

AI VALUE IN THE REAL WORLD

Why AI Productivity Gains Do Not Automatically Become Business Results

A structured review of 20 studies and cases - plus a practical measurement model for B2B teams.

The productivity chain has four links.

Faster task -> accepted work -> freed capacity -> business result

Most public evidence measures the first link. Very little measures all four.

Executive findings

The 20-source review finds substantial evidence that AI can accelerate bounded tasks. It finds much thinner evidence that the saved time survives review, is redeployed productively, and changes revenue, cost, retention or another business outcome.

75% include a baseline or comparison

Fifteen sources compare AI-assisted work with a control, earlier period or non-adopter group. The strength of those baselines varies from randomized trials to vendor before-and-after cases.

5% capture verification inside the measured workflow

Only one source in the coded set clearly measures realistic end-to-end task time against a quality threshold that includes the friction of using and checking AI. Most quality studies score the final output but do not price the reviewer time required to reach it.

40% track a business or operational outcome

Eight sources connect AI to an outcome beyond individual speed or output quality. Several are self-reported surveys or vendor cases, so traceability does not automatically mean causal proof.

Bottom line

AI productivity is a task-level input. Business value appears only when quality holds, freed capacity is deliberately redeployed, and a downstream outcome changes.

The conversion chain

A productivity claim becomes a business claim only after it crosses three additional gates.



Where the chain usually breaks

A faster first draft may still create source checks, brand corrections, handoffs and rework. Even accepted work may not reduce cost if the freed time is absorbed by more low-priority production. Finally, more activity does not guarantee demand, conversion, retention or margin.

This is the productivity paradox in operational form: local efficiency can improve while the business result remains unchanged.

How the 20 sources were coded

The review combines experiments, field data, surveys and public enterprise cases. Each source was coded against three questions using only information reported publicly.

BASELINE / CONTROL

Yes when a source contains a control group, pre-period, non-adopter comparison or other explicit reference point. A baseline can be present without proving causality.

REVIEW COST CAPTURED

Yes only when the measured workflow includes the human effort needed to inspect, correct or validate the AI-assisted result against a real acceptance threshold.

BUSINESS RESULT

Yes when the source traces AI to revenue, cost, customer operations, labor outcomes or another organizational result beyond individual speed and output quality.

EVIDENCE STRENGTH

Randomized and causal designs support stronger conclusions than self-reported surveys. Vendor cases can be useful operationally, but usually lack independent validation.

NO POOLED ROI

The percentages are not averaged. The tasks, tools, samples and outcome definitions are too different for a defensible universal ROI estimate.

The measurement gap in 20 public sources

Teamulate coding of the selected evidence set. Counts reflect reported measurement, not overall study quality.

15 / 20

Baseline or comparison



1 / 20

Review cost captured



8 / 20

Business result tracked



The rarest measurement is the most important one.

Without review and correction time, a productivity study may measure how quickly AI creates something - not how quickly a business receives work it can safely use.

Strong task-level gains exist

Several controlled studies show material gains on clearly scoped tasks. These results establish capability, but not a universal rate of return.

PROFESSIONAL WRITING

Noy and Zhang report roughly 40% less completion time and 18% higher rated quality for professional writing tasks with ChatGPT.¹

CUSTOMER SUPPORT

Brynjolfsson, Li and Raymond find a 15% average increase in issues resolved per hour across 5,172 support agents, with larger gains for less experienced workers.²

CONSULTING - INSIDE THE FRONTIER

In a 758-person experiment, GPT-4 users completed 12.2% more tasks, finished 25.1% faster and produced more than 40% higher quality on suitable tasks.³

SOFTWARE DEVELOPMENT

A controlled GitHub Copilot experiment reports 55.8% faster completion of a defined coding task.⁴

ONLINE RETAIL

Seven field experiments produced sales effects ranging from 0% to 16.3%. Four applications showed positive effects, while three did not.⁹

The same technology can also create zero or negative gains

The evidence changes when tasks are complex, implicit, unfamiliar to the model or embedded in a high-quality real workflow.

19% SLOWER

METR found experienced open-source developers took 19% longer with early-2025 AI tools on real issues in repositories they knew well.⁵

PERCEPTION WAS WRONG

Before the METR tasks, developers expected to be 24% faster. Afterward, they still believed they had been 20% faster despite the measured slowdown.⁵

OUTSIDE THE FRONTIER

Consultants using AI were 19 percentage points less likely to reach the correct answer on a task outside the model's capability frontier.³

CONTEXT CHANGES THE EFFECT

A 2026 meta-analysis of 23 programming studies found a moderate positive average effect, but substantial heterogeneity and smaller gains in open-source and enterprise settings.⁸

SKILL STILL MATTERS

A 2026 experiment found that the ability to elicit, filter and verify AI output predicted who captured value; structured scaffolding reduced outcome variance.⁷

More activity is not automatically more value

Three widely cited signals answer three different questions.

21.2%

more productivity-app actions

Observed among high-use AI adopters in M365 trace data

1.4%

of total work hours reportedly saved

National worker survey estimate

0%

measured productivity difference

S&P 500 deep-adoption analysis

Interpretation

Application actions show changed behavior. Self-reported time savings show perceived efficiency. Firm productivity compares organizational output. These signals can all be accurate while pointing in different directions because they sit at different points in the conversion chain.

Enterprise surveys show the conversion gap

Adoption is widespread, but financial results remain concentrated. Survey findings are directional because outcomes are self-reported.

80% VS 37%

McKinsey's 2026 survey reports that 80% perceive individual productivity improvement, while 37% report at least some EBIT impact.¹⁴

56%

In PwC's survey of 4,454 CEOs, 56% report neither higher revenue nor lower costs from AI in the prior 12 months.¹⁶

12%

Only one in eight PwC respondents report both additional revenue and lower costs from AI.¹⁶

SHALLOW USE

The national U.S. adoption study estimates that only 1% to 5% of total work hours were AI-assisted, despite broad awareness and rapid adoption.¹¹

WHY THIS MATTERS

A tool can feel useful to many people and still be too lightly or inconsistently integrated to change an enterprise metric.

Enterprise cases are useful - with limits

Public cases show how value can appear operationally. They also tend to be authored by vendors or partners and rarely disclose full counterfactuals or implementation cost.

KLARNA

The published case reports customer resolution time falling from 11 minutes to under 2, a 25% drop in repeat inquiries, and an estimated \$40 million profit improvement.¹⁷

MORGAN STANLEY

The case reports document access rising from 20% to 80%, more than 98% advisor-team adoption, and an explicit eval framework with expert grading and daily regression tests.¹⁸

MODERNA

The case reports 750 internal GPTs within two months and 40% of weekly active users creating GPTs. It describes human-led validation but does not quantify a company-level financial result.¹⁹

RB2B / INTERCOM

The case reports 65% automatic resolution and more than 132 hours saved monthly. It does not publicly price the knowledge-maintenance and oversight work behind the result.²⁰

READ CASES AS LEADS

Use case studies to identify mechanisms and metrics worth testing. Do not treat them as transferable ROI forecasts.

Evidence matrix 01-10

Teamulate coding: baseline/control, review cost captured, and a result beyond individual speed or output quality.

# / Source	Type / setting	Baseline	Review cost	Business result	Reported signal
01 Noy & Zhang	RCT Writing	Yes	No	No	Time -40%; quality +18%
02 Brynjolfsson et al.	Field rollout Support	Yes	No	Yes	Throughput +15%
03 Dell'Acqua et al.	RCT Consulting	Yes	No	No	25.1% faster in-frontier
04 Peng et al.	Experiment Coding	Yes	No	No	Task 55.8% faster
05 METR	RCT Coding	Yes	Yes	No	19% slower
06 Cybernetic Teammate	Field RCT Innovation	Yes	No	No	Higher-quality output
07 Cruces et al.	RCT Problem solving	Yes	No	No	Education gap narrowed
08 Maier et al.	Meta-analysis Coding	Yes	No	No	$g=0.33$; high heterogeneity
09 Fang et al.	Field RCTs Online retail	Yes	No	Yes	Sales 0% to +16.3%
10 Humlum & Vestergaard	Survey + admin Labor market	Yes	No	Yes	Near-zero earnings/hours effect

Evidence matrix 11-20

Teamulate coding: baseline/control, review cost captured, and a result beyond individual speed or output quality.

# / Source	Type / setting	Baseline	Review cost	Business result	Reported signal
11 Bick et al.	National survey Workplace	No	No	No	1.4% of total hours saved
12 Yu et al.	Difference-in-diff. M365 activity	Yes	No	No	Actions +21.2%
13 Yu et al. S&P; 500	Longitudinal Enterprise	Yes	No	Yes	No productivity difference
14 McKinsey	Executive survey Enterprise	No	No	Yes	80% individual; 37% EBIT
15 Microsoft WTI	Worker survey Knowledge work	No	No	No	Oversight remains human
16 PwC CEO Survey	CEO survey Enterprise	No	No	Yes	56%: no revenue/cost benefit
17 Klarna case	Vendor case Support	Yes	No	Yes	11 to <2 min; \$40m estimate
18 Morgan Stanley case	Vendor case Advisory	Yes	No	No	Document access 20% to 80%
19 Moderna case	Vendor case Enterprise	No	No	No	Adoption and use cases
20 RB2B / Intercom case	Vendor case Support	Yes	No	Yes	132+ hours/month reported

Teamulate missing-measurement map

The handoff between one metric and the next is where claimed ROI most often disappears.



GAP 1

Speed without acceptance

A draft is faster, but reviewer time and correction are missing.

GAP 2

Acceptance without capacity

The asset passes, but no one verifies that usable hours were released.

GAP 3

Capacity without redeployment

Time is saved, but it is not assigned to a higher-value activity.

GAP 4

Activity without outcome

More work is produced, but customer or financial metrics do not move.

A measurement model for B2B marketing

Track one recurring workflow through all four links. Keep the brief, sources, quality gate and reviewer comparable.

Layer	Primary metric	Required companion	Decision
Task	Active human minutes	Tool cost	Is the task faster?
Quality	First-pass acceptance	Rework + issue log	Is the result usable?
Capacity	Hours released	Hours redirected	Did work actually change?
Business	One downstream KPI	Baseline + time window	Did the business benefit?

Example for a B2B campaign email

Measure human minutes to approved email, first-pass acceptance, factual or brand corrections, and variable AI cost. Then verify whether the released time was redirected to list quality, offer testing or sales follow-up. The downstream metric might be qualified replies or meetings - not email volume.

Minimum credible claim

“For this workflow, under these conditions, AI reduced human time to accepted work by X%, while first-pass acceptance changed by Y points and the selected business KPI changed by Z.”

Sources and scope notes

Review completed in September 2026. The set spans controlled studies, real-work field evidence, enterprise surveys and public cases. This is a structured evidence review, not a systematic meta-analysis.

1. Noy, S. and Zhang, W. [Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence](#). Science (2023).

2. Brynjolfsson, E., Li, D. and Raymond, L. [Generative AI at Work](#). NBER / arXiv (2023).

3. Dell'Acqua, F. et al. [Navigating the Jagged Technological Frontier](#). SSRN (2023).

4. Peng, S. et al. [The Impact of AI on Developer Productivity: Evidence from GitHub Copilot](#). arXiv (2023).

5. Becker, J. et al. [Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity](#). METR (2025).

6. Dell'Acqua, F. et al. [The Cybernetic Teammate](#). HBS Working Paper / SSRN (2025).

7. Cruces, G. et al. [Does Generative AI Narrow Education-Based Productivity Gaps?](#). arXiv (2026).

8. Maier, S. et al. [A Meta-analysis of the Effect of Generative AI on Productivity and Learning in Programming](#). arXiv (2026).

9. Fang, L. et al. [Generative AI and Firm Productivity: Field Experiments in Online Retail](#). arXiv (2025).

10. Humlum, A. and Vestergaard, E. [Large Language Models, Small Labor Market Effects](#). NBER (2025).

11. Bick, A., Blandin, A. and Deming, D. [The Rapid Adoption of Generative AI](#). Management Science / NBER (2026).

12. Yu, Y. et al. [Adoption of Generative AI in the Workplace](#). arXiv (2026).

13. Yu, Y. et al. [AI Adoption in S&P 500 Firms](#). arXiv (2026).

14. McKinsey [The State of AI in 2026](#). Report (2026).

15. Microsoft [2026 Work Trend Index](#). Report (2026).

16. PwC [29th Global CEO Survey](#). Report (2026).

17. OpenAI [Klarna](#). Customer case.

18. OpenAI [Morgan Stanley Uses AI Evals to Shape the Future of Financial Services](#). Customer case.

19. OpenAI [Moderna](#). Customer case.

20. Intercom [How RB2B Created a Lean, Scalable Support System](#). Customer case (2025).

Interpretation

Coding uses publicly reported methods and outcomes. 'Business result' covers financial, customer, operational or labor outcomes beyond individual speed and output quality. A 'Yes' does not imply causal proof; vendor cases remain reported claims rather than independently audited findings.

About Teamulate

Teamulate is a managed AI marketing department for growing B2B companies, combining AI execution with human judgment to produce controlled, measurable marketing work.