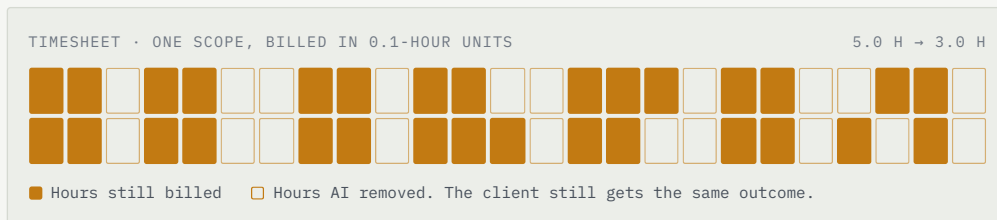


THE PARADOX

Get 40% faster at your job. Earn 40% less for doing it.

Beyond the Billable Hour. That drop is a property of the arithmetic of billing by the hour, not a market condition. This paper shows why AI makes that arithmetic self-defeating, what the controlled evidence says, and what breaks in the operating model underneath.

Bryan Musto CHIEF REVENUE OFFICER



00

EXECUTIVE
SUMMARY

Consulting has priced time for as long as the modern firm has existed. That made sense when expertise and execution capacity were scarce. A configuration took a week because it took a week, and hours billed were a fair proxy for value delivered.

AI removes that scarcity. When a team can configure, test, migrate data and document in a fraction of the hours the same scope used to need, a model built on hourly billing starts to punish the firms that get better at the job. Under time and materials, the firm that adopts AI fastest earns less for the same outcome.

40%

fewer hours to reach the same outcome means 40% less revenue under time and materials.

1.67×

the hourly rate a firm would need just to bill the same total. Clients will not pay it.

0%

of the value the client receives disappears. Only the hours do.

This paper argues that the shift toward fixed-fee, outcome-linked and risk-shared models is the necessary response to a delivery model where AI compresses effort while the value delivered stays the same. It draws on controlled field experiments, contract theory, and the history of other industries that made the same move. The argument in eight lines:

- I The hour was a safe unit because it didn't shrink.** For sixty years consulting was a textbook case of Baumol's cost disease: prices rose with wages because productivity barely moved.²
- II AI is the first real productivity shock to knowledge work,** and controlled studies now measure it: 25% to 56% less time on in-scope tasks.^{3,4,5}
- III The gains are jagged.** Some workstreams collapse, others barely move, and outside the frontier AI makes experts worse.^{3,8}
- IV Under hourly billing, every efficiency gain transfers to the client.** Contract theory has predicted this since 1979.¹²
- V Under outcome pricing, delivery variance replaces utilization as the source of margin.** The firm that is most predictable can charge the least and still earn the most.
- VI Other industries made this move before us,** from jet engines to energy retrofits to Medicare. Each needed measurable outcomes, instrumented delivery data, and a balance sheet able to carry risk.
- VII The disruption arrives in waves, with 2027 to 2029 decisive.** A measurable trend in AI task length, plus historical adoption lags, lets firms plan against dates instead of hype.^{25,26}
- VIII The pyramid compresses,** and with it utilization, leverage, graduate hiring, partner economics and how the market values the firm.

01

WHERE WE
STARTED

The model time built

The economics of traditional consulting, and systems integration in particular, rest on a simple structure: a pyramid of staff at different seniority and rate levels, billing hours against a project, with margin coming mainly from utilization rather than outcome. A partner or architect designs the solution. A larger base of mid-level and junior staff builds, tests and documents it. The firm bills for all of it.

David Maister formalized this in 1982. Profit per partner is the product of margin, productivity and *leverage*, the ratio of staff to partners.¹ He also sorted engagements into three types. **Brains** work is novel and bought for expertise. **Grey-hair** work is bought for experience. **Procedure** work is familiar, repeatable and bought for efficient execution. Procedure work supports the widest pyramids, because it is the work juniors can do.

AI attacks procedure work first. That is exactly where the pyramid keeps its base.

None of this was cynical. It reflected a real constraint. Knowledge work in enterprise software was labor-intensive in a way that resisted compression. Data mapping, regression testing, configuration and documentation took roughly the time they took, and clients could audit inputs even when they couldn't audit output quality in real time.

Economists have a name for sectors like this. William Baumol showed in 1967 that when a service's productivity stays flat while wages rise across the economy, its price must keep rising.² The string quartet needs four players in 2026 just as it did in 1826. For decades consulting behaved the same way. The hour was a stable unit because an hour of configuration work produced about the same output year after year, so rate cards could rise with wages and nobody had to ask what an hour was worth.

That constraint is what's changing.

02

THE
EVIDENCE

What controlled studies actually show

Most commentary on AI productivity is anecdote or vendor marketing. The claims in this paper rest on something firmer: randomized and field experiments with real professionals doing real work. One of the most cited studies used consultants from a strategy firm as its subjects.

In 2023, researchers from Harvard Business School, Wharton and MIT ran a preregistered experiment with 758 Boston Consulting Group consultants.³ On tasks inside AI's capability frontier, consultants using GPT-4 finished 12.2% more tasks, 25.1% faster, at more than 40% higher quality. On a task deliberately designed to sit just outside that frontier, consultants with AI were 19 percentage points *less* likely to reach the correct answer than those working without it. The authors called this the **jagged technological frontier**.

FIGURE 1

The measured effect of generative AI on professional work

Controlled and field studies, each on its own metric. Positive values are an improvement for the worker using AI. The lab effects are large; the two negative results and the small job-level figure are why delivery has to be measured before it is priced.

■ Improvement ■ Deterioration

Coding with Copilot

Peng et al. 2023 · n = 95 · time to complete

56% faster

Professional writing tasks

Noy & Zhang 2023 · n = 453 · time per task

40% faster

Developers at three firms

Cui et al. 2025 · n = 4,867 · tasks completed

+26% tasks

BCG consultants, inside the frontier

Dell'Acqua et al. 2023 · n = 758 · time per task

25% faster

Customer-support agents

Brynjolfsson, Li & Raymond 2025 · n = 5,179 · resolutions/hour

+14%

Workers across 11 exposed occupations

Humlum & Vestergaard 2025 · n ≈ 25,000 · hours saved

2.8% saved

Experienced developers, own codebases

Becker et al. / METR 2025 · n = 16 · time to complete

19% slower

BCG consultants, outside the frontier

Dell'Acqua et al. 2023 · correct answers

-19 pts

-20% 0% +20% +40% +60% +80%

Sources: Noy & Zhang (2023)⁴; Peng et al. (2023)⁵; Dell'Acqua et al. (2023)³; Cui et al. (2025)⁶; Brynjolfsson, Li & Raymond (2025)⁷; Humlum & Vestergaard (2025)⁹; Becker et al., METR (2025)⁸. Time reductions are shown as positive effects. Metrics differ by study and are not directly comparable; the figure shows direction and order of magnitude.

Three findings matter for the commercial argument.

First, the task-level gains are large and replicated. Professionals writing business documents cut time by 40% and raised quality by 18%.⁴ Developers completed a standard coding task 55.8% faster.⁵ Across nearly 5,000 developers at Microsoft, Accenture and a Fortune 100 company, AI access raised completed tasks by 26%.⁶ These are the activities that fill the base of a delivery pyramid.

Second, the gains are not automatic. In a 2025 randomized trial, experienced open-source developers working in their own mature codebases were 19% *slower* with AI tools, while believing they had been 20% faster.⁸ A study of 25,000 Danish workers found that chatbots saved an average of 2.8% of work hours, with no measurable effect on earnings.⁹ Task-level speed does not become firm-level economics unless the work is redesigned around it.

Third, perception is unreliable. If developers misjudge their own speed by nearly 40 points, a firm cannot price outcomes on intuition. It has to instrument delivery and price from the data. That requirement shapes the rest of this paper.

03

UNEVEN
COMPRESSION

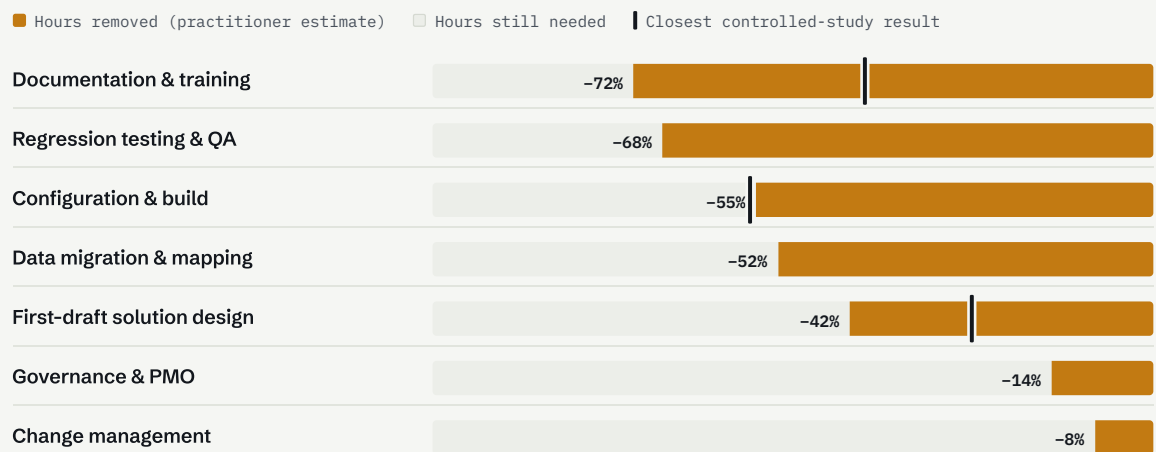
What AI changes in delivery

The change isn't that consultants type faster. Entire categories of manual, hour-consuming work are collapsing, and they are not collapsing evenly. The workstreams that compress most are the codified ones: documentation, testing, configuration, data mapping. The ones that barely move depend on trust, persuasion and accountability.

FIGURE 2

How far each workstream compresses

Every row starts at the same pre-AI effort. The shaded segment is the effort that disappears in AI-augmented delivery. The black tick marks the closest controlled-study result, where one exists.



Practitioner estimates are directional, not a benchmark, and reflect agentic tooling applied across a whole workstream. Controlled-study ticks measure single tasks: writing time -40%⁴, coding time -55.8%⁵, consulting-task time inside the frontier -25.1%³.

The pattern lines up with the evidence. Codified work, where output can be checked against a specification, compresses sharply. Judgment-heavy work that sits outside the frontier does not, and that is where AI can mislead an unwary team. None of this eliminates the need for skilled architects, delivery leads or change management. Judgment, client context and risk ownership stay human work. But the hours needed to reach a given outcome are falling, and falling at different rates in each workstream.

That unevenness matters commercially. A blended hourly rate hides it. A priced outcome exposes it, and rewards the firm that knows its own compression curve better than its competitors do.

04

THE BROKEN
INCENTIVE

Why time and materials can't survive this

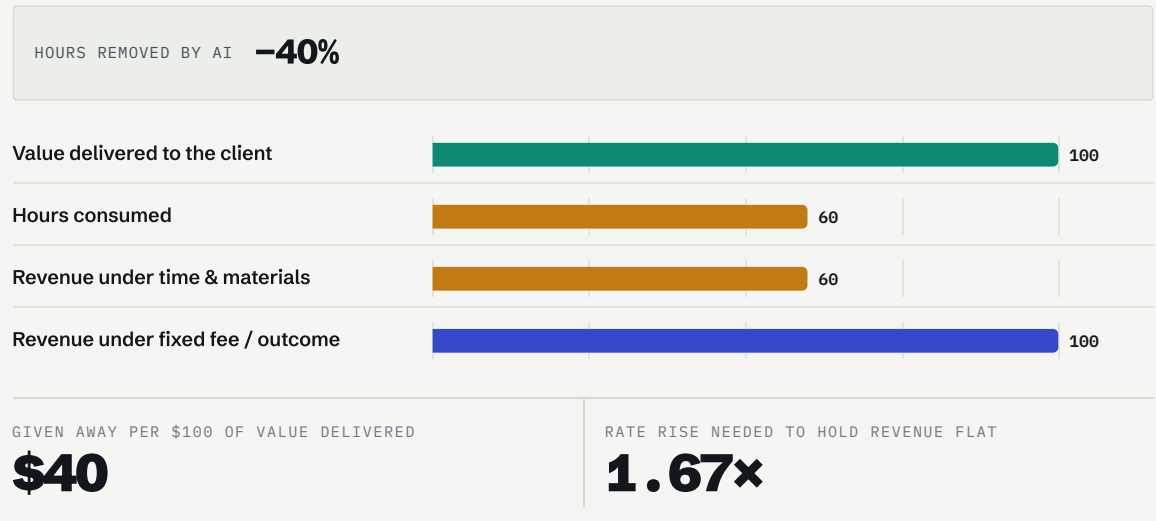
Under time and materials, a firm that uses AI to deliver a project in 40% fewer hours also earns about 40% less revenue for the same outcome. It has two escape routes. It

can raise rates, which clients resist, or it can expand scope to absorb the freed capacity, which erodes trust and looks like padding.

FIGURE 3

The efficiency trap

Shown at a 40% efficiency gain. The value delivered to the client holds flat. Under time and materials, revenue falls with the hours.



Index: pre-AI engagement = 100. Rate multiplier = $1 \div (1 - \text{share of hours removed})$.

A billing model that pays for hours discourages the firm from using the very technology that makes it better at its job.

This is not a new insight. It is one of the oldest results in contract theory. Bengt Holmström showed in 1979 that principals pay for inputs when outputs are hard to observe, and that such contracts are inefficient whenever better output signals exist.¹² Hourly billing is a contract on observable effort. It made sense when clients couldn't judge configuration quality in real time and effort was a fair proxy for progress. AI breaks the proxy, because effort and output are no longer tied together. At the same time AI makes output easier to observe, since automated testing, telemetry and data-quality tooling can now verify what "done" means. Theory predicts the result: contracts should migrate from inputs to outcomes.

Healthcare offers the closest parallel. Fee-for-service medicine paid for procedures rather than health, and for decades the incentive to add volume was a central critique of the system.²² Consulting's hourly model is fee-for-service by another name.

It is also a form of the innovator's dilemma. Christensen and colleagues warned in 2013 that consulting was on the cusp of disruption, because modular, productized offerings could unbundle the traditional engagement.¹⁴ The firms with the most reason to resist AI-driven efficiency are those whose commercial model depends on hours, even as their clients expect AI to be part of delivery. Sophisticated buyers

already ask why a team using AI-assisted delivery still bills like it's 2015. That question will get louder.

05

WHO KEEPS
THE GAIN

The surplus goes somewhere

When AI removes 40 hours from a 100-hour engagement, it creates a surplus. The commercial model decides who keeps it. The figure below follows one engagement through four states, using a simple assumption: a 40% delivery margin before AI, and AI tooling and reusable IP that cost the firm 4 points of revenue to run.

FIGURE 4

Where the efficiency surplus lands, per \$100 of pre-AI value

Each bar is \$100 of value delivered to the client. Segments show what the client pays for delivery cost, what the firm keeps as profit, and what the client keeps as a price reduction.

■ Delivery cost ■ Firm profit ■ Client saving

Before AI • time & materials

Price \$100 • firm profit \$40



With AI • time & materials

Price \$60 • firm profit \$24



With AI • fixed fee, year one

Price \$100 • firm profit \$60



With AI • fixed fee after competition

Price \$82 • firm profit \$42



Illustrative model. Pre-AI: price \$100, cost \$60. With AI, hours fall 40% and labor cost falls to \$36; the IP-led firm adds \$4 of tooling cost. The competitive fixed fee assumes rivals force the price down to \$82.

The last row matters most, because it answers the obvious objection. Competition will not let a firm keep every dollar of the surplus for long. Prices for a defined outcome will fall as more firms can deliver it. But even after competition takes its share, the outcome-priced firm earns **\$42** on the engagement while the time-and-materials firm earns **\$24**. The client is better off in both cases. Outcome pricing does not let a firm hoard the gain. It lets the firm share the gain on its own terms instead of surrendering all of it by default.

06

VARIANCE IS
THE NEW MARGIN

Why predictability, not speed, wins fixed-fee work

Here is the point most commentary on outcome pricing misses. When a firm prices a fixed fee, it is not pricing its average effort. It is pricing its *tail*. The fee has to cover the engagement that runs long, and in enterprise technology, long tails are the norm.

1 in 6

IT projects in Flyvbjerg and Budzier's study of 1,471 projects was a "black swan", with an average cost overrun of 200% and a schedule overrun of almost 70%. The average overrun across all projects was 27%.¹⁵

A firm that estimates each engagement from scratch carries a wide, right-skewed effort distribution. To price a fixed fee safely it must cover, say, the 80th percentile of that distribution, which puts it well above its own median. A firm with reusable accelerators and instrumented delivery has a narrower distribution. It can price closer to its median, undercut the bespoke firm, and still earn more.

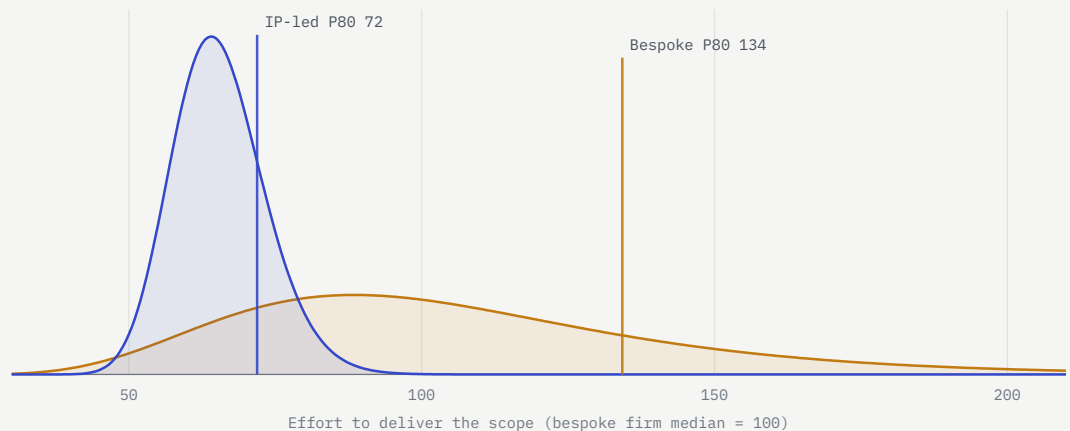
FIGURE 5

Two firms bid the same fixed-fee scope

Distribution of effort to deliver the same outcome. Each firm must price at or above its 80th-percentile effort to be profitable on four engagements out of five. Shown with the IP-led firm at $\sigma = 0.12$.

IP-LED FIRM'S DELIVERY VARIANCE (Σ) **0.12**

■ Bespoke firm: median 100, σ 0.35 ■ IP-led, AI-augmented firm: median 65



BESPOKE FIRM · SAFE FIXED FEE (P80)

134

IP-LED FIRM · SAFE FIXED FEE (P80)

72

Model: lognormal effort distributions, a standard shape for project cost overruns. Units are effort indexed to the bespoke firm's median = 100. $P80 = \text{median} \times e^{0.8416\sigma}$. Parameters are illustrative.

This changes what a consulting firm competes on. Under hourly billing the firm with the best people, fully utilized, wins. Under outcome pricing the firm with the **tightest delivery distribution** wins, because it can price aggressively without betting the margin on luck.

It also names the main risk of moving too early. Oil-industry engineers described the **winner's curse** in 1971: in competitive bidding under uncertainty, the winner is disproportionately the bidder who underestimated cost the most.¹⁶ A firm that moves to fixed fees without variance data will win the bids it should have lost. The answer is to sequence the transition, which Section 10 lays out.

The emerging alternatives

There isn't a single outcome-based model. There is a spectrum, and different engagement types belong at different points on it. Two properties decide where: how measurably "done" can be defined, and how predictably the firm can deliver it.

LOWER RISK

Fixed fee / value-based

A defined scope and set of deliverables priced as one fee, independent of hours worked. Firms with low delivery variance can price this with confidence; firms estimating from historical hourly benchmarks cannot.

MEASURED

Outcome-linked milestones

Payment tied to measurable milestones: go-live by a date, an SLA met in production, a data-migration accuracy threshold. Payment follows results rather than hours logged.

SHARED UPSIDE

Gain-share / risk-share

The firm takes a share of a measured business result: shorter loan-origination cycle time, faster quote-to-close, recovered revenue leakage. Fully aligned, but both parties must agree on measurement and attribution. Collars and caps keep the risk bounded.

ONGOING

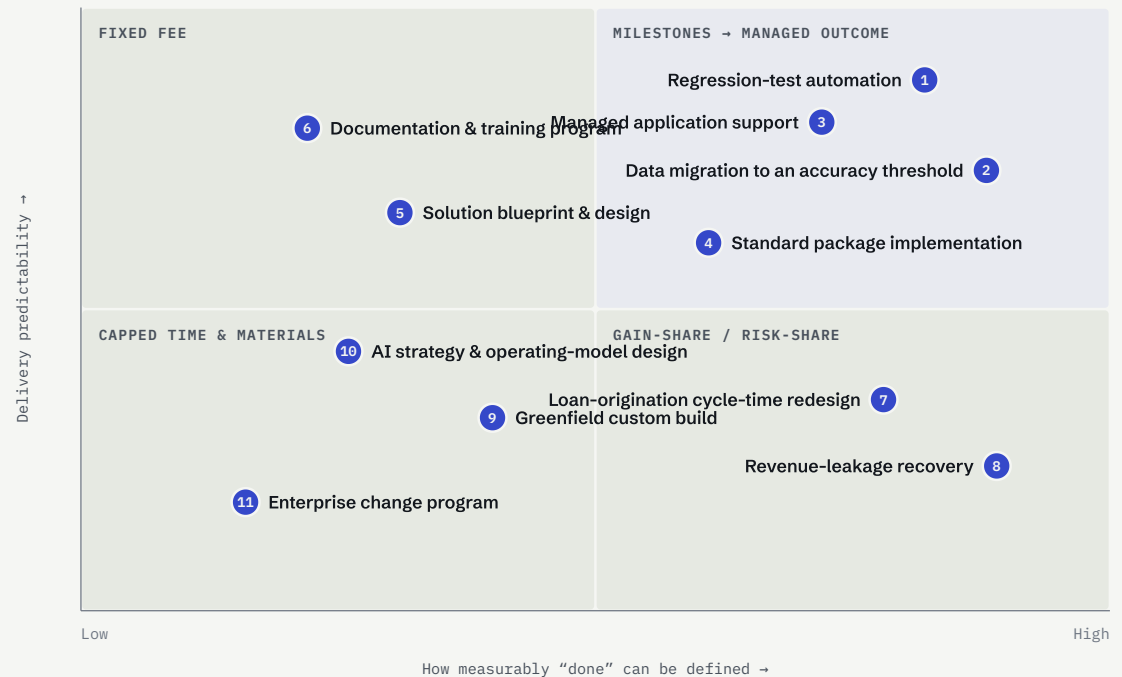
Managed outcome subscription

A recurring fee tied to the sustained performance of a system or process, such as uptime, accuracy or adoption. Delivery becomes a managed service with performance guarantees built in.

FIGURE 6

The pricing readiness map

Where common engagement types sit on the two dimensions that decide how they can be priced. Move right by defining measurable outcomes; move up by building reusable IP and delivery data.



- | | |
|--|---|
| 01 Regression-test automation
Managed outcome | 02 Data migration to an accuracy threshold
Outcome milestones |
| 03 Managed application support
Managed outcome | 04 Standard package implementation
Outcome milestones |
| 05 Solution blueprint & design
Fixed fee | 06 Documentation & training program
Fixed fee |
| 07 Loan-origination cycle-time redesign
Gain-share with collar | 08 Revenue-leakage recovery
Gain-share |
| 09 Greenfield custom build
Capped T&M | 10 AI strategy & operating-model design
Capped T&M |
| 11 Enterprise change program
Capped T&M | |

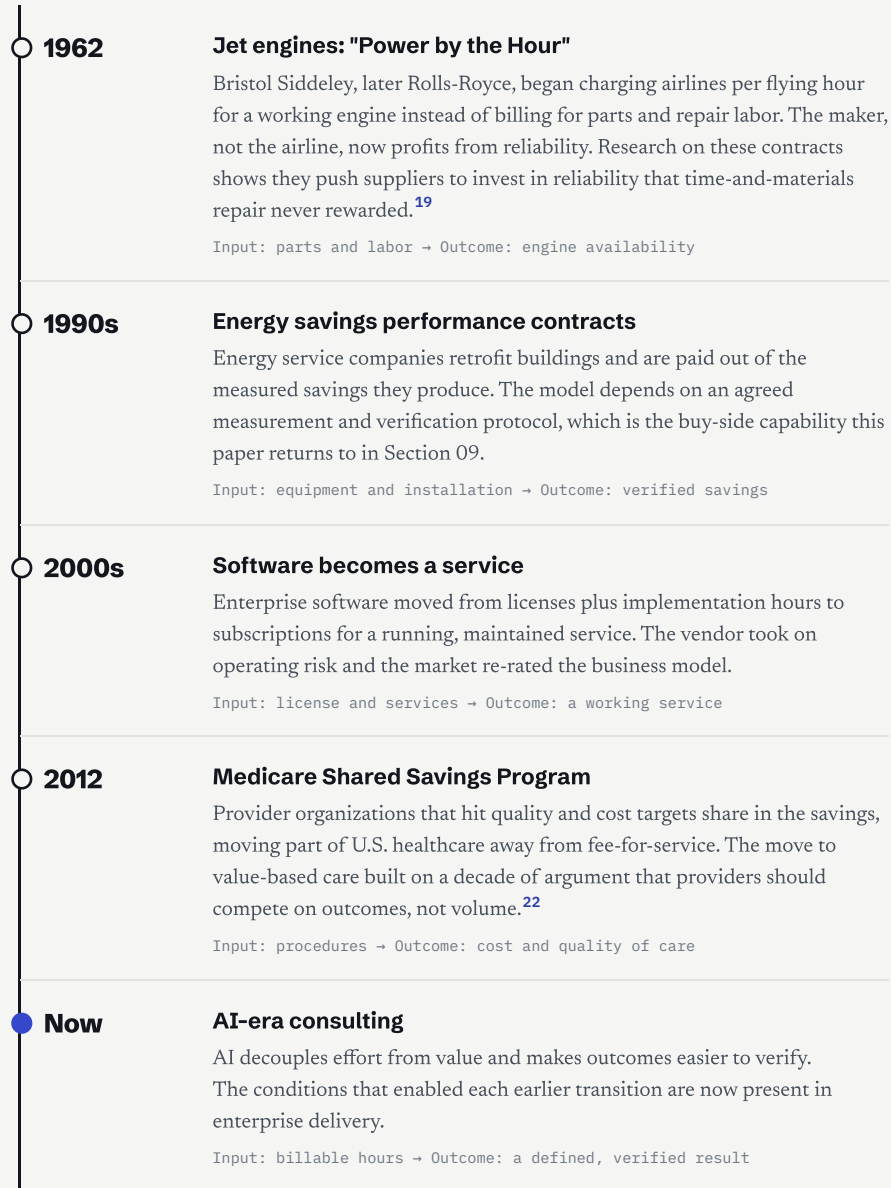
Placements are the author's judgment, offered as a starting framework for classifying a firm's own portfolio.

The map also says what not to do. Holmström and Milgrom showed that when an agent has several tasks and only some are measurable, strong incentives on the measured task pull effort away from the rest.¹³ Put a gain-share on a change program with a fuzzy outcome and you get a firm that games the metric. The bottom-left of the map should stay on time and materials, with caps, until the outcome can be defined. Outcome pricing is a discipline of choosing where it fits, not a slogan applied everywhere.

Other industries made this move first

Consulting is late to this transition, not first. Several industries have moved from paying for inputs to paying for results, and the pattern is consistent enough to learn from.

FIGURE 7 • FROM PAYING FOR INPUTS TO PAYING FOR RESULTS



The common pattern: every transition needed three things. An outcome both sides could measure, delivery data good enough to price risk, and a party with the balance sheet to carry it. Research on defense outcome contracts finds the same prerequisites.²⁰

What this requires of firms, and of buyers

Moving toward outcome pricing is not only a sales or contracting change. It requires real changes to how a firm operates.

Productized, reusable IP. Accelerators and pre-built industry configurations reduce delivery variance enough to price outcomes with confidence, instead of treating every engagement as a fresh estimate. Section 06 shows why this is the core asset.

A different financial model. Firms need the balance sheet and risk appetite to underwrite fixed-fee and gain-share risk, instead of passing all delivery-time risk to the client by the hour. Building reusable IP is an intangible investment, and Brynjolfsson, Rock and Syverson show that such investments depress measured returns before they raise them.²¹ Plan for the J-curve.

A different talent model. Smaller, more senior, AI-augmented delivery teams replace large pyramids of junior staff whose main value was billable hours.

A different sales motion. Selling a defined outcome and a commercial structure around it, instead of a rate card and an estimate of hours.

Firms that already bring proprietary, productized IP into engagements have a real head start, because their delivery predictability is already close to what outcome pricing requires.

The buyer's side of the table

The shift isn't one-sided. Clients have work to do as well.

Define "done" and "good" in measurable terms: SLAs, accuracy thresholds, cycle-time targets, not just a list of activities and a level of effort.

Evolve procurement and legal processes built around hourly rate cards toward frameworks that can evaluate and contract for outcome-based and risk-shared structures.

Build the capability to measure outcomes. Gain-share and outcome milestones only work if both sides trust the measurement.

There is also a behavioral reason buyers will accept this. Marketing research has documented a consistent **flat-rate bias**: customers regularly choose a fixed price even when paying by usage would cost them less, because it removes uncertainty and the discomfort of a running meter.^{17,18} An hourly engagement is a taxi meter on a strategic program. A firm that removes the meter is selling certainty, and certainty commands a premium, provided the firm's variance lets it afford one.

Buyers who build this capability early will get better terms and better-aligned partners than those who keep buying consulting the way they did a decade ago.

10

SEQUENCING

A practical path forward

For firms making this shift, a phased approach works better than converting everything at once. Each phase ends at a gate: evidence the firm should have before it takes on more risk.

FIGURE 8 • FOUR PHASES, EACH WITH AN EVIDENCE GATE

PHASE 1	PHASE 2	PHASE 3	PHASE 4
Pilot on bounded scope Fixed fee on a well-understood engagement type with accelerators behind it, before extending to novel scopes. GATE TO PHASE 2 A handful of fixed-fee engagements delivered, with actual effort tracked against the bid.	Instrument delivery Capture actual effort by workstream and the drivers of variance, meaning what separated the smooth projects from the rest. GATE TO PHASE 3 A measured P50 and P80 effort per offering, and a known compression curve per workstream.	Formalize offerings Accelerator-driven packages with defined scope, defined outcomes, and fixed-fee-plus-contingency pricing. GATE TO PHASE 4 The P80-to-P50 ratio narrowing release over release, and outcome metrics the client already trusts.	Extend risk-share Gain-share structures on strategic accounts, once there is enough delivery data to underwrite that risk responsibly. ONGOING DISCIPLINE Every gain-share carries a collar, an agreed baseline, and an independent measurement method.

“I’ve spent the better part of two decades on both sides of this transition, inside global systems integrators and inside the institutions buying their services. I don’t think outcome-based pricing is a nice-to-have positioning statement. I think it’s the only commercial model that survives AI-driven delivery in the long run.”

BRYAN MUSTO

11

WHAT BREAKS
UNDERNEATH

The second-order effects

If the commercial model changes, the operating model underneath it can’t stay the same. A firm doesn’t swap its billing mechanics and leave staffing, career paths, governance, deal-making and its market valuation untouched.

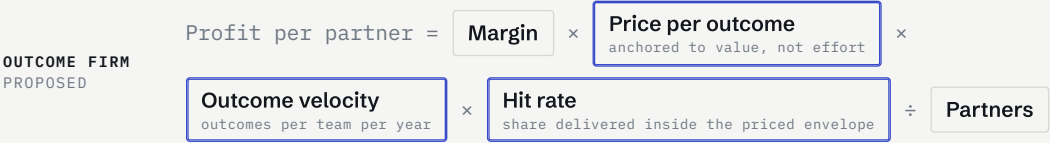
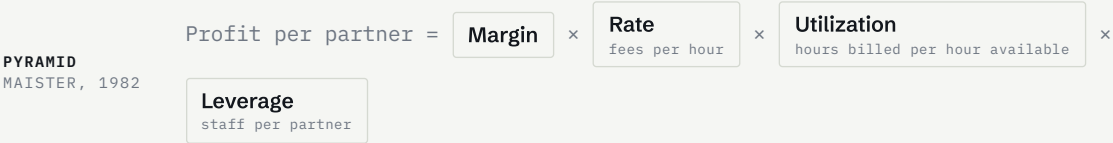
Utilization

Utilization has been consulting’s central profitability lever for decades: keep billable staff busy and margin follows. That lever weakens under outcome pricing, because

revenue no longer depends on hours logged. What replaces it is closer to outcome velocity: how quickly and predictably a team turns a defined scope into a defined result. A firm can be fully utilized and still lose money on a fixed-fee engagement that runs long.



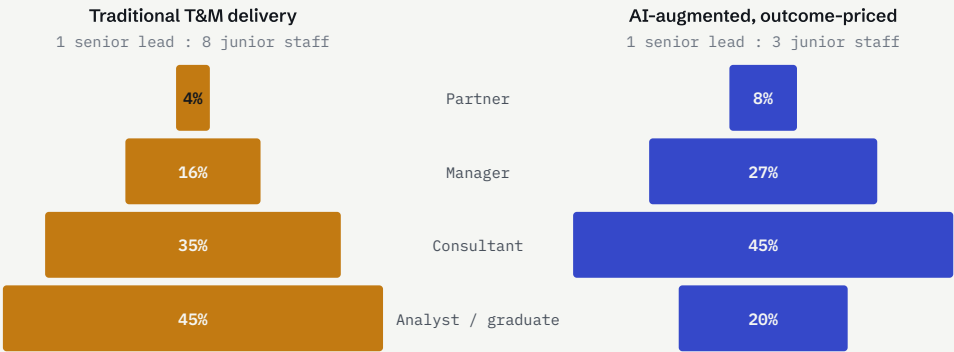
The same logic rewrites Maister's profit equation. Two of its terms lose their force, and two new ones take over.



Leverage ratios

Traditional leverage, a few senior staff supervising a much larger base of juniors billing hours, followed from the same economics as utilization: junior hours were profitable hours. When delivery is AI-assisted and priced on outcome, the profitable unit of work shifts toward judgment and the configuration of AI tooling. The pyramid becomes a diamond.

FIGURE 9
From pyramid to diamond: delivery headcount mix by level
Share of delivery headcount at each level, and the resulting ratio of junior staff to each senior lead.



Illustrative and directional, not a market forecast.

Hiring graduates

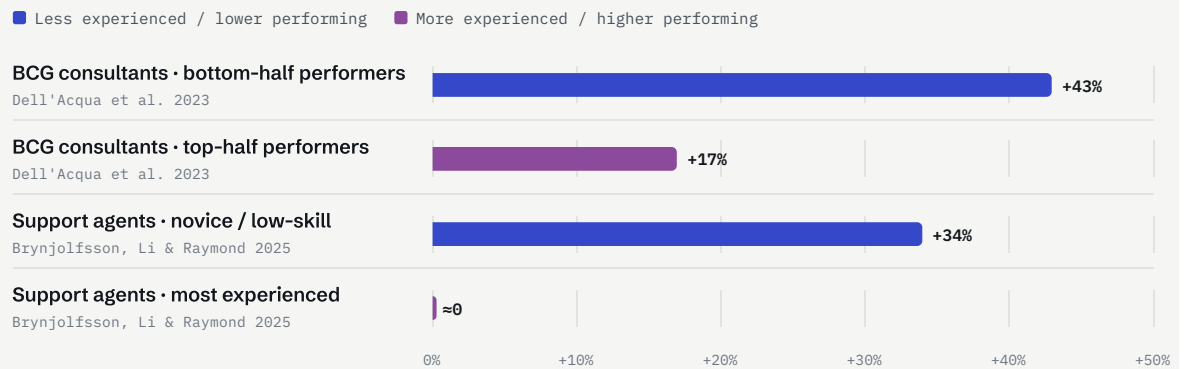
The apprenticeship model worked like this: hire large graduate classes, let them learn the craft on billable projects, promote the best over several years. It was subsidized by the fact that junior hours were profitable to bill. If junior work compresses, so does the volume of entry-level hiring the model can support.

The evidence adds a twist I call the **apprenticeship paradox**. AI helps the least experienced workers most. In the BCG study, below-average performers improved 43% with AI, against 17% for above-average ones.³ Among customer-support agents, novices gained 34% while the most experienced saw almost no change.⁷ AI compresses the skill gap between junior and senior. That makes juniors more productive, and it also weakens the seniority-tiered rate card that made them profitable.

FIGURE 10

AI helps the least experienced most

Productivity or quality gain from AI access, by experience or skill level, in two field studies.



-13%

Relative decline in employment for workers aged 22 to 25 in the most AI-exposed occupations since late 2022, while employment for older workers in the same jobs held up. Payroll data covering millions of U.S. workers.¹⁰

The graduate pipeline will narrow, and the role will change: less rote configuration and testing, and structured exposure to client judgment and AI-assisted delivery earlier in a career. Firms that stop hiring juniors altogether will face a senior-talent shortage in ten years. The ones that redesign the apprenticeship will own the next generation of delivery leaders.

Partner models

Partner compensation and governance have followed the same logic as leverage. Partners earned on the margin generated by the staff beneath them, and equity reflected books of billable business. Outcome pricing changes what a book of business means: a portfolio of outcome commitments and shared risk, not a backlog of

billable hours. Partner groups will likely get smaller and more accountable for the risk they underwrite.

Acquisitions

Consulting M&A has traditionally been priced on headcount, utilization and backlog: buy the book of business and the pyramid that delivers it. That calculation changes when the valuable asset is productized IP and a record of pricing and delivering outcomes. Expect acquirers to pay premiums for firms with defensible accelerators and outcome-pricing track records, and to discount pure staffing firms whose main asset is headcount that AI makes less scarce every year.

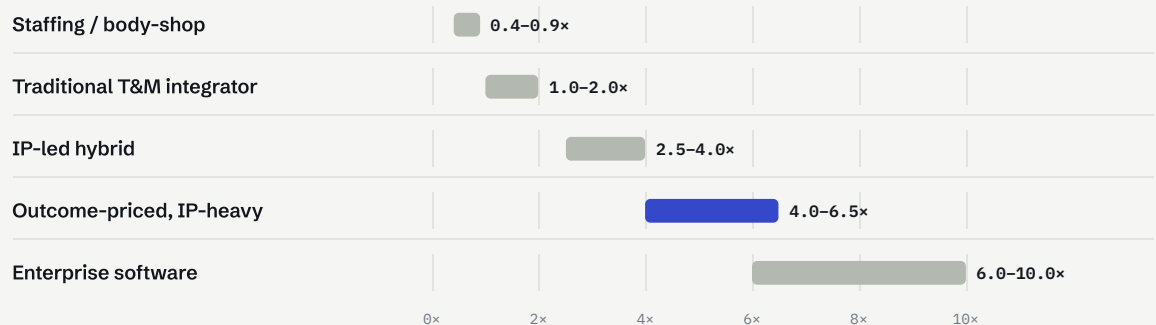
Valuation multiples

This is the clearest through-line of all the second-order effects. Traditional time-and-materials integrators have been valued closer to staffing and services businesses. A firm re-engineered around productized IP and outcome pricing starts to look more like a software or platform business: higher margin, more predictable revenue, less linear to headcount. Multiples should re-rate accordingly. Not all the way to software, but meaningfully closer.

FIGURE 11

Where the multiple sits: enterprise value to revenue

Indicative ranges by business model. The re-rating opportunity is the gap between a traditional integrator and an outcome-priced, IP-heavy firm.



Indicative ranges for discussion, not drawn from transaction data. Multiples vary with growth, margin and market cycle.

12

FORECAST:
WHEN IT
ARRIVES

A timeline for the disruption

Every executive I brief asks the same question: *when?* Most forecasts answer with a vibe. This one rests on two things that can be measured and argued with. The first is how long a task AI can complete on its own, which is rising on a steady, published trend. The second is how long enterprises have historically taken to reorganize work around a new general-purpose technology.

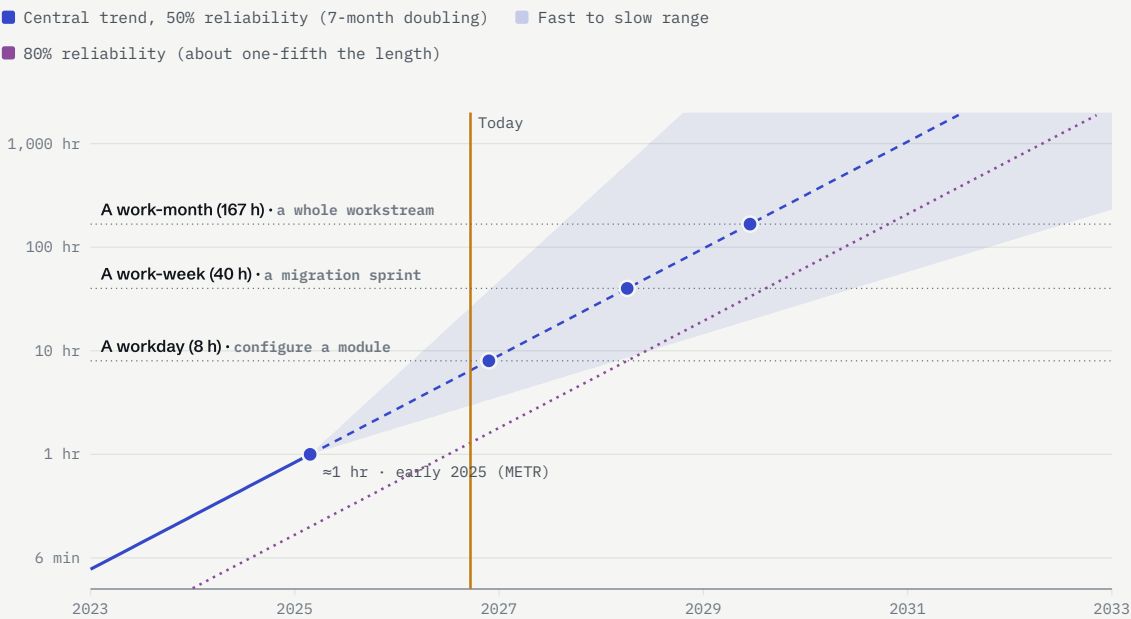
The capability clock

In 2025 the research group METR measured the length of software and reasoning tasks, timed by how long they take skilled humans, that frontier AI models complete with 50% reliability. That "time horizon" had doubled roughly every seven months since 2019, and stood at about one hour in early 2025.²⁵ Consulting work comes in similar units. A test script is an hour. Configuring a module is a day. A migration sprint is a week. A full workstream is a month. Project the trend forward and you get a date by which AI can plausibly handle each unit.

FIGURE 12

When AI can complete a unit of consulting work on its own

Length of task (in human working time) that frontier AI completes with 50% reliability. Log scale: each gridline is ten times the one below. The band spans a 4-month to 12-month doubling time around the 7-month central trend.



A WORKDAY (8 H)

2026

central · range 2026–2028

A WORK-WEEK (40 H)

2028

central · range 2026–2030

A WORK-MONTH (167 H)

2029

central · range 2027–2032

Anchor: Kwa et al. / METR (2025), ≈1 hour 50%-horizon in early 2025, doubling ≈7 months over 2019–2025.²⁵ Crossing dates are extrapolations, not predictions of any lab's roadmap. METR's tasks are cleaner than client work, so real engagements will lag. Update this chart as new measurements are published.

Capability is not disruption. Paul David showed that electric motors took roughly four decades to raise factory productivity, because the gains only came once factories were redesigned around the new power source.²⁶ Brynjolfsson, Rock and Syverson document the same lag for modern general-purpose technologies.²¹ In enterprise services, the lag runs through procurement cycles, multi-year master agreements, regulatory review and partner incentives. My working assumption is a lag of **18 to 36 months** from "AI can do this reliably" to "clients refuse to pay hours for it", with regulated industries at the long end.

The AI maturity map

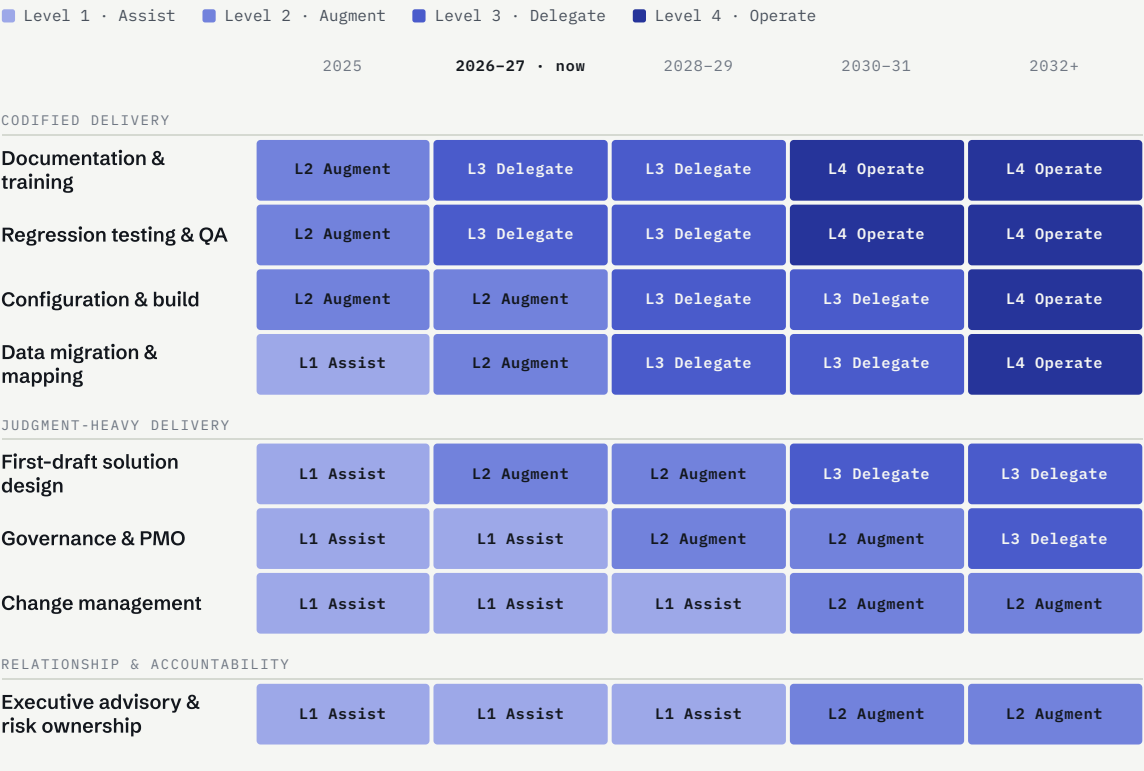
Combining the capability clock with that lag gives a maturity forecast for each workstream. I use four levels, defined by who does the work and who is accountable for it.

■ LEVEL 1 Assist AI drafts and suggests. People do the work and bill for it. Hours fall at the margin.	■ LEVEL 2 Augment AI produces most of the output; people review all of it. Hours fall sharply and the efficiency trap bites.
■ LEVEL 3 Delegate Agents run the workflow end to end. People handle exceptions and sign off. Hours stop being a sensible unit of price.	■ LEVEL 4 Operate Agents run it continuously as a managed service against SLAs. The firm sells the outcome and owns the accountability.

FIGURE 13

AI maturity map: forecast level of delivery by workstream

Typical level in mainstream enterprise engagements, not at the leading edge, which runs about one period ahead. Darker means more of the work is done by AI.



Author's forecast. Derived from the capability clock (Figure 12) plus an 18–36 month adoption lag, and from the compression evidence in Section 03. Relationship and accountability work stays at Level 1–2 by design: clients pay a person to own the risk.

Five waves of disruption

As workstreams climb the maturity levels, the effects arrive in waves. Each wave starts when enough of the engagement reaches a new level for clients and competitors to

notice. The windows overlap, and each has leading signals a firm can watch for.

FIGURE 14

The disruption timeline for the consulting model

The pale band is the plausible window; the solid bar is the most likely period of peak impact.

■ Plausible window ■ Most likely peak | Today (September 2026)

Wave 1 • Task compression

Documentation, testing and code hours fall inside engagements.

Wave 2 • Commercial repricing

Procurement shifts scope from rate cards to fixed-fee and outcome terms.

Wave 3 • Pyramid compression

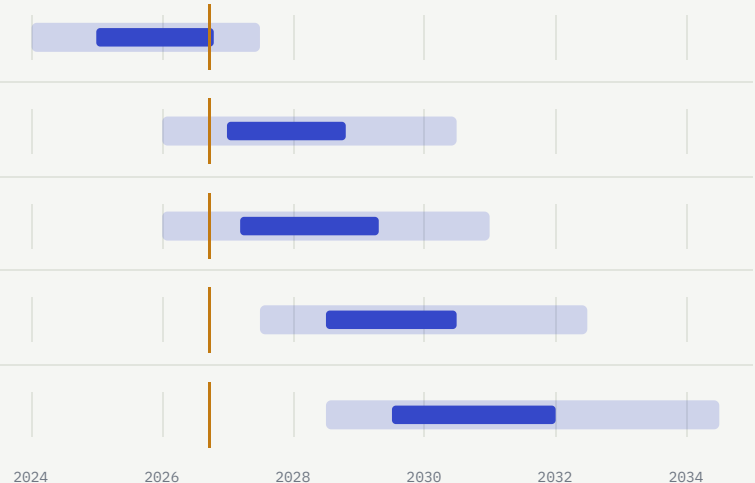
Graduate intake falls; leverage ratios move from about 1:8 toward 1:3.

Wave 4 • Agentic workstreams

Agents deliver whole workstreams; managed-outcome contracts spread.

Wave 5 • Re-rating & consolidation

Valuations split between IP-led and headcount-led firms; M&A follows.



Author's forecast, built on Figures 12 and 13. Windows reflect mainstream enterprise buyers; regulated sectors (banking, insurance, public sector) run 12-24 months later.

2027-29

The decisive window. Commercial repricing and pyramid compression overlap, and AI can plausibly deliver a work-week of effort on its own. Firms that have not instrumented delivery and built outcome offerings by then will be repricing under pressure instead of by design.

Signals to watch

In client RFPs: requests for AI-adjusted estimates, productivity credits, or a fixed-fee option next to the rate card.

In rate cards: discounts that concentrate on analyst and consultant tiers while senior rates hold.

In hiring: smaller graduate classes and fewer entry-level postings in exposed roles, which payroll data already shows.¹⁰

In firm disclosures: the share of revenue from fixed-fee, managed-service and outcome-linked contracts.

In capability research: each new time-horizon measurement. If the doubling time stays near seven months, the dates in Figure 12 hold. If it shortens, move every wave earlier.

The strongest case against this argument

An argument worth making should survive its best counterarguments. Here are the five I hear most often from partners, CFOs and procurement leaders.

OBJECTION 1

"The productivity numbers are hype."

Some of them are. The METR trial and the Danish labor-market study are real warnings, and Daron Acemoglu's estimate of AI's aggregate macroeconomic effect is modest.^{8,9,24} But the argument here doesn't need a 40% gain. Any sustained compression in delivery effort, even 10%, creates the same trap under hourly billing. Uncertainty about the size of the gain is a reason to instrument delivery before underwriting risk. It is not a reason to keep a billing model that loses money on every gain that does materialize.

OBJECTION 2

"Cheaper consulting will create more demand. Volume will save us."

Possibly. Jevons observed in 1865 that more efficient coal engines increased total coal consumption.²³ AI may well expand the market for transformation work. But more demand at a lower price per outcome does not rescue revenue per hour, and the firms that capture new volume will be those whose capacity scales without adding headcount. A Jevons effect helps outcome-priced firms more than hourly ones.

OBJECTION 3

"Outcomes can't be measured or attributed."

For some work that is true, and the readiness map in Figure 6 says so. Contract theory warns that rewarding a single measured outcome can distort effort on everything else.¹³ The answer is to match the model to the measurability: fixed fees where deliverables are clear, milestones where "done" is testable, gain-share only where a business metric has a clean baseline, and capped time and materials where none of those hold yet.

OBJECTION 4

"Fixed fees just move all the risk onto the firm."

Yes, and that is the product. Clients pay a premium for certainty; the flat-rate bias literature shows it.¹⁷ The question is whether the firm's delivery variance lets it earn that premium without being ruined by the tail. That is why Figure 5 matters: risk transfer is profitable for a firm with a narrow distribution and dangerous for one without.

OBJECTION 5

"Clients will just negotiate hourly rates down. Time and materials survives at a lower price."

For commodity work, that is exactly what will happen. A time-and-materials firm then becomes a price-taker in a shrinking market, competing on rate against offshore capacity and AI tooling that clients can run themselves. That is the staffing-firm end of Figure 11. Survival at a lower rate is the outcome this paper is trying to help firms avoid.

AI doesn't just change how consulting firms deliver work. It changes what clients should pay for, and how. The firms that move first from billing time to pricing outcomes will capture the value AI creates. The firms that don't will spend the next several years discovering that their own efficiency gains were given away for free.

ABOUT THE AUTHOR

Bryan Musto

APPLIED AI IN DELIVERY

OUTCOME-BASED PRICING

REVENUE STRATEGY

Bryan Musto is a Chief Revenue Officer and go-to-market leader working where applied AI meets enterprise delivery. His focus is what AI-assisted and agentic delivery does to the commercial models built on top of it. He is CRO of Northlight Solutions Group, where he leads productized, AI-augmented delivery and the outcome-based pricing structures it makes possible.

He writes and advises on AI-driven delivery economics, pricing-model design and the compression of the consulting pyramid. His background includes senior roles at Capgemini Invent, Vanguard and BNP Paribas, spanning both sides of the buy-sell relationship for large enterprise technology programs. He holds an MBA from the Wharton School of the University of Pennsylvania.

[linkedin.com/in/bryanmusto](https://www.linkedin.com/in/bryanmusto)

REFERENCES

- 1 Maister, D. H. (1982). Balancing the professional service firm. *Sloan Management Review*, 24(1), 15–29.
- 2 Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review*, 57(3), 415–426.
- 3 Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper* 24-013.
- 4 Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.
- 5 Peng, S., Kalliamvakou, E., Cihon, P., & Demirel, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv:2302.06590*.
- 6 Cui, Z. K., Demirel, M., Jaffe, S., Musolf, L., Peng, S., & Salz, T. (2025). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Working paper*, SSRN 4945566.
- 7 Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.
- 8 Becker, J., Rush, N., Barnes, E., & Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. *METR*, arXiv:2507.09089.
- 9 Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects. *NBER Working Paper* 33777.
- 10 Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. *Stanford Digital Economy Lab*.
- 11 Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702), 1306–1308.
- 12 Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, 10(1), 74–91.
- 13 Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7, 24–52.
- 14 Christensen, C. M., Wang, D., & van Bever, D. (2013). Consulting on the cusp of disruption. *Harvard Business Review*, 91(10), 106–114.
- 15 Flyvbjerg, B., & Budzier, A. (2011). Why your IT project may be riskier than you think. *Harvard Business Review*, 89(9), 23–25.
- 16 Capen, E. C., Clapp, R. V., & Campbell, W. M. (1971). Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, 23(6), 641–653.
- 17 Lambrecht, A., & Skiera, B. (2006). Paying too much and being happy about it: Existence, causes, and consequences of tariff-choice biases. *Journal of Marketing Research*, 43(2), 212–223.
- 18 Prelec, D., & Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science*, 17(1), 4–28.
- 19 Kim, S.-H., Cohen, M. A., & Netessine, S. (2007). Performance contracting in after-sales service supply chains. *Management Science*, 53(12), 1843–1858.
- 20 Ng, I. C. L., Maull, R., & Yip, N. (2009). Outcome-based contracts as a driver for systems thinking and service-dominant logic in service science: Evidence from the defence industry. *European Management Journal*, 27(6), 377–387.
- 21 Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372.
- 22 Porter, M. E., & Teisberg, E. O. (2006). *Redefining Health Care: Creating Value-Based Competition on Results*. Harvard Business School Press.
- 23 Jevons, W. S. (1865). *The Coal Question*. Macmillan.
- 24 Acemoglu, D. (2024). The simple macroeconomics of AI. *NBER Working Paper* 32487.
- 25 Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., et al. (2025). Measuring AI ability to complete long tasks. *METR*, arXiv:2503.14499.
- 26 David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review*, 80(2), 355–361.

A note on method. Figures 1 and 10 and the statistics in Sections 02, 06 and 11 report published results. Figures 2, 4, 5, 6, 9 and 11 are models or practitioner estimates, and Figures 12 to 14 are forecasts, all labeled as such, meant to make the argument's mechanics visible rather than to forecast specific numbers.