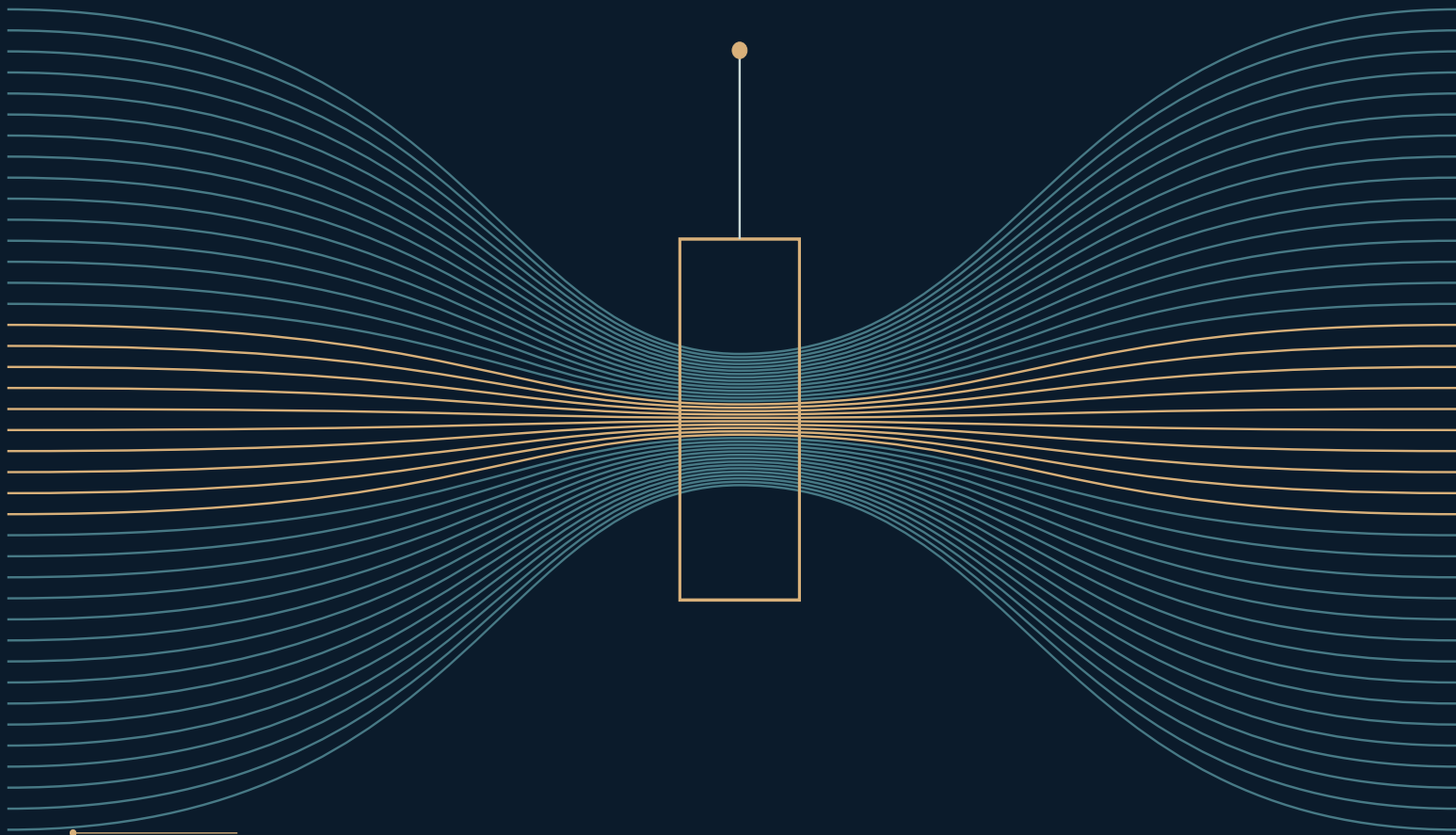


STRATEGY / OCTOBER 2026

The AI productivity question needs a better denominator

Why promising task results and modest business effects can both be true



THE DECISION IN BRIEF

An AI task gain becomes enterprise value only when management changes how the released capacity is used and verifies the resulting outcome.

- 01

Positive, negative and null AI findings often measure different things. Writing speed, resolved cases, developer time and recorded working hours cannot be combined into one productivity percentage.
- 02

Trace the same work from task performance through the complete workflow to an enterprise result. Review effort, demand and downstream capacity determine whether faster preparation changes anything that the business values.
- 03

Give each claimed gain a conversion decision: redeploy capacity, expand a valued service, improve assessed quality or change expenditure. Report what is observed, what that decision could realize and what remains untested.

THE MAYDEN VIEW

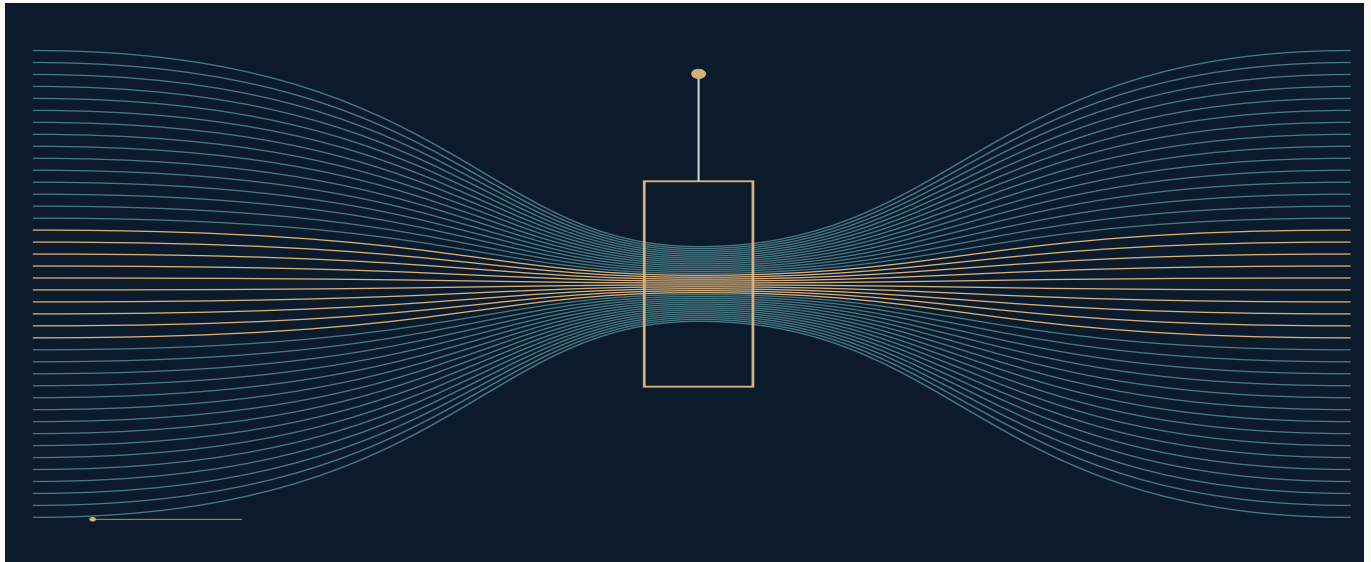
Released time becomes enterprise value when someone makes an operating decision that uses it.

INSIDE THIS PERSPECTIVE

The argument at a glance	3
01 / Read the evidence at its scale	4
02 / Trace the conversion	5-6
Exhibit 01 / Trace a faster draft to an enterprise result	7
03 / Design a credible comparison	8-9
Exhibit 02 / Different studies measure different effects	10
04 / Report realised value	11
The leadership agenda	12
Method and references	13

ONE DECISION, FOUR CONNECTED QUESTIONS

The constraint determines the result



PRODUCTIVITY BOTTLENECK / CONCEPTUAL ARTWORK

01 Read the evidence at its scale

A faster draft needs a conversion decision

02 Trace the conversion

Negative and null findings are equally decision-relevant

03 Design a credible comparison

Include the work that the tool creates

04 Report realised value

The case for moving faster remains credible

Released time becomes enterprise value when someone makes an operating decision that uses it.

CHAPTER 01

01

Read the evidence at its scale

A faster draft needs a conversion decision

An analyst finishes a draft faster. The business benefits only if the draft is accepted and the released capacity changes an outcome worth buying. It may support more decisions, better analysis or less expenditure. It may also move effort to a senior reviewer. The productivity question is therefore what happens to the complete work, not simply how quickly the first output appears.

Specify the ratio before claiming the gain: output, input, quality threshold and observation period. Drafts per hour, accepted decisions per hour and cost per resolved case answer different questions. None can stand in for the others without an explicit link. The denominator determines which management decision the result can support.

mayden's interpretation is that some apparent contradictions reflect a change of measurement level. AI improves an activity while another part of the system determines delivery. It can also create new work that consumes the released time. Test both explanations locally - and retain the possibility that the tool does not help the intended work.

Read positive findings within their tested scope

Noy and Zhang's randomized experiment involved 453 professionals performing selected writing tasks. It reported a 40% reduction in completion time and an 18% increase in assessed quality. ^[1] The design establishes a result in those assignments. It does not establish that a department can reduce its entire cost base by the same percentage. That distinction becomes material when writing is only one stage before fact checking, approval and implementation.

A field study of 5,172 support agents reported a 15% average increase in issues resolved per hour, with differing effects by experience. ^[2] This moves closer to an operational outcome, but remains one deployment and company context. The relevant transfer question is whether the new organization has comparable work, evidence, users and incentives. An effect becomes useful for planning when its mechanism matches the proposed change, rather than merely because the occupation has a similar name.

Before borrowing an external result, compare task boundaries, worker experience, tool version, quality assessment and surrounding support. Those differences define what the local trial must resolve. A study offers a starting expectation; it does not supply the organization's operating estimate.

CHAPTER 02

02

Trace the conversion

Negative and null findings are equally decision-relevant

METR's early-2025 randomized study covered 16 experienced open-source developers and 246 issues in their own repositories. AI access increased completion time by 19% in that setting. ^[3] Its small, specialized population and tool period limit generalization. It is nevertheless a useful warning against equating impressive benchmark performance or perceived helpfulness with measured speed. The appropriate response is to examine the conditions, not to turn the finding into a universal claim that AI slows developers.

The same researchers' February 2026 update says selection effects make their follow-up a weak basis for estimating the current effect. Developers and tasks that benefit most may be

missing because participants do not want to work without AI. ^[4] This matters for internal evaluation: participation can become selective precisely as tools become useful. A clean-looking comparison is not persuasive if the groups differ in ways that determine the outcome.

Humlum and Vestergaard linked Danish adoption surveys covering about 25,000 workers across 11 occupations to administrative records. Their April 2025 paper found no significant effects on earnings or recorded hours over its early observation period. ^[5] These labour-market outcomes differ from task speed and company profit. The finding constrains claims about rapid economy-wide transmission without proving that individual workers or firms obtained no benefit. Keep the outcome and horizon attached to the result.

CHAPTER 02

02

Trace the conversion

Follow the gain to its business consequence

Follow four levels of evidence. Task performance asks whether accepted work takes less effort or reaches a higher standard. Workflow performance asks whether the complete process improves. Enterprise value asks how that change affects cost, revenue, service or risk. Distribution asks who receives the benefit. This is mayden's analytical bridge; each level requires its own comparison.

A result at one level does not automatically establish the next. Suppose a hypothetical analyst prepares a useful first draft faster. The team may spend the released time exploring more alternatives, which could improve a decision without changing headcount. Alternatively, the draft may create more verification work for a senior colleague. Both possibilities belong in the evaluation. Counting only the analyst's time credits the system with a gain before examining whether it moved work elsewhere.

A randomized field experiment spanning 66 firms and 7,137 knowledge workers helps locate the conversion gap. In the second half of the six-month experiment, the authors report two fewer weekly email hours among the 80% of treated workers who used the tool, but no detected shift in task quantity or composition from individual-level provision. ^[8] This is evidence of a bounded time benefit, not proof of enterprise savings or of the returns to coordinated redesign.

Exhibit 01 turns the analyst example into a conversion ledger. Follow the same cases through downstream verification and record the decision that uses any released capacity. Lower expenditure requires a changed spending commitment; more output requires demand and capacity; better quality requires an assessed improvement. If the link remains unproven, report a task benefit with unresolved conversion. Then examine how any realized value is shared among employees, customers and owners.

ILLUSTRATIVE DESIGN · NO MEASURED RESULTS

Trace a faster draft to an enterprise result

Illustrative decision record

Conversion	Case record	Missing evidence	Accountable owner
Task change	Accepted analyst draft; assisted versus existing process	Quality and effort on comparable cases	Analyst manager
Downstream effort	Verification, revision and exception time	Whether work moved to a senior colleague	Workflow owner
Capacity action	Reassign capacity to an identified queue, or change spending	Demand, staffing commitment and ability to redeploy	Operations with finance
Enterprise outcome	Accepted additional work, service improvement or changed expenditure	Durability and who receives the benefit	Business sponsor

THE MANAGEMENT IMPLICATION

A task gain earns an enterprise claim only when each conversion is evidenced.

Illustrative analyst workflow, not measured gains. Record the same cases and period; do not add hours, output and money into one index. Source: mayden analysis. Proposed framework or illustrative scenario.

CHAPTER 03

03

Design a credible comparison

Include the work that the tool creates

New systems introduce setup, checking, maintenance and exception work. Some of this is an initial investment; some remains in steady operation. The productivity J-curve research models complementary intangible investments and their effect on measured productivity. ^[6]

The implication for a company is to separate the transition account from the operating account. The theory does not license indefinite losses or establish when a particular investment will pay back.

A useful operating account follows representative cases from intake to accepted completion. Include retries, discarded outputs, reviews, corrections and time spent recovering from failures. Track work done by adjacent

teams as well as direct users. Record the model and workflow version because a later result may reflect a different intervention. These requirements are practical: without them, a deployment can look increasingly efficient as more of its costs disappear into somebody else's budget.

Treat quality-adjusted output carefully. A single weighted index may conceal a serious defect behind gains in speed or convenience. Report a primary outcome alongside separate guardrails for material errors and customer consequences. If a composite measure is necessary, publish the weights and show how the conclusion changes under reasonable alternatives. A ratio that can reverse its meaning after a minor change in weighting is too fragile to carry the investment case alone.

CHAPTER 03

03

Design a credible comparison

Design the comparison before the rollout changes behaviour

UK guidance addresses causal evaluation and the context of changing AI systems. ^[7] Begin with the decision the test must support. Feasibility asks whether the task can be performed. Operating evaluation asks whether ordinary users can deliver accepted work. Economic evaluation asks whether the complete result warrants its cost. A persuasive demonstration does not answer all three.

Where feasible, randomize access or rollout timing among comparable units and preserve a clear baseline. Where randomization is impractical, use a credible comparison and document concurrent changes in demand, staffing, policy and case mix. Do not label a before-and-after difference causal without examining these explanations. An evaluation can still be useful when causal certainty is limited, provided the decision is proportionate to the strength of the evidence and the remaining uncertainty is visible.

Measure participation as well as outcomes. Who declined the tool, which cases were excluded

and which assignments were abandoned? A result based only on completed AI-assisted cases can omit the very failures that determine operational value. Capture reasons for non-use without assuming resistance is irrational. Users may identify risks, missing context or additional work that the evaluation team has overlooked. Their experience is evidence to investigate, not a substitute for measured performance.

Team-level effects deserve a separate check. If assisted employees share better drafts or advice with a comparison group, the measured difference between groups may understate the overall benefit. If the assisted team consumes scarce reviewers that the comparison team also needs, the comparison may be distorted in the other direction. Document these shared dependencies before selecting the unit of assignment. The appropriate evaluation unit may be a whole team or workflow rather than an individual. This does not require pretending that the workplace is a laboratory; it requires explaining where the comparison is informative and where interference limits the conclusion.

RESEARCH COMPARISON · INCOMPATIBLE MEASURES

Different studies measure different effects

Separate study outcomes; no pooled estimate

Study / design	Unit and observation boundary	Reported result	Comparison limit
Writing; randomised experiment [S1]	453 professionals; selected assignments	40% less task time; 18% higher assessed quality	AI access versus no access; not whole-job productivity
Support; field study [S2]	5,172 agents; one deployment	15% more issues resolved per hour	Study estimates assistance effect; heterogeneous experience groups
METR; randomised early-2025 study [S3]	16 developers; 246 issues in their own repositories	19% longer completion time	AI allowed versus disallowed; specialised work and dated tools
Danish survey and administrative records [S5]	About 25,000 workers; 11 occupations; early observation period	No significant earnings or recorded-hours effects	Labour-market outcome; not a task-speed experiment

THE MANAGEMENT IMPLICATION

The relevant question is which mechanism and measurement boundary resembles the intended deployment.

Effects use different metrics, populations and designs. Do not average them. METR's February 2026 follow-up identifies selection bias and does not establish a clean current estimate. Observation boundaries are not interchangeable follow-up horizons. See each source locator for its estimation design; no missing interval is inferred. Sources: [1], [2], [3], [4], [5]

CHAPTER 04

04

Report realised value

The case for moving faster remains credible

A measurement programme can become an obstacle when it demands certainty before any useful action. For low-consequence, reversible work, modest evidence may justify wider access while learning continues. A company should not spend more on evaluating a small convenience improvement than the decision warrants. The standard should rise with consequence, commitment and difficulty of reversal. Rigour means choosing a suitable burden of proof, rather than applying the most elaborate method to every tool.

There is also a risk that current metrics miss new value. A team may attempt a useful analysis it previously could not afford, or serve a customer segment it previously excluded. Comparing only the old task can miss this expansion. Evaluate the new proposition directly: does someone value the additional service, can the organization deliver it reliably and what does it cost? Novel output is not automatically valuable, but it deserves a measurement design that can recognize value when present.

External findings also age. A dated negative study should not be used to dismiss a later capability; a dated positive study should not justify a changed system without validation. Keep an evidence expiry condition tied to material changes in the tool, task or user population. This avoids both permanent scepticism and permanent optimism. The organization maintains a current argument about its own deployment instead of borrowing a fixed position from a fast-moving literature.

Fund the missing link in the value account

Give management three separate accounts. Observed results identify population, comparison and period. Plausible conversion identifies the funded operating change that could turn capacity into value. Untested opportunity identifies a proposition and the next test. This distinction shows where capital and management attention are needed without requiring every initiative to claim completed savings.

Make operations and finance reconcile the account. Operations establishes sustainable throughput and usable capacity; finance establishes changed expenditure. Commercial leaders test additional demand, while people leaders examine workload and capability. Resolve incompatible definitions before aggregating results across functions.

The next investment should address the break in the conversion chain. If task performance is weak, improve or stop the tool. If the workflow is constrained, redesign the handoff. If capacity has no valued use, revise the economic claim. Fund the link that the evidence identifies rather than another headline productivity target.

FROM INSIGHT TO ACTION

Decisions to take forward

BUSINESS SPONSOR

Accept only the claim measured

Evidence to require: Population, comparison and observation period

OPERATIONS

Choose a documented redeployment

Evidence to require: Downstream effort and usable capacity

FINANCE

Recognise cash savings only when spending changes

Evidence to require: Changed expenditure versus released hours

mayden connects senior judgement, research and engineering to turn strategic choices into working capabilities.

EVIDENCE YOU CAN INSPECT

Research notes

Method and boundaries

Narrative comparison of eight primary studies and methodological sources. Contrasting findings are retained; incompatible outcomes are not statistically pooled.

Selected experimental, field and administrative evidence to distinguish task, workflow and labour-market outcomes. Sources reopened 10 October 2026.

This is not a systematic review of all productivity literature. Studies differ in date, tools, populations and causal design. The evidence bridge is mayden analysis, not a validated score or forecast.

Primary sources and author research

[1] Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence

Noy and Zhang; author manuscript / July 2023. Abstract; experiment design. Selected short tasks, not whole jobs or company profit.

[Read the source](#)

[2] Generative AI at Work

Brynjolfsson, Li and Raymond; arXiv v 2 / 6 November 2024. Abstract. One company; effects vary by experience.

[Read the source](#)

[3] Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

METR / July 2025. Abstract; Section 3. Small specialized population and early 2025 tools; not a current universal estimate.

[Read the source](#)

[4] We are Changing our Developer Productivity Experiment Design

METR / 24 February 2026. Opening findings; selection-effects discussion. Does not provide an unbiased current effect estimate.

[Read the source](#)

[5] Large Language Models, Small Labor Market Effects

Humlum and Vestergaard; University of Chicago BFI / 14 April 2025. Abstract; introduction; administrative outcomes through June 2024. Danish context and short horizon; earnings/hours are not task speed or company profit.

[Read the source](#)

[6] The Productivity J-Curve: How Intangibles Complement General Purpose Technologies

Brynjolfsson, Rock and Syverson; AEJ Macroeconomics / January 2021. Abstract. No guarantee of returns to an individual programme.

[Read the source](#)

[7] Guidance on the Impact Evaluation of AI Interventions

HM Treasury and Evaluation Task Force / Updated 15 May 2026. Evaluation design and context. Company application is mayden interpretation.

[Read the source](#)

[8] Shifting Work Patterns with Generative AI

Dillon, Jaffe, Immorlica and Stanton; American Economic Review: Insights / Forthcoming; accessed 10 October 2026. Publisher abstract; NBER w33795 revised November 2025 corroborates version. Access randomized; two-hour abstract figure refers to users, not an unconditional all-worker treatment effect. No measured enterprise profit; vendor-employed authors.

[Read the source](#)