

AI BENEFIT REALISATION · PARTS 1 TO 4

Measuring and funding AI with the finance function.

A baseline, an attribution method, a cost per outcome and a gate that can stop funding. Four controls the CFO and the Chief AI Officer sign together.

Terence Kok and Dr. Christina Chua

For CFOs, Chief AI Officers and investment committees

ABOUT THE AUTHORS

We wrote this series together.



Terence Kok

Executive Director, AI Governance & Assurance Practice, Orion Five Engineering

ISO/IEC 42001 Lead Auditor. Programme governance from infrastructure, where funding is released in stages against evidence.

BRINGS

The governance view: gates, decision rights and the record a board, auditor or regulator will ask for.



Dr. Christina Chua

Director, Institutional Sales, Moomoo Financial Singapore

More than ten years in financial services, taking new products to clients through product, compliance, risk and technology teams.

BRINGS

The finance view: what a CFO can sign, book and review, from the baseline to the monthly cost bridge.

The series does not discuss Moomoo, and the views in it are not Moomoo's.

THIS SESSION

Five parts to this session.

01

The joint framework

Why AI benefits go unproven, and four controls the two offices sign together.

PART 1

02

The baseline

What to measure, for how long, and who signs before the first model call.

PART 2

03

Attribution

Separating the AI's effect from everything else that changed.

PART 3

04

The cost per outcome

Tagging every model call, and the monthly AI cost review.

PART 4

05

What to do this quarter

Actions, committee questions and what the series covers next.

THE ARGUMENT IN ONE SLIDE

Most AI programmes cannot say what they returned.

25%

of AI initiatives delivered the return chief executives expected, and 16% scaled enterprise-wide

30%

of generative AI projects Gartner expected to be abandoned after proof of concept by the end of 2025

98%

of FinOps practitioners now manage AI spend, up from 31% in 2024

5–30×

more tokens per task for agentic models than for a standard chatbot

Sources: IBM CEO Study, May 2025; Gartner, July 2024 and March 2026 (forecasts); State of FinOps 2026.

01

The joint framework.

The CAIO runs the work and the CFO owns the numbers. Neither owns the chain from starting point to realised benefit.

PART 1 · 10 OCTOBER 2026



WHY AI BENEFITS GO UNPROVEN

The evidence of underperformance is consistent.

25%

of AI initiatives delivered expected ROI over the last few years

IBM CEO Study, 2025, 2,000 CEOs

30%+

of generative AI projects abandoned after proof of concept by end-2025

Gartner forecast, July 2024

40%+

of agentic AI projects cancelled by the end of 2027

Gartner forecast, June 2025

63%

of finance organisations found AI implementation slower than expected in 2025

Gartner, May 2026

"Counts of pilots, tools rolled out or use cases in production show that finance is moving, but they do not prove that AI is delivering the value." Gartner finance research, May 2026.

Gartner's reasons in 2024: poor data quality, inadequate risk controls, escalating costs or unclear business value. Grades: reported (surveys) and forecast.

BELIEF AND MEASUREMENT

The best-known figures record what executives believe.

MIT NANDA · July 2025

"95% of organizations are getting zero return"

Based on 52 interviews, a survey of 153 leaders and a review of public announcements.

THE AUTHORS' OWN CAVEAT

"Directionally accurate based on individual interviews rather than official company reporting."

Wharton and GBK · October 2025

Three in four leaders report positive returns

An executive survey, later the same year.

WHAT BOTH HAVE IN COMMON

Neither is a measurement. The gap between belief and measurement is the problem this series addresses.

Grades: MIT NANDA reported, contested; Wharton reported, executive survey.

THE OWNERSHIP GAP

A benefit claim travels between two offices.

CHIEF AI OFFICER

Owns the use case, model, data and delivery team

76% of organisations now have one, up from 26% a year earlier

"The model cut handling time by a fifth."

The CFO cannot find the saving in the cost base.

CHIEF FINANCIAL OFFICER

Owns the budget, the benefits register and the board numbers

62% of CFOs say their role expanded into technology or AI strategy

NEITHER OFFICE IS WRONG. EACH IS A GAP BETWEEN THEM.

Time was released

and nobody decided what to do with it

The baseline was taken

in a quiet month

The inference bill grew

faster than the saving

Sources: IBM CEO Study, May 2026; IBM CFO Study, September 2026 (reported, executive surveys). The three gaps are our reading of deployments we have seen.

FOUR CONTROLS, SIGNED JOINTLY

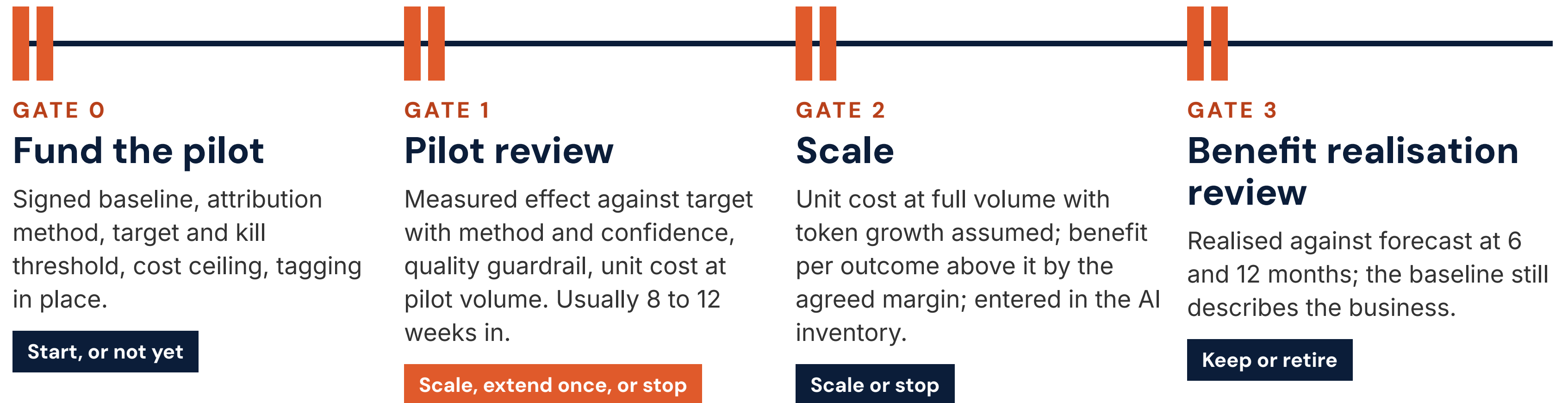
One office prepares each control, and the other signs it.

CONTROL	CAIO PREPARES	CFO SIGNS	WHEN
1 Baseline	Metric definitions, data source, measurement period, natural variance	The baseline record and the target, as a finance figure	Before the first production model call
2 Attribution	The method, the control group or comparison, the sample size	The grade of evidence each claimed benefit may carry	At pilot approval, before results exist
3 Unit cost	Tagged inference, retrieval and evaluation spend per use case	The fully loaded cost per outcome, including human review	Monthly, from the first day of the pilot
4 Stage-gates	The results against target, with method and confidence	The decision to scale, extend once, or stop	At each gate, on a fixed date

Our proposed split of responsibilities. The point is that neither office can pass a gate alone.

STAGE-GATES THAT STOP FUNDING

Four gates, each with criteria written before the work.



| A second extension at Gate 1 should need the investment committee.

Our proposal, adapted from Cooper's stage-gate systems (1990) and the UK government's Gateway reviews, where Gate 5 is dedicated to benefit realisation.

KILL CRITERIA

Four kill criteria for every pilot charter.

- 1 The measured effect, at the agreed confidence, is below the kill threshold.
- 2 The unit cost at projected volume exceeds the unit benefit.
- 3 The quality guardrail is breached and cannot be fixed within the extension.
- 4 The agreed attribution method could not be run, so the benefit cannot be measured.

Teams resist the fourth most, because the pilot may well be working. A benefit that cannot be measured cannot be defended to a board, an auditor or a regulator.

MAKING STOPPING NORMAL

A pilot stopped with a clean measurement has done its job.

Report the stop rate

A measure of the gates' health

A portfolio in which nothing has ever been stopped does not mean every pilot succeeded. It means the gates are not being used.

Gartner, 2026

"Cutting weak ideas early"

Gartner advised CFOs to stop looking for a single ROI formula and build a balanced portfolio. Mature finance uses such as payables automation generally return their investment within nine to ten months.

Sources: Gartner press releases, March and September 2026 (reported, analyst survey). The stop-rate measure is our proposal.

THE SINGAPORE CONTEXT

Singapore funds AI adoption and asks for outcomes.



A joint CFO and CAIO gate, with written criteria and a record of every decision, is evidence of the oversight MAS expects.

Sources: MDDI, February 2024; Budget 2026 statement; MAS media release, October 2026. Hollow points are dates still to come, or our finding.

02

The baseline.

A baseline taken after the pilot starts is a reconstruction, and reconstructions flatter the pilot.

PART 2 · 17 OCTOBER 2026

AN OLD HABIT, RARELY APPLIED TO AI

Few organisations measure AI against a starting point.

<1 in 5

organisations tracked KPIs for their gen AI solutions. KPI tracking had the most bottom-line impact of the twelve practices tested.

McKinsey State of AI, March 2025

37%

of organisations attribute any EBIT impact to AI, essentially unchanged from the year before

McKinsey State of AI, August 2026

Initiation

"should be taken during project initiation so that performance is not impacted by any element of project implementation"

UK IPA benefits guide, on the baseline

Grades: McKinsey reported, executive surveys; UK Infrastructure and Projects Authority guide, 2017, government guidance.

WHY A LATE BASELINE FLATTERS THE PILOT

Four reasons, and none needs bad faith.

01

People misremember the starting point

Capable people who had just done the work misjudged the AI's effect by a wide margin.

02

Pilots go where performance is worst

Regression to the mean: an unusually bad month is followed by a better one, with or without AI.

03

New tools look better at first

Novelty effects: early results differ from the settled effect because users react to the change itself.

04

Measures shift once the project is known

Teams tidy data, change a process or redefine a metric "in anticipation of the intervention".

Sources: METR, 2025; Barnett et al., 2005; Kohavi, Tang and Xu, 2020; HM Treasury guidance on evaluating AI interventions, 2024. The grouping is ours.

METR, JULY 2025 · RANDOMISED, 16 DEVELOPERS

They were slower, and believed they were faster.

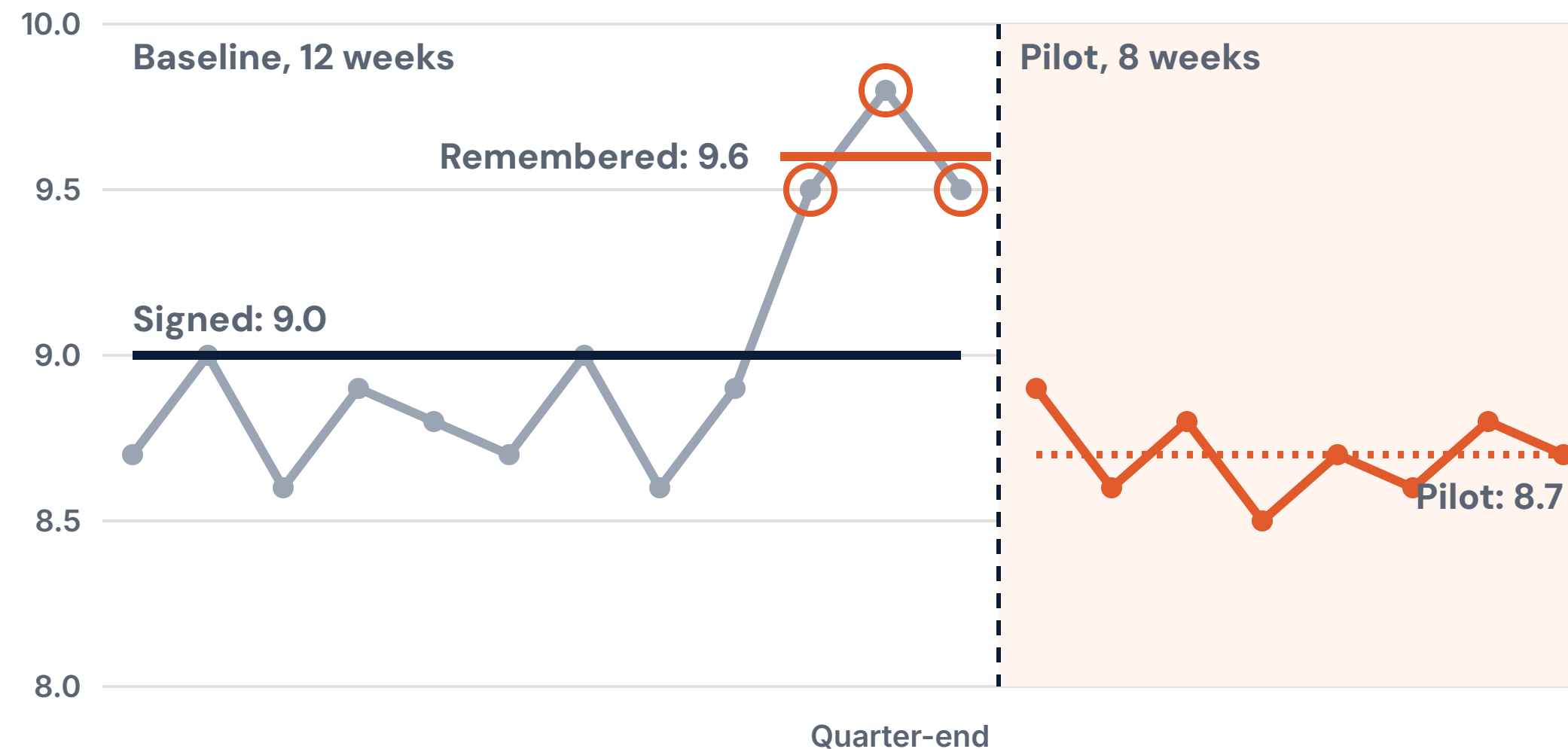


METR says the results no longer reflect current tools. The perception gap is the point: a baseline gathered from recollection inherits the same error.

Source: METR, July 2025, and its February 2026 note. Grade: measured, randomised experiment, early-2025 tools.

THE SAME EIGHT PILOT WEEKS, TWO BASELINES · ILLUSTRATIVE

The reconstructed baseline nearly triples the gain.



AGAINST THE SIGNED BASELINE

3%

better. Twelve weeks including a quarter-end. **The pilot goes back for a longer run.**

AGAINST A RECONSTRUCTION

9%

better. Three busy weeks everyone remembers. **The pilot gets scaled.**

Average handling time, minutes. Illustrative example built for the post: the signed baseline covers twelve weeks including a quarter-end; the reconstruction uses the three busy weeks everyone remembers.

CHOOSING WHAT TO MEASURE

Pair every speed measure with a quality measure.

Use case	Primary outcome	Quality guardrail	System of record
Contact-centre assistant	Calls resolved per agent hour	Repeat contact within 7 days; complaint rate	Telephony and case-management timestamps
Invoice processing	Cost per invoice processed	Payment errors; duplicate payments caught late	Accounts payable ledger and workflow logs
Claims assessment	Claims closed per assessor per week	Reopened claims; leakage found at audit	Claims system and audit sample
Credit memo drafting	Days from application to credit decision	Memos returned by the credit committee	Loan origination system
Month-end close	Working days to close	Post-close adjustments	General ledger and close calendar

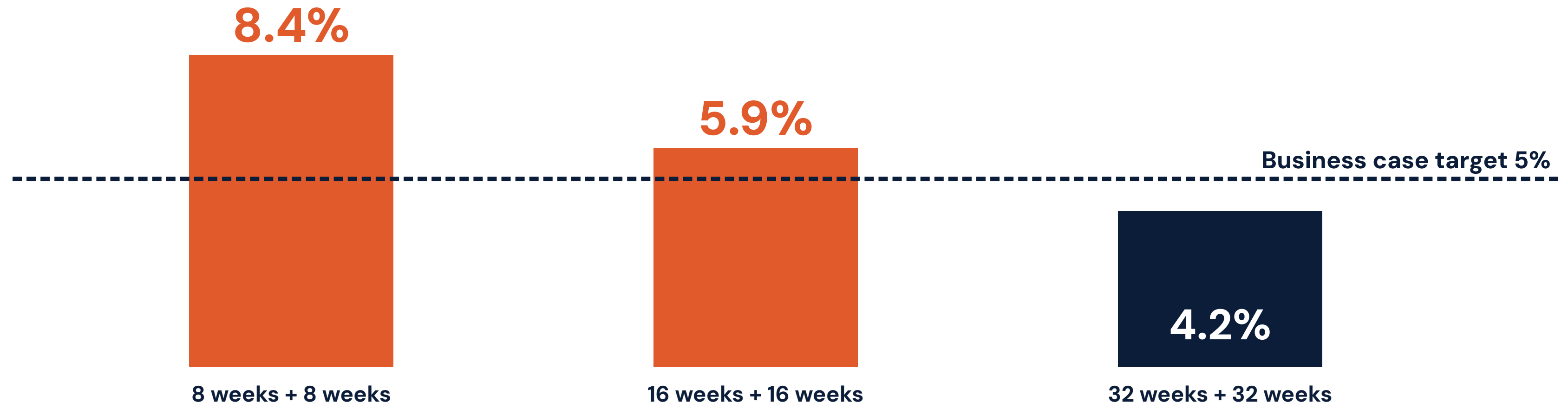
Our examples. In 5,172 support agents, productivity rose 15% on average and 30% for novices, while the most experienced saw "small gains in speed and small declines in quality" (Brynjolfsson, Li and Raymond, QJE 2025).

NATURAL VARIANCE DECIDES THE PILOT LENGTH

A 5% target can hide inside the noise.

Smallest detectable change $\approx 2.8 \times \text{weekly spread} \times \sqrt{(1/\text{baseline weeks} + 1/\text{pilot weeks})}$

80% power at 5% significance · weekly spread 6% of the average



The rule and the 8.4% example are from the post (standard two-sample power calculation, simplified). The 16- and 32-week bars are our arithmetic with the same rule. Correlated or trending weeks make the real figure larger.

WHEN THE TARGET SITS INSIDE THE NOISE

Four responses, decided before approval.

01

Run longer

More weeks narrow the band, but slowly. Doubling both periods cuts the detectable change by only about 30%.

02

Add a control group

A staggered rollout or holdout measures the same weeks, so shared movement cancels out. Part 3.

03

Use the baseline in the analysis

CUPED cut variance "by about 50%" at Bing. It needs a baseline at team or case level, recorded before the pilot.

04

Accept it cannot be measured

A legitimate finding, and the fourth kill criterion. Cheaper at approval than at Gate 1.

Source for CUPED: Deng, Xu, Kohavi and Walker, WSDM 2013 (measured, company experiment platform).

THE ONE-PAGE BASELINE RECORD · CONTACT-CENTRE EXAMPLE

The baseline fits on one page, signed by finance.

Use case identifier	CC-ASSIST-01, the same ID used to tag model calls and costs	Primary outcome	Calls resolved per agent hour, from the case system
Data source and owner	Telephony platform and CRM; Head of Contact Centre Operations	Measurement period	13 weeks to 30 September, including quarter-end
Average and weekly spread	4.1 calls per hour; weekly spread 6%	Trend and seasonality	Flat trend; quarter-end weeks 8 to 10% lower
Smallest detectable change	7.5% before-and-after over an eight-week pilot	Quality guardrail	Repeat contacts within 7 days no more than 12%
Expected disbenefits	Escalations to supervisors; supervisor review hours	Current cost to serve	S\$9.40 per resolved call
Target and kill threshold	Target 12%; kill below 5%	Plan for released time	Absorb volume growth for two quarters, then no backfill of leavers
Signatures and freeze date	Signed and frozen before the first production model call	Change log	Empty at signature; every later change logged with a reason

Our proposed template. The example figures are illustrative.

WHO SIGNS AND WHEN

Six steps to the first production model call.



MAS's final guidelines expect AI performance thresholds to be "clearly defined, documented, and mutually agreed upon by business owners, developers and reviewers". The signed baseline applies the same discipline to the benefit side.

Sequence: our proposal. MAS Guidelines on AI Risk Management, 7 October 2026, para 5.14(a). No Singapore guidance requires a benefits baseline for AI today.

WHEN THERE IS NO CLEAN HISTORY

Three better answers than a reconstruction.

01

A shadow measurement period

Instrument the process and measure it for four to eight weeks before the AI goes live.

It delays the pilot. It is the only way to get a baseline on the same definition as the results.

02

Sampled measurement

Measure a random sample of cases by hand, such as the time taken on a hundred randomly chosen claims.

A small random sample is more reliable than a large estimate.

03

A control group from day one

If history cannot be measured, measure a comparison group during the pilot.

The concurrent group becomes the reference point.

Never accepted: the pilot team's estimates, vendor benchmarks or another organisation's published results.

TIME RELEASED NEEDS A PLAN

Most pilots release time before they release cash.

~3%

of work hours saved by Danish workers using AI chatbots, self-reported. The study found "precise null effects on earnings and recorded hours", ruling out effects above 2%.

Humlum and Vestergaard, NBER WP 33777, revised March 2026

STATE THE TYPE OF BENEFIT IN THE BASELINE

Capacity	to absorb growth
-----------------	------------------

Cost avoided	hiring that no longer happens
---------------------	-------------------------------

Cost removed	spend that leaves the budget
---------------------	------------------------------

Revenue	income the AI adds
----------------	--------------------

Time released is reported as capacity until cost leaves the budget.

Grade: measured, administrative data and survey. The four benefit types and the reporting rule are our proposal.

03

Attribution.

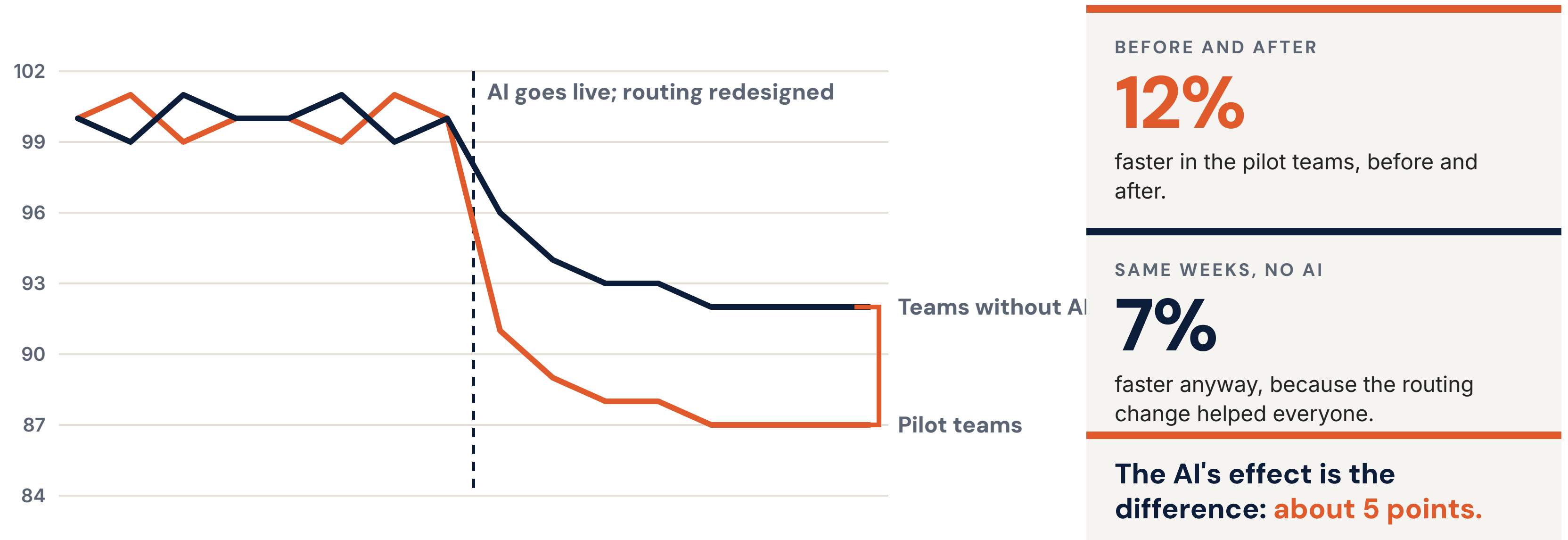
Revenue rose and the AI went live in the same quarter.
The only reliable answer is a comparison group
measured over the same weeks.

PART 3 · 24 OCTOBER 2026



ONE PILOT, READ TWO WAYS · ILLUSTRATIVE

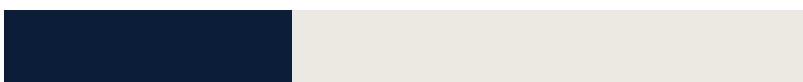
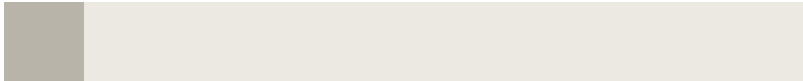
A 12% gain becomes 5 points next to a comparison group.



Handling time indexed to 100 before go-live. Illustrative example built for the post: call routing was redesigned in the same quarter and helped every team.

FIVE METHODS, STRONGEST FIRST

The methods give very different confidence.

Randomised holdout	Some teams or cases are randomly kept without the AI. The difference is the effect.		Highest
Staggered rollout	Waves; each compared with the teams still waiting.		High
Matched comparison	Similar teams without the AI, chosen before the pilot.		Moderate
Trend-adjusted before and after	Against the baseline, corrected for trend and season.		Low
Self-reported estimate	Useful for deciding what to measure next.		Not counted

Our ranking. Every claimed benefit carries its method and grade, and benefits of different grades are reported separately, never summed.

RANDOMISED HOLDOUT

Three design decisions go in the pilot charter.

01

The unit of randomisation

Where the AI is used and its effects stay contained: the individual for a drafting assistant, the team for a tool that changes how work is shared.

Randomising whole teams or sites prevents most leakage, such as held-out agents copying drafts from colleagues.

02

The size of the holdout

Large enough to detect the target. 10 to 20% of units is common in practice.

The right figure comes from the detectable-change calculation in Part 2.

03

The cost

The held-out group goes without the benefit for the pilot. That forgone benefit is the price of the evidence.

Usually small next to scaling a tool on a number that turns out to be wrong.

Leakage and interference: Kohavi, Tang and Xu, 2020. The holdout share is our practice.

CUI ET AL., MANAGEMENT SCIENCE · THREE RANDOMISED TRIALS

Even large, well-run trials give a wide range.

4,867

software developers

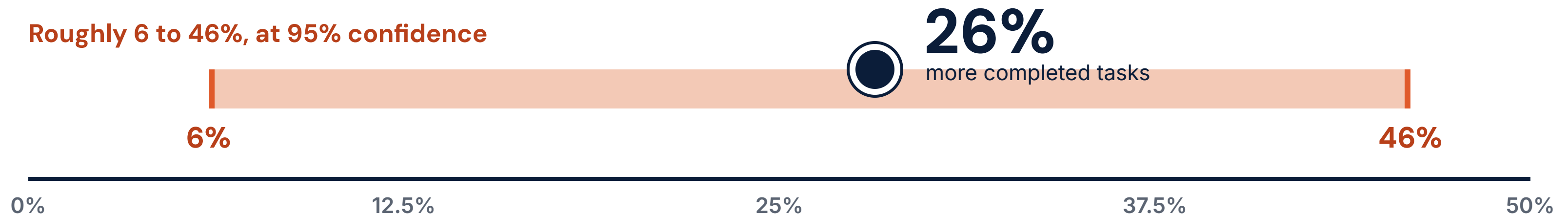
3

experiments, at Microsoft, Accenture and a Fortune 100 firm

Noisy

each experiment, on its own

Roughly 6 to 46%, at 95% confidence



A single-team pilot without a comparison group could not produce a reliable estimate at all.

Estimate 26.08%, standard error 10.3 points (measured, randomised experiments). The interval is our arithmetic from the published estimate and standard error.

Randomisation shows the harms a before-and-after hides.

Inside the AI frontier

12.2% **25.1%**

more tasks completed faster

With significantly better quality.

Outside the AI frontier

19%

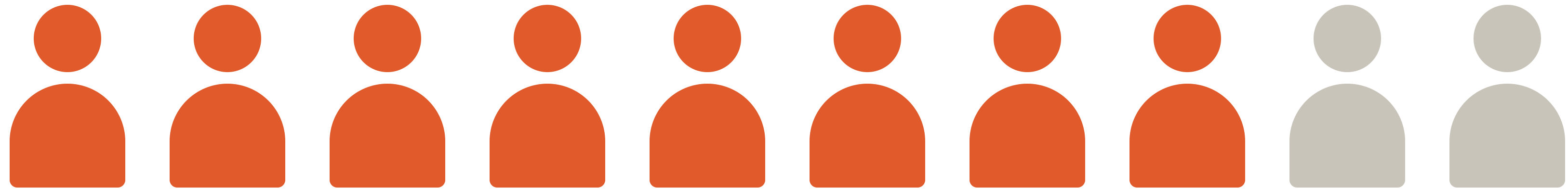
less likely to reach a correct answer

Without the random assignment, this finding would have been invisible. A pilot that measured only the tasks the AI handled well would have reported a clean gain.

Grade: measured, randomised experiment.

DILLON ET AL., 2025 · 7,137 WORKERS IN 66 FIRMS, RANDOMISED

Book the effect per licence paid for.



● 80% used the tool: about two fewer hours a week on email ● 20% did not use it, and the licence was still paid

$$\begin{array}{|c|} \hline \mathbf{2 \text{ hours}} \\ \hline \text{per user} \\ \hline \end{array} \times \begin{array}{|c|} \hline \mathbf{80\%} \\ \hline \text{who used it} \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{1.6 \text{ hours}} \\ \hline \text{per licence, roughly} \\ \hline \end{array}$$

Statisticians call the per-licence figure the intention-to-treat effect. It is the one a business case needs, because the organisation pays per licence. The vendor's case study will usually quote the larger number.

Source: Dillon, Jaffe, Immorlica and Stanton, arXiv 2504.11436 v4 (measured, randomised field experiment). 1.6 hours is our arithmetic, assuming non-users save nothing.

A COMMON SHORTCUT

Comparing heavy and light users does not measure the tool.

WHAT THE COMPARISON MIXES

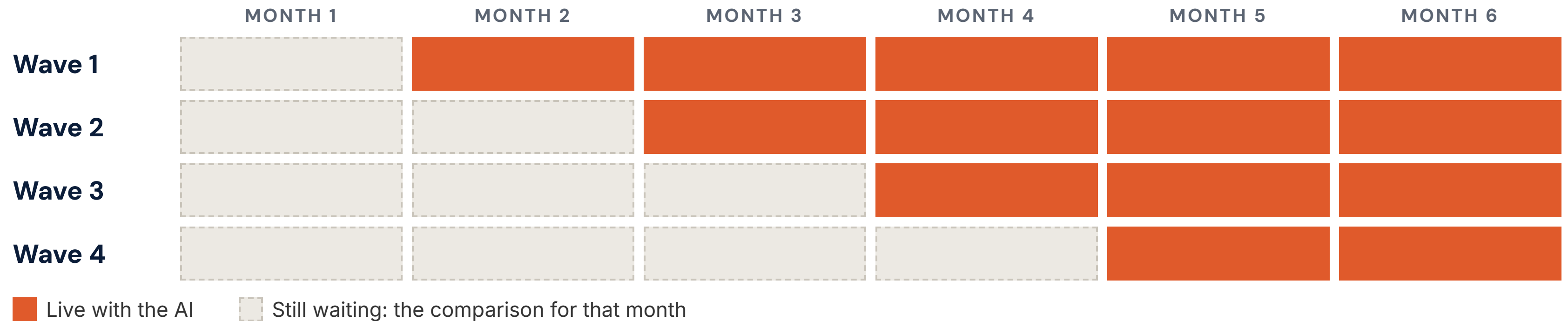
Heavy users chose to use the tool, and they are often the people who were already faster, more curious or less busy. The difference mixes the AI's effect with the difference between the people.

WHAT IT IS GOOD FOR

Deciding where to train. It should never be booked as a benefit.

STAGGERED ROLLOUT · SCHEMATIC

Each wave is compared with the teams still waiting.

**Draw the wave order by lot**

It becomes a stepped-wedge trial, with nearly the confidence of a holdout.

Compare only with teams not yet started

Early teams used as the comparison bias the standard regression when effects grow over time (Goodman-Bacon, 2021).

Brynjolfsson, Li and Raymond used a staggered rollout for 5,172 agents: 15% average gain, 30% for novices. Estimators: Callaway and Sant'Anna, 2021. Wave order rule: HM Treasury, Magenta Book.

WHEN THE AI CANNOT BE SPLIT BY USER

Randomise by time block, or build a synthetic comparison.

Switchback test

For routing, pricing and dispatch

Run the AI in some time blocks and the old rules in others, with the schedule drawn at random, across several sites.

THE CATCH

The number of independent blocks decides the precision, so switchbacks need weeks of blocks. DoorDash described the method for its dispatch algorithm in 2018.

Synthetic control

For a change in one unit

A weighted mix of untreated units whose history matches the treated unit before the change, such as the other countries of a regional insurer.

THE RULE

Fix the weights before the results are seen. A synthetic control chosen afterwards can be made to show almost anything.

Sources: Kastelman and Ramesh, DoorDash Engineering, 2018 (reported); Abadie, Diamond and Hainmueller, JASA 2010 (measured, methods literature).

THE WEAKEST COUNTED METHOD, AND OVERLAPPING CHANGES

When nothing else is possible, say so in the grade.

Trend-adjusted before and after

Interrupted time series

Fit the weekly history, project it forward, and measure the departure. It needs a long regular history, an expected shape set in advance, and seasonal adjustment.

WHERE IT FAILS

The reorganisation, the price change and the AI launch are announced together. A control series helps considerably.

Several changes at once

Do not split the total by negotiation

Every sponsor would then claim the full effect elsewhere.

THREE BETTER OPTIONS

Sequence the changes; use a factorial design; or report one combined benefit with one owner.

CHOOSING A METHOD FOR EACH USE CASE

The method follows from how the AI's effects spread.

Use case	Unit of comparison	Recommended method	Fallback
Drafting assistant for relationship managers	Individual or team	Randomised holdout	Staggered rollout, randomised wave order
Contact-centre agent assistant	Team or site	Staggered rollout, randomised wave order	Matched comparison of similar sites
Call routing or pricing model	Site and time block	Switchback test	Trend-adjusted, with a control series
Claims triage in one country of a regional insurer	Country	Synthetic control from the other countries	Trend-adjusted before and after
Organisation-wide finance close assistant	The whole organisation	Trend-adjusted before and after	Report as capacity, graded low

Our examples. The recommended method is the strongest the use case usually allows; the fallback applies when the first is impractical.

WHAT EACH METHOD COSTS

Finance should see the method's cost at approval.

Method	Main cost	Grade
Randomised holdout	Benefit forgone by the holdout during the pilot	Highest
Staggered rollout, randomised order	Later waves wait; analyst time for a modern estimator	High
Switchback test	Weeks of blocks to reach enough independent units	High
Synthetic control	Data preparation; weights fixed in advance	Moderate
Matched comparison	Risk that the groups differ in ways nobody measured	Moderate
Trend-adjusted before and after	Confounding by anything launched at the same time	Low

A sponsor who asks to waive a holdout states the alternative and the grade it supports, in the charter. A waiver does not remove the evidence requirement.

04

The cost per outcome.

The price per token falls every quarter while the AI bill rises. Account for it like any other variable cost.

PART 4 · 29 OCTOBER 2026

WHY THE BILL RISES WHILE PRICES FALL

Prices fall fast, and consumption grows faster.

9–900×

yearly fall in the price of a fixed level of performance, depending on the task

Epoch AI, March 2025

~40×

a year, to match GPT-4 on PhD-level science questions

Epoch AI, March 2025

5–30×

more tokens per task for agentic models than a standard chatbot

Gartner forecast, March 2026

>5×

rise in inference cost per agentic workflow through 2028, the "inference paradox"

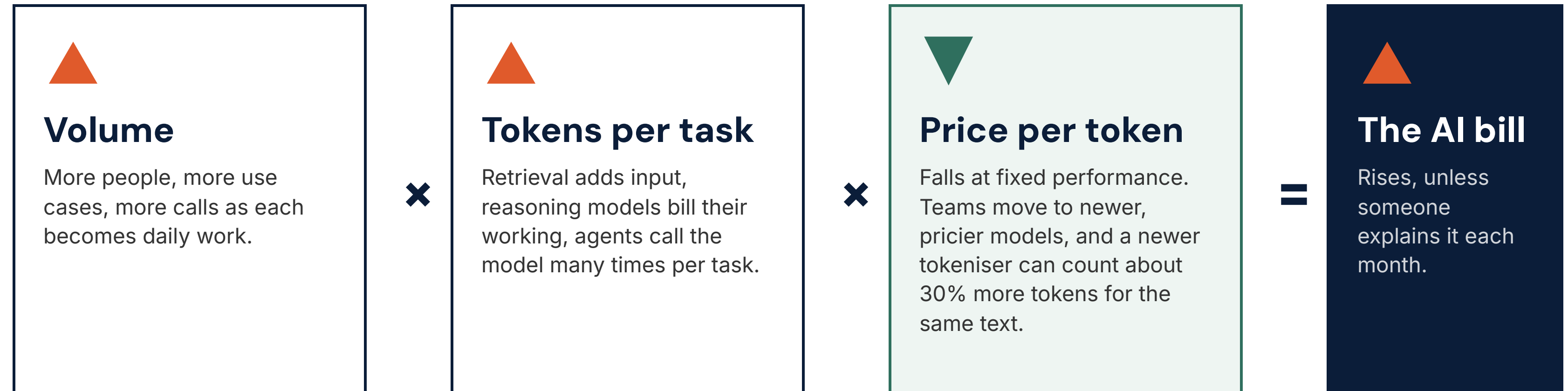
Gartner forecast, 2026

Nobody would accept "the price per kilowatt-hour fell" as an explanation of a rising electricity bill.

Grades: Epoch measured, benchmark price analysis; Gartner forecasts, read through syndicated copies of the release text.

THE BILL IS A PRODUCT OF THREE NUMBERS

Only one of the three is falling.



Tokeniser: Anthropic pricing documentation, October 2026 (newer tokeniser, Claude 4.7 onwards). Agent token use: Gartner, March 2026 (forecast).

ONE ID ON BOTH SIDES OF THE LEDGER

The benefit and the cost carry the same key.

Benefits register		AI gateway cost log	
Use case	CC-ASSIST-01	Use case	CC-ASSIST-01
Owner	Head of Contact Centre Ops	Calls this month	Tagged at the call
Benefit per reply sent	\$S\$1.68	Full running cost	\$S\$19,000
Method and grade	Staggered rollout · High	Cost per reply sent	\$S\$0.48

UNIT MARGIN, A SIMPLE JOIN
\$S\$1.20 per reply

Figures from the post's illustrative worked example. Shared costs are allocated by a driver agreed before the quarter starts.

FOUR TAGGING RULES

Tagging is a small engineering job with four rules.

01**Tag at the call**

One internal gateway rejects calls without a use-case ID and records tokens, model and cost. A key per use case gives a second check.

02**Allocate shared costs by a declared driver**

Requests or tokens, chosen before the quarter, and not changed when a use case's numbers look poor.

03**Keep build and run apart**

Development, testing and evaluation runs are build cost, reported apart from the live service.

04**Show unused commitments**

The unused part of committed spend is its own platform line, so a purchasing decision does not hide in unit costs.

FinOps Foundation's FOCUS specification (1.3, December 2025) gives billing data from several providers one format, and the Foundation has a workstream on AI token charges.

THE COST-PER-OUTCOME NUMERATOR · ILLUSTRATIVE

Model tokens are 32% of the running cost.

Model tokens Including failed calls and retries		6,100
Retrieval and compute Vector database, search, orchestration		1,800
Monitoring and guardrails Logging, filters, tracing, alerts		900
Evaluation runs Scheduled quality tests		700
Oversight and upkeep Reviewers sampling drafts; prompt and knowledge-base staff		7,400
Shared platform Gateway and platform team, by requests		2,100
Total per month, \$\$		19,000

Illustrative worked example: a drafting assistant for customer-email replies. Count reviewer time here only if it sits outside the process whose benefit was measured.

THE DENOMINATOR · ILLUSTRATIVE

Divide by the outcome the benefit was measured in.

S\$19,000

57,000 drafts

S\$0.33

Flatters the use case: a draft nobody sends
still costs money.

S\$19,000

40,000 replies sent

S\$0.48

The unit the benefit was measured in.

UNIT MARGIN PER REPLY

Benefit: 2.4 min × S\$0.70 **S\$1.68**

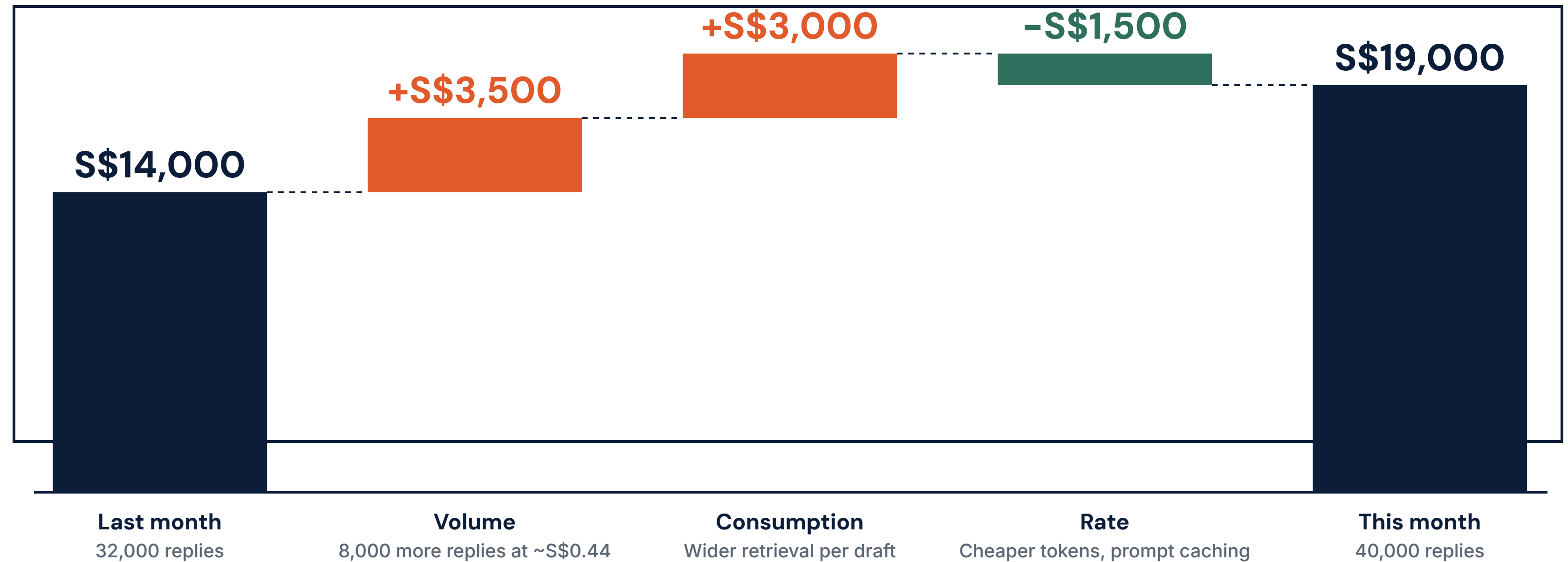
Cost per reply sent **-S\$0.48**

Earned per reply **S\$1.20**

Illustrative. A use case whose cost per outcome is above its benefit per outcome loses money on every transaction, however impressive the total benefit looks.

EXPLAINING THE CHANGE EACH MONTH · ILLUSTRATIVE

A bridge splits the change into volume, consumption and rate.



Replies rose from 32,000 to 40,000; retrieval was widened; prompt caching was turned on. Cost per reply rose from about \$0.44 to \$0.48.

THE SHARE NOBODY HAS TAGGED, AND THE LEVERS

Untagged spend should fall every month.

WHERE UNTAGGED SPEND COMES FROM

- Keys that predate the gateway
- Personal accounts expensed to a cost centre
- AI features bundled into subscriptions
- Pilots that skipped the register

Aim: below a tenth within two quarters. Each line is tagged, registered with an owner, or switched off.

LEVERS THAT LOWER COST PER OUTCOME

Route by difficulty	Small model for simple tasks. Gartner's main defence against the inference paradox.
Prompt caching	Cache reads at 10% of base input price on most current Anthropic models.
Batch processing	50% cheaper for work that can wait.
Limit output	Output tokens cost five times input.
Cap agent steps	Removes the long tail of runaway tasks.

Prices from Anthropic's pricing documentation, October 2026; other providers differ. US-only inference costs 1.1×: a governance choice whose cost belongs in the line. The one-tenth aim is our judgement.

THE MONTHLY AI COST REVIEW

One page, three numbers, owners present.

01

Cost per outcome against benefit per outcome

One row per live use case, with last month and the trend.

A negative unit margin is flagged for the next stage-gate.

02

The bridge

Volume, consumption per outcome and rate.

One line from the owner for any consumption change over 10%.

03

The untagged share

Percentage of AI spend with no use-case ID.

With the three largest untagged items.

Two habits from the finance side

Accrue AI costs from gateway logs at month end; do not wait for the invoice.

Budget and alert on cost per outcome, as well as total spend.

Our proposal. The 10% threshold and the accrual habit are our practice.

05

What to do this quarter.

The controls are cheap. The expensive thing is scaling a use case on a number nobody can defend.

SEVEN ACTIONS

Seven actions for this quarter.

1

Name the joint owners

A delivery owner and a finance owner for every use case.

2

Baseline before the next pilot

The one-page record, from systems of record, signed by finance.

3

Agree attribution at approval

Method, unit, estimator and grade in the pilot charter.

4

Tag every model call

The register ID on every call, from day one.

5

Build the cost-per-outcome line

Inference, retrieval, evaluation, review and platform share.

6

Write the gates and kill criteria

Fixed dates; one extension without the committee.

7

Report benefits by grade

Holdout and rollout results in one column, estimates in another, never summed.

QUESTIONS FOR THE INVESTMENT COMMITTEE

Five questions to put to management in writing.

- 1** Which of our AI use cases had a signed baseline before they started?
- 2** For each claimed benefit, which group over which weeks tells us what would have happened without the AI?
- 3** Is each figure per licence paid for, or per active user?
- 4** What does one outcome cost us today for each use case, and what share of our AI spend has no owner?
- 5** When did we last stop funding an AI pilot, and what did we learn from it?

Drawn from the question lists at the end of each part.

WHAT TO TAKE AWAY

One lesson from each part.

01

Make the chain a joint product

Four controls, each prepared by one office and signed by the other, so neither can pass a gate alone.

03

Choose the comparison first

A group measured over the same weeks, written into the charter before results exist, booked per licence.

02

Sign the baseline first

One page, a full business cycle, the natural spread, and a plan for released time, frozen before the first model call.

04

Put the cost on the benefit's line

Same ID, same outcome unit, reviewed monthly on three numbers.

WHERE WE COULD BE WRONG

Four limits we set out in the series.

PART 1 · STRATEGIC BETS**Unit economics can stop the wrong work**

New products and market position need a gate judged on option value.

PART 2 · HOLDOUTS EVERYWHERE**The historical baseline would matter less**

We would still keep the record for the cost to serve, the guardrail and released time.

PART 3 · CHEAP PILOTS**Precision may cost more than it is worth**

A low-graded comparison may do, as long as the grade travels with the number.

PART 4 · SMALL ESTATES**A gateway and bridge may be overkill**

We would still tag every call from the start; adding tags later means rebuilding the history.

How strong each claim is: surveys and forecasts.

Claim	Source	Grade
25% of AI initiatives delivered expected ROI; 16% scaled	IBM CEO Study, May 2025, 2,000 CEOs	Reported, executive survey
At least 30% of GenAI projects abandoned after proof of concept by end-2025	Gartner, July 2024	Forecast
"95% of organizations are getting zero return"	MIT NANDA, July 2025; 52 interviews, 153 survey responses	Reported, contested
76% have a CAIO, up from 26%; 62% of CFOs' roles expanded into AI	IBM CEO Study, May 2026; IBM CFO Study, September 2026	Reported, executive surveys
Less than one in five tracked KPIs for gen AI; 37% attribute any EBIT to AI	McKinsey State of AI, March 2025 and August 2026	Reported, executive surveys
98% of practitioners manage AI spend, from 31% in 2024	State of FinOps 2026, 1,192 respondents	Reported, practitioner survey
Price at fixed performance falls 9x to 900x a year	Epoch AI, March 2025	Measured, with the authors' caveat
5 to 30x more tokens per task for agents; >5x cost per agentic workflow through 2028	Gartner, March 2026 and 2026	Forecast
MAS AI risk guidelines final 7 October 2026; thresholds mutually agreed	MAS Guidelines on AI Risk Management, para 5.14(a)	Measured, regulator record
No Singapore guidance requires an AI benefits baseline	Our search of MAS, IMDA and GovTech publications	My call

Experiments, guidance and our own proposals.

Claim	Source	Grade
758 consultants; 12.2% more tasks, 25.1% faster; 19% less likely correct outside the frontier	Dell'Acqua et al., Organization Science, 2026	Measured, randomised experiment
5,172 agents; 15% average, 30% for novices; quality declines for most experienced	Brynjolfsson, Li and Raymond, QJE, 2025	Measured, staggered rollout
16 developers, 19% slower, believed 20% faster	METR, July 2025; METR says results are out of date	Measured, early-2025 tools
26.08% more tasks, standard error 10.3%, 4,867 developers	Cui et al., Management Science, 2026	Measured, randomised experiments
80% of treated workers used the tool; about two fewer hours a week on email	Dillon et al., arXiv 2504.11436 v4, 2025	Measured, randomised field experiment
CUPED cut variance by about 50% at Bing	Deng, Xu, Kohavi and Walker, WSDM 2013	Measured, company platform
About 3% of hours saved; no earnings or hours effects above 2%	Humlum and Vestergaard, NBER WP 33777, rev. March 2026	Measured, administrative data
Baseline at initiation; evaluation methods and parallel trends	UK IPA benefits guide, 2017; HM Treasury Magenta Book, 2026	Measured, government guidance
3% against 9%; 12% against 5 points; S\$19,000 cost line and bridge	Illustrative examples built for the posts	Illustrative
Decision rights, baseline record, gates, kill criteria, method costs, review	Our proposal	My call

Where a study has a published version and an earlier working paper with different numbers, we used the published one. The worked examples are illustrative, and the decision rights, baseline record, gates, cost lines and review are our proposal, adapted from UK benefits and evaluation guidance.

FOR DISCUSSION

Four questions for the room.

- 1** Which AI benefit in our last board pack would survive an audit committee's questions about its method?
- 2** If we froze a baseline for our next pilot today, who would pull the data, and who would sign it?
- 3** Which pilot would we stop tomorrow if we applied our own kill criteria honestly?
- 4** Could we say this month which use case ran which part of our AI bill?

SOURCES · 1 OF 3

Works cited in Part 1.

- 1 IBM. (2025). IBM study: CEOs double down on AI while navigating enterprise hurdles.
newsroom.ibm.com
- 2 Gartner. (2024). Gartner predicts 30% of generative AI projects will be abandoned after proof of concept by end of 2025.
gartner.com
- 3 Gartner. (2025). Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027.
gartner.com
- 4 Challapally, A., Pease, C., Raskar, R., and Chari, P. (2025). The GenAI divide: state of AI in business 2025. MIT NANDA.
webberwentzel.com
- 5 Wharton School and GBK Collective. (2025). Accountable acceleration: gen AI fast-tracks into the enterprise.
knowledge.wharton.upenn.edu
- 6 Gartner. (2026). Finance AI press releases, 24 March, 28 May and 24 September 2026.
gartner.com
- 7 IBM. (2026). IBM study: CEOs are reshaping C-suite roles for the AI era.
newsroom.ibm.com
- 8 IBM. (2026). IBM study: as AI scales enterprise-wide, CFOs play an expanded role in transformation.
newsroom.ibm.com
- 9 Infrastructure and Projects Authority. (2017). Guide for effective benefits management in major projects.
apm.org.uk
- 10 Dell'Acqua, F., et al. (2026). Navigating the jagged technological frontier. Organization Science, 37(2), 403-423.
doi.org
- 11 Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics, 140(2), 889-942.
doi.org
- 12 DBS Group. (2026). Annual report 2025: letter from the chairman and CEO.
dbs.com
- 13 Stanford HAI. (2025). AI Index report 2025.
hai.stanford.edu
- 14 Epoch AI. (2025). LLM inference prices have fallen rapidly but unequally across tasks.
epoch.ai
- 15 Gartner. (2026). Gartner predicts that by 2030 performing inference on an LLM with 1 trillion parameters will cost GenAI providers over 90% less than in 2025.
gartner.com
- 16 FinOps Foundation. (2026). State of FinOps 2026.
data.finops.org

SOURCES · 2 OF 3

Works cited in Parts 1 and 2.

- | | |
|---|---|
| <p>17 FinOps Foundation. (2026). FinOps for AI overview.
finops.org</p> <hr/> <p>18 Storment, J.R. (2026). Token economics: the atomic unit of AI value. FinOps Foundation.
finops.org</p> <hr/> <p>19 FinOps Foundation. FOCUS: the FinOps open cost and usage specification.
focus.finops.org</p> <hr/> <p>20 Linux Foundation. (2026). Linux Foundation announces Tokenomicon, a new conference for the economics of AI.
linuxfoundation.org</p> <hr/> <p>21 IFRS Foundation. IAS 38 Intangible assets.
ifrs.org</p> <hr/> <p>22 IFRS Interpretations Committee. (2021). Configuration or customisation costs in a cloud computing arrangement.
ifrs.org</p> <hr/> <p>23 Cooper, R.G. (1990). Stage-gate systems: a new tool for managing new products. Business Horizons, 33(3), 44-54.
doi.org</p> <hr/> <p>24 UK Government. (2026). Gate review 5: operations review and benefit realisation.
gov.uk</p> <hr/> | <p>25 Monetary Authority of Singapore. (2026). MAS sets out supervisory expectations on responsible AI adoption by financial institutions.
mas.gov.sg</p> <hr/> <p>26 Ministry of Digital Development and Information. (2024). AI initiatives launched to uplift Singapore's economic potential.
mddi.gov.sg</p> <hr/> <p>27 Ministry of Finance, Singapore. (2026). Budget 2026 statement: harness AI as a strategic advantage.
singaporebudget.gov.sg</p> <hr/> <p>28 HM Treasury. (2026). The Magenta Book: central government guidance on evaluation.
gov.uk</p> <hr/> <p>29 Frontier Economics for HM Treasury. (2024). Guidance on the impact evaluation of AI interventions.
assets.publishing.service.gov.uk</p> <hr/> <p>30 McKinsey & Company. (2025). The state of AI: how organizations are rewiring to capture value.
mckinsey.com</p> <hr/> <p>31 McKinsey & Company. (2026). The state of AI in 2026: on the road to ROI.
mckinsey.com</p> <hr/> <p>32 METR. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity.
metr.org</p> <hr/> |
|---|---|

SOURCES · 3 OF 3

Works cited in Parts 2, 3 and 4.

- 33 METR. (2026). We are changing our developer productivity experiment design.
metr.org
- 34 Barnett, A.G., van der Pols, J.C., and Dobson, A.J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215-220.
doi.org
- 35 Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy online controlled experiments: a practical guide to A/B testing*. Cambridge University Press.
doi.org
- 36 Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *WSDM 2013*.
doi.org
- 37 Monetary Authority of Singapore. (2026). Guidelines on artificial intelligence risk management.
mas.gov.sg
- 38 Humlum, A., and Vestergaard, E. (2026). Still waters, rapid currents: early labor market transformation under generative AI. NBER Working Paper 33777.
nber.org
- 39 HM Treasury. (2026). Magenta Book Annex A: analytical methods for use within an evaluation.
gov.uk
- 40 Cui, K.Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., and Salz, T. (2026). The effects of generative AI on high-skilled work: evidence from three field experiments with software developers. *Management Science*.
doi.org
- 41 Dillon, E.W., Jaffe, S., Immorlica, N., and Stanton, C.T. (2025). Shifting work patterns with generative AI. *arXiv 2504.11436*.
arxiv.org
- 42 Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254-277.
doi.org
- 43 Callaway, B., and Sant'Anna, P.H.C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200-230.
doi.org
- 44 Kastelman, D., and Ramesh, R. (2018). Switchback tests and randomized experimentation under network effects at DoorDash. DoorDash Engineering.
careersatdoordash.com
- 45 Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493-505.
doi.org
- 46 Lopez Bernal, J., Cummins, S., and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, 46(1), 348-355.
doi.org
- 47 Gartner. (2026). Gartner predicts AI inference costs per agentic workflow will increase more than fivefold through 2028. Press release.
gartner.com
- 48 Anthropic. (2026). Pricing. Claude Platform documentation.
platform.claude.com

COMING NEXT IN THE SERIES

Two parts still to come.

Part 5

Stage-gates for AI pilot funding

Gates with kill criteria written in advance, decision rights, and a template for the stop decision.

Part 6

Reporting AI benefits to the board

Realised against forecast, net of run cost, with the confidence of each figure stated. A one-page report that survives an audit committee.

Parts 5 and 6 are provisional and may be renamed before they are written.

AI BENEFIT REALISATION · PARTS 1 TO 4

Thank you.

Read the full series at
terencekok.com/blog/series/ai-benefit-realisation/

A gate that stops a weak pilot early is the cheapest control an AI programme will ever buy.

Terence Kok and Dr. Christina Chua
terencekok.com



Read the series