose the unit version and composition rather than pretending the nominal unit is invariant. Disclosure at scale is a data standardization problem with a solved precedent: the unit metadata, model, version, token class, default change flags, must travel in the billing data itself, exactly the shape of problem the open cost and usage specification movement solved for cloud billing and the stated direction of its AI extension, so a firm's unit event log is partly assembled from the pipe rather than by hand.

6.5 Separate performance measurement from decision measurement

Boards and management teams routinely ask two different questions with one ROI number.

The first is ex post: how did prior AI investment perform? That requires matched vintages, realized benefits, economic capital, and a stable counterfactual.

The second is ex ante: what is the expected value of the next commitment from today's starting point? That requires marginal cost, expected incremental benefit, risk, option value, and the alternatives currently available.

The two calculations overlap, but they are not interchangeable. A program can have poor historical returns and still present a positive incremental opportunity if the remaining investment is small and the capability is now proven. A program can have excellent historical returns and a poor next dollar return if the best workloads have already been captured or marginal costs have risen.

Historical performance should inform the forecast by updating beliefs about adoption, implementation speed, cost, and value. It should not mechanically determine the scale decision.

6.6 The ratio integrity test

Before a material AI ratio is used for budgeting, incentive compensation, procurement, or investment approval, management should be able to answer seven questions.

Integrity question	Failure if unanswered	Typical repair
Is the reference stable?	Baseline or benchmark no longer represents the current economic object	Version the reference; crosswalk or break the series
Is the population comparable?	Routing or eligibility changed the case mix	State the estimand; randomize at the margin; log the assignment
Is the unit economically stable?	Token, task, or action content drifted	Normalize to outcomes or version and disclose the unit
Are cost, benefit, capital, and period matched?	Numerator and denominator belong to different vintages or owners	Reconcile the split ledger; use investment cohorts
Is this a performance measure or a decision measure?	Historical ROI is being used as the marginal investment rule	Separate ex post and ex ante views
Is the rate single valued for the unit?	Identical work is priced differently by capacity lane or commitment state, invisibly in the bill	Construct and disclose an effective rate; treat per request figures as allocations
Is the outcome definition stable and quality conditioned?	Output increased by moving rework, error, risk, or customer effort outside the measurement boundary	Incorporate quality, rework, recontact, risk, and acceptance into the outcome definition

The test is intentionally simple: not a replacement for analysis, a tripwire for the moment a familiar ratio is about to be trusted without evidence that its terms remain comparable. Three design conditions keep it from becoming attestation. It is applied by an owner who does not own the number, since every channel in section 5 applies to the test itself. Each answer is evidenced by operational diagnostics rather than assertion, population stability statistics, unit event counts, baseline age, with materiality thresholds the firm sets and files alongside the series break rules of section 7. And the test is scoped: it presumes the measure was worth selecting, the validation question the performance measurement literature owns, and audits only whether a chosen measure remains interpretable.

7. Boundary conditions, lineage, and open problems

7.1 What is not new

The paper's mechanisms have clear intellectual ancestors. Campbell and Goodhart describe the deterioration of measures once they become targets. Lucas warns that relationships estimated under one regime may not survive a policy change that alters behavior. The performativity literature adds the loop: models as engines shaping the markets they were built

to describe, the self confirming dynamic's nearest ancestor, with the engine here an operating ratio steering operating design, honest agents throughout. Selection bias and Simpson's paradox explain how aggregates can reverse or mislead when population composition differs. Hedonic methods address products whose quality changes faster than nominal units reveal. The productivity J curve explains why general purpose technologies can depress measured productivity while complementary intangible investment is being built. Software and computing markets have repeatedly changed licensing models, product quality, and price structures, and the technology benchmarking tradition itself warned long ago that spend ratios rest on unstable comparison bases, an instability it answered in its own era by diversifying the base.

The management side has its own canon, conceded with equal fullness. The strategic performance measurement literature documented firms adopting nonfinancial measures without causal models linking them to value, and validating almost nothing; the performance paradox literature showed measures losing variance and informativeness through use, gaming, selection, learning, and suppression, with firms cycling to fresh ones; the multitask literature explained output improving by pushing unmeasured work outside the boundary; the target setting literature showed ratchets institutionalizing whatever the baseline contained; and comparability itself is financial accounting's oldest virtue, with standard machinery, retrospective restatement, prospective application, disclosed breaks, for references that change. The collective departure: that canon explains measures degrading through agents' responses to being measured, and the standard setters' machinery stops at the financial statements; the mechanisms here run through the technology's own operation, with honest agents everywhere, at the operating ratio level no standard reaches.

Measurement economics itself must be conceded most fully of all, because it owns the largest share of this territory. Changing economic objects under stable labels is that profession's modern history: seventy years of hedonic adjustment; the new goods problem, handled daily as products enter and exit the baskets; chain linking, invented precisely because fixed bases mislead, the reference re drawn annually; and a permanent revision apparatus, benchmark revisions, published methodological changes, the national accounts' own production boundary re drawn generation by generation, software capitalized in 1993, R&D in 2008, data assets argued now. Two things survive that concession, and they are the paper's claim. Exogeneity inverts: statistical offices measure economies they do not operate, no census shrinks the economy by counting it, while the firm deploys the intervention that consumes its bases and sorts its populations, a measured object endogenous to the measurer, which the entire statistical apparatus is built to assume away. And the apparatus is absent: everything the profession solved, it solved with institutions, methods manuals, hedonic programs, revision cycles, methodological independence, and at the operating ratio layer the closest analog, the commercial benchmarking industry that served for decades as a private statistical office, taxonomies, peer groups, normalization rules, annual cycles, ran on the two assumptions this era breaks, annual cadence and stable units. The apparatus is not absent but outrun, which is the stronger claim.

Those precedents are not qualifications to be hidden; they establish that each mechanism has a known theoretical home. The contribution claimed here is a firm level synthesis around AI value measurement: multiple mechanisms can affect the same ratio simultaneously; the AI intervention itself can generate the change in the measurement object; the change can

occur inside ordinary governance cycles; and reliance on the distorted measure can feed back into deployment and allocation, changing the next observation.

7.2 What the construct does and does not claim

AI distortion is not a claim that AI economics are unknowable. Nor is it a claim that every AI metric is invalid. Some measures are robust precisely because their references are governed constants or because the outcome is carefully conditioned and versioned.

The construct is also not a substitute for uncertainty analysis. A perfectly specified ratio can still describe an investment with highly uncertain future value. Conversely, a low variance estimate can be systematically distorted if its population or denominator is economically mismatched. Precision and validity are different properties.

One scope condition belongs here because it bounds when the whole exercise binds. These ratios are instruments for detecting waste, which is the right instrument for a market where supply is abundant and the enemy is spending more than necessary. Practitioner reporting through 2026 describes conditions moving the other way, toward constrained capacity where access is allocated by commitment and relationship as much as by price. In such a market the binding question changes: a firm that improves its cost per unit while losing access to the units it needs has optimised a number that was no longer the one that mattered. That does not make the measures wrong, and a distorted ratio is no more useful under scarcity than under abundance. It does mean the ratio answers a question the firm should confirm is still its question.

Finally, the paper does not argue that management should always fully load shared costs into every use case. The appropriate boundary depends on the decision. The governing principle is matching: the economic objects placed on either side of a ratio should correspond to the question the ratio is intended to answer.

7.3 Adjacent work

Four bodies of current work sit beside this paper without overlapping it. The token and agent valuation literature builds frameworks for what tokens and agents contribute, marginal productivity, workflow position, and benchmark human workflows; it is the value side of the same coin, and its benchmark workflow construct is a formal ally of the counterfactual disciplines here. The national accounts debate, welfare from free goods and the construction of price indices under rapid quality change, is the macro sibling of the unit drift mechanism, cited and deliberately not joined. The evaluation literature's documented imbalance, technical metrics dominating economic ones in agentic AI assessment, describes which dimensions get measured, where this paper concerns whether the measures taken can be trusted. And the practitioner standards work now underway, cost and value definitions, full cost reference models, unit economics guidance, is the substrate this paper's disciplines are built to inform: the definitions being written this year are denominator and unit choices, and they are cheapest to harden before they set. The operating model home already exists: the FinOps practice, its mission formally broadened from cloud value to technology value, runs unit economics and cross owner cost reconciliation as established capabilities, and several of section 6's disciplines describe work that practice's AI extension will own. External pricing consumes the same distorted operating ratios, ARR based and AI native metrics feeding multiples, a

valuation spillover this paper notes and does not pursue. Closest of all is this paper's own family: the companion papers' machinery consumes exactly these ratios, a burdened cost per outcome in its cost chain, realized against claimed in its ex post loop, and the disciplines here are the terms that machinery assumes.

7.4 Open problems

Several issues remain materially unresolved.

Shared platform allocation. There is no universally correct rule for allocating common AI infrastructure, model, governance, and data costs across applications. Activity based approaches can improve transparency but may create false precision when use cases share option value and capacity.

Economic useful lives. Accounting depreciation schedules do not necessarily correspond to the economic life of models, chips, workflow integrations, or proprietary data assets. Internal ROIC requires assumptions about economic depreciation that can change rapidly as capability and prices evolve.

Option value. Foundational AI investments can create future applications that are not forecastable at approval. Ignoring that option value can create false stop decisions; invoking it without discipline can turn strategic value into an unfalsifiable plug. A useful framework must distinguish bounded experimentation from open ended justification.

Quality adjusted outcome units. The ideal denominator is often an outcome, but outcomes themselves can drift. A resolved support case, accepted code change, or completed claim needs a quality definition stable enough to preserve comparability.

Percentile governance for stochastic execution. Once agentic unit cost is treated as a distribution, the remaining question is which point of the curve to budget and price to: expected cost, a stated percentile, or outcome level contribution. The distribution itself is a discipline; the percentile is the open choice.

Market pricing on distorted units. External pricing consumes the same operating ratios, and a priced metric invites being managed to, Goodhart at market level; whether reference versioning, unit disclosure, and outcome normalization can travel into externally reported and priced measures, and who would enforce them there, is unresolved.

A venue now exists for the reference distribution problem. The external benchmarking entry above supposes a source of reference distributions that does not yet exist. Since this paper's first version, a vendor neutral body has formed under the Linux Foundation with working groups on production, consumption and value, and has published draft guidance. Two of its stated open questions are two of this paper's: who defines and who runs benchmarking, and how to value work for which no baseline exists. That converts an open problem into a live venue, and it makes the custody question, who holds sensitive comparative data and under what governance, the operative one rather than the existence question.

Detection ahead of disclosure. Methods for establishing that an endpoint's behavior has changed are published, cheap, and in several cases released as working code. There is no agreed vocabulary for what constitutes a material change, no standard obliging its disclosure, and no convention for reporting instrument stability alongside a ratio that depends on it. The gap is in standards rather than in technique, which suggests the constraint is institutional

and that the venue for closing it is a standards body rather than a research programme.

External benchmarking. Cross vendor comparisons become increasingly difficult when commercial units are incompatible and vendor capabilities evolve at different rates. The benchmark may need to move from price per vendor unit to a standardized workload and quality adjusted outcome. Pooled practitioner community benchmarking is the plausible source for the reference distributions this problem and the empirical validation problem both require.

Empirical validation. The disciplines are stated as testable propositions: firms adopting reference versioning, population conditioning, and vintage matching should show smaller gaps between claimed and realized value and fewer measurement driven reversals than firms that do not; whether they do is an open association question in the performance measurement tradition.

Series break rules. Firms need practical thresholds for when a historical series should be restated, crosswalked, or broken. Too many breaks destroy continuity; too few preserve fiction. Official statistics' re basing and revision practice is the working model.

These open problems do not weaken the case for rebuilding ratios. They define where the rebuilding process must remain explicit about assumptions rather than manufacturing comparability.

8. Conclusion: the arithmetic survives

The three numbers in the opening did not lie because their observations were false. They lied because management assumed that the same label continued to describe the same economic object.

The January baseline was accurate but no longer represented the September process. The assistant and human service averages were accurate but described populations the AI system had sorted differently. The procurement benchmark was accurate for the commercial unit it measured, but that unit no longer captured the way the product was being consumed.

This distinction matters because firms are moving quickly from AI experimentation to AI governance. Return measures, unit economics, productivity ratios, procurement benchmarks, and budget controls will increasingly determine which systems are scaled, which are stopped, which managers are rewarded, and which vendors are selected. Once those measures become institutionalized, their assumptions become part of the operating system of the firm.

Conventional financial arithmetic remains useful. ROI, ROIC, NPV, payback, cost per outcome, and productivity can all support sound decisions. But their validity depends on relationships that the formulas themselves cannot test. The formula does not know that the population self sorted, that the denominator shrank because of the intervention, that the task changed meaning, that the commercial unit migrated, or that today's benefits belong to yesterday's capital.

That is the core of AI distortion. The failure is not arithmetic. It is unobserved change in the relationship among the things being counted.

The response is therefore constructive rather than nihilistic. Version the reference. Condition the population. Match costs, benefits, capital, and time. Normalize to stable outcomes. Separate historical performance from the economics of the next decision. Where continuity

cannot be defended, break the series rather than preserve a false comparison.

Reliable value measurement requires more than accurate observations. It requires stable relationships, or measurement systems deliberately engineered to reveal when those relationships have changed.

AI does not make ratios obsolete. It makes their construction consequential.

References

Factual claims are stated as of 25 August 2026 and verified against the project's research base on that date. Fast moving market figures carry their approximate dates in text and are cited as directional evidence; none is load bearing alone.

404 Media. 2026. Reporting on Microsoft's internal AI usage measures and division token budgets. August 2026, with corroborating coverage.

Abdirahman, M., D. Coyle, R. Heys, and W. Stewart. 2022. "A Comparison of Approaches to Deflating Telecoms Services Output." *Economie et Statistique / Economics and Statistics* 530-31.

Albrecht, A. J. 1979. "Measuring Application Development Productivity." *Proceedings of the IBM Applications Development Symposium*, 83-92.

Appenzeller, G. 2024. "Welcome to LLMflation: LLM Inference Cost Is Going Down Fast." Andreessen Horowitz, a16z.com.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111(12): 4088-4118.

Athey, S., and G. W. Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27): 7353-7360.

Athey, S., and S. Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89(1): 133-161.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang. 2019. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects." NBER Working Paper 26463.

Atlassian. 2026. Fiscal 2026 earnings materials and coverage of the first reported seat decline.

Boston Consulting Group. 2026. Analysis of hybrid AI and human support programs.

Brynjolfsson, E., A. Collis, W. E. Diewert, F. Eggers, and K. J. Fox. 2025. "GDP-B: Accounting for the Value of New and Free Goods." *American Economic Journal: Macroeconomics* 17.

Brynjolfsson, E., D. Rock, and C. Syverson. 2021. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13(1): 333-372.

Campbell, D. T. 1976. "Assessing the Impact of Planned Social Change." Occasional Paper 8, Public Affairs Center, Dartmouth College.

Chen, L., M. Zaharia, and J. Zou. 2024. "How Is ChatGPT's Behavior Changing over Time?" *Harvard Data Science Review* 6(2).

Corrado, C., C. Hulten, and D. Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661-685.

Coyle, D. 2014. *GDP: A Brief but Affectionate History.* Princeton University Press.

Coyle, D. 2025. *The Measure of Progress: Counting What Really Matters.* Princeton University Press.

Coyle, D., and D. Nguyen. 2018. "Cloud Computing and National Accounting." ESCoE Discussion Paper 2018-19, Economic Statistics Centre of Excellence.

Decagon. 2026. Support metrics glossary. decagon.ai.

Epoch AI. 2025-2026. LLM inference price trends. Public data series, epoch.ai.

Evaluation review. 2026. "Measurement Imbalance in Agentic AI Evaluation." Systematic review of 84 studies, arXiv preprint.

Erol, M. H., and coauthors. 2025. "Cost-of-Pass: An Economic Framework for Evaluating Language Models." arXiv:2504.13359.

FASB. *Accounting Standards Codification* Topic 250: Accounting Changes and Error Corrections. Financial Accounting Standards Board.

FinOps Foundation. 2024-2026. *FOCUS: The FinOps Open Cost and Usage Specification*, versions 1.0 through 1.4. focus.finops.org.

FinOps Foundation. 2025. *The FinOps Framework.* finops.org.

FinOps Foundation. 2026. Practitioner reporting on AI spend against budget. Mid 2026 briefing.

Forrester Research. 2025. *Predictions 2026.* October 2025.

Forrester Research. 2026. *2027 Budget Planning Guides.* July 15, 2026.

Gao, I., P. Liang, and C. Guestrin. 2025. "Model Equality Testing: Which Model Is This API Serving?" arXiv:2410.20247.

Gao, Y., M. Wang, and Y. L. Yu. 2026. "Token Arena: A Continuous Benchmark Unifying Energy and Cognition in AI Inference." arXiv:2605.00300. The authors operate the benchmark described.

Gartner. 2026. Survey and analyst coverage: CFO AI budget alignment, August 2026; agentic token multiplier estimates; junior tier hiring analyses.

Fuller, M. 2026. "FinOps and Tokenomics: How Do They Fit Together?" FinOps Foundation Insights, 25 August 2026. Licensed CC BY 4.0.

Goodhart, C. A. E. 1975. "Problems of Monetary Management: The U.K. Experience." In *Papers in Monetary Economics*, Reserve Bank of Australia.

Hausman, J. A. 1996. "Valuation of New Goods under Perfect and Imperfect Competition." In *The Economics of New Goods*, eds. Bresnahan and Gordon. University of Chicago Press.

Hicks, J. R. 1940. "The Valuation of the Social Income." *Economica* 7(26): 105-124.

Holmstrom, B., and P. Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7: 24-52.

IASB. 2018. *Conceptual Framework for Financial Reporting.* International Accounting Standards Board.

Ittner, C. D., and D. F. Larcker. 1998. "Are Nonfinancial Measures Leading Indicators of Fi-

nancial Performance? An Analysis of Customer Satisfaction." *Journal of Accounting Research* 36 (Supplement): 1-35.

Ittner, C. D., and D. F. Larcker. 2003. "Coming Up Short on Nonfinancial Performance Measurement." *Harvard Business Review* 81(11): 88-95.

Leshin, J., M. Shah, I. Timmis, and D. Kang. 2026. "Behavioral Fingerprints for LLM Endpoint Stability and Identity." arXiv:2603.19022. The authors operate the monitoring system described.

Lev, B., and F. Gu. 2016. *The End of Accounting and the Path Forward for Investors and Managers.* Wiley.

Linux Foundation. 2026. Tokenomics Foundation launch materials. August 4, 2026, linuxfoundation.org.

Lucas, R. E. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1: 19-46.

MacKenzie, D. 2006. *An Engine, Not a Camera: How Financial Models Shape Markets.* MIT Press.

McKinsey & Company. 2025-2026. *The State of AI* survey waves.

Meyer, M. W., and V. Gupta. 1994. "The Performance Paradox." *Research in Organizational Behavior* 16: 309-369.

Murphy, J. E. 2026a. "AI Capital Allocation and Governance: A Portfolio Model." Working paper. ssrn.com/abstract=7095978.

Murphy, J. E. 2026b. "No Exchange Rate: Carbon in the Economics of AI Compute." Working paper. ssrn.com/abstract=7216759.

Nordhaus, W. D. 2007. "Two Centuries of Productivity Growth in Computing." *Journal of Economic History* 67(1): 128-159.

NVIDIA. 2026. GTC 2026 keynote remarks on engineer token budgets.

Patterns. 2023. Synthesis review of metric failure, drawing the Campbell, Goodhart, and Lucas lineage. Cell Press.

Poyar, K. 2026. Survey of some 230 software companies' pricing models. Growth Unhinged.

PricingSaaS. 2026. Pricing and packaging change tracker, the 500 largest software companies.

Rubin, H. A. Circa 2009. The technology intensity method: technology spend measured against revenue and operating cost together. Practitioner benchmarking literature.

Salesforce. 2026. Agentforce pricing materials and 2026 earnings coverage.

ServiceNow. 2026. Earnings coverage on the share of new business priced other than by seat.

Simpson, E. H. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society, Series B* 13(2): 238-241.

Stanford HAI. 2026. *Artificial Intelligence Index Report 2026.* Stanford Institute for Human-Centered AI.

Storment, J. R., and M. Fuller. 2023. *Cloud FinOps: Collaborative, Real-Time Cloud Value*

Decision Making. 2nd ed. O'Reilly Media.

"The Price of Progress: Inference Prices and Frontier Capability." 2026. arXiv:2511.23455.

Wager, S., and S. Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523): 1228-1242.

Zhu, and coauthors (New York University Tandon). 2026a. "AI Tokenomics." arXiv:2606.24616.

Zhu, and coauthors (New York University Tandon). 2026b. "Agentomics." arXiv:2606.14769.

Zylo. 2026. Buyer survey on consumption pricing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used large language model assistants, including Anthropic's Claude and OpenAI's ChatGPT, for drafting support, source-verification support, critical review, and editorial iteration under the author's direction. After using these tools, the author reviewed and edited all content, independently verified the sources and the claims they support, and takes full responsibility for the content of this publication.

Comments are invited; this is an explicitly provisional working draft.