Analytical perspective, 15 November 2026. A narrative synthesis and proposed practice, not peer-reviewed research.

AI assistance in knowledge work: separating adoption, workflow gains and realised value

Mr. Gómez · Analytical perspective · 15 November 2026

Abstract

Generative AI can improve knowledge work. Published evidence shows faster and better professional writing in randomised tasks, customer-support gains under a quasi-experimental staggered rollout, and higher-rated product-innovation proposals relative to solo work without AI in a field experiment. Other evidence records worse answers outside a tested frontier, limited changes in measured office work patterns, and near-null average earnings and recorded-hours effects. These findings concern different interventions, populations, dates and outcomes. They are not interchangeable verdicts on whether AI works.

This paper asks how an enterprise can test whether AI assistance improves a defined knowledge-work workflow, and distinguish that improvement from adoption, perceived time savings and realised financial value. It argues for evaluating acceptable outcomes across the full production boundary, including downstream recipients' checking and rework, then documenting what happened to any capacity released. Quality, service, accessibility, learning and sustainable workload can justify investment without becoming cash savings.

Existing methods already supply the foundations, so no new branded framework is warranted. The proposed contribution connects them through three reporting artefacts that are not field-validated. The appropriate decision is not always to scale, nor always to stop. It is to identify which claim the evidence supports and what remains untested.

For enterprise practitioners, particularly CTOs, CIOs, CFOs and operating leaders. Not a systematic review, a new measurement theory or an original field study. Product and company names identify public research settings, not endorsements.

1. The number that answers a different question

Australian Treasury's evaluation of its 2024 Microsoft 365 Copilot trial makes a useful distinction. Participants found applications for basic administrative work, but the evaluation found no clear evidence that the tool improved final work outcomes during the trial. It did not explicitly quantify task time savings. Its calculation of approximately 13 minutes a week for an APS6 employee was the time required to offset the licence cost, not an observed saving and not a calculation of the full cost of introducing AI (Australian Centre for Evaluation, 2025).

A licence-cost threshold asks what would need to happen for one expense to be covered. An evaluation asks what actually changed. The lesson is not that the Treasury trial failed: a short trial can identify promising uses without establishing their eventual value. The problem begins when a conditional threshold becomes a measured benefit, or an easier drafting experience becomes an organisation-wide productivity claim.

Consider a research brief. Faster drafting may help its author, but the reviewer may still need to trace every source, correct unsupported conclusions and return the document. If approval takes longer, work has moved rather than disappeared. Conversely, a better first draft may genuinely reduce both people's effort and improve the decision it supports. These are hypothetical possibilities, not findings from a client engagement.

The research question is therefore operational rather than promotional: **How can an enterprise test whether AI assistance improves a defined knowledge-work workflow, and distinguish that improvement from adoption, perceived time savings and realised financial value?** Answering it requires four separate claims and an explicit decision about what value means, without assuming that value exists only when payroll falls.

2. How the evidence was selected and read

This is a targeted narrative synthesis, with primary sources checked in November 2026. Selection started from a pre-identified list of studies, then prioritised identifiable empirical designs beyond coding, government evaluations relevant to Australian enterprises, and sources that challenged a simple positive story. Designs, samples and results were checked against publisher or author material and official reports, and published versions replace earlier working-paper estimates where available. Working papers are not treated as peer-reviewed merely because NBER hosts them.

No exhaustive search protocol, registered inclusion rule or formal risk-of-bias instrument was used. Publication bias, English-language selection, sponsor access and the choice of familiar studies can distort this sample. Effects are not pooled, because their denominators are incompatible. Randomisation supports attribution within a setting, not universal applicability, and self-reports do not observe the alternative world without AI. Commercial links are disclosed where relevant; they neither automatically invalidate a result nor remove the need for independent replication. They include BCG collaboration and co-authors (Dell'Acqua, McFowland, et al., 2026); P&G gifts to an HBS institute, earlier consulting and company co-authors (Dell'Acqua, Ayoubi, et al., 2026); Microsoft involvement and employee authors (Dillon et al., 2025); a financial model commissioned by Microsoft (Forrester Consulting, 2024); and vendor-produced capability evidence and documentation (Microsoft, 2026; OpenAI, 2025).

3. Four claims, not one productivity number

Adoption means people use an available tool. A monthly active user, a daily user and a licence holder are different denominators. Bick et al.'s US task-level survey analysis finds widespread but shallow use: by May 2026, 45% of workers reported using generative AI for their jobs, with substantial variation across tasks. This is adoption evidence, not an estimate of output caused by AI (Bick et al., 2025, 2026).

Task improvement means a specified task changes in time, effort or quality. Even a credible causal effect applies only to the tested task and intervention, not to what happens when recipients must verify the result.

Realised operational capacity or service improvement concerns the workflow. Capacity released is not automatically capacity redeployed, and fragmented minutes may not support another useful assignment.

Financial return concerns attributable spending, margin and investment over a stated horizon. It requires a further reconciliation, not a salary multiplier applied to claimed minutes.

These are **different claim types**, not rungs on a single evidence-strength ladder. Causal credibility is a separate axis: an excellent randomised capability test can answer a narrow question convincingly, while a weak before-and-after profit chart answers an important one unconvincingly.

Claim type	Relevant evidence	Attribution question and boundary
Adoption	Defined usage telemetry or representative survey	Who uses what? Use is not itself a productivity effect.
Task effect	Randomised task access, expert assessment, time records	Did the intervention change the specified task? Include quality and actual comparator.
Workflow or organisational outcome	Accepted output, all participants' effort, queues, service and staff measures	Did the entire process improve, rather than move work to recipients? A counterfactual is still needed.
Financial or other realised benefit	Finance records or pre-agreed service, accessibility and workload indicators	What changed because of the intervention, for whom, net of relevant costs? Importance does not confer causal credibility.

4. What the empirical evidence establishes

4.1 Real gains, with boundaries

Noy and Zhang randomly exposed half of 453 college-educated professionals to ChatGPT during incentivised writing tasks. The 2023 published experiment reports 40% less completion time and an 18% improvement in assessed quality. This is causal evidence for bounded professional writing, not evidence that an employer's entire reporting process became 40% faster or its profit rose by a corresponding amount (Noy & Zhang, 2023).

Brynjolfsson et al.'s published customer-support study analyses the staggered introduction of an assistant, principally during 2020 and 2021, using 5,172 agents and approximately three million chats. Issues resolved per hour increased by 15% on average. Less experienced and lower-skilled agents improved speed and quality; the most experienced and highest-skilled had small speed gains and small quality declines. The main estimate is quasi-experimental, not a randomised assignment of all 5,172 agents (Brynjolfsson et al., 2025).

The final published *Cybernetic Teammate* paper reports a 2024 P&G experiment comparing solo and paired professionals, with and without AI, on real early-stage innovation challenges. It covers 791 participants and 550 submitted solution units; 776 participants supplied complete data including post-task surveys. Table 2's simple specification estimates quality gains of 0.373 standard deviations for AI-assisted individuals and 0.392 for AI-assisted teams, against 0.245 for teams without AI, all relative to solo work without AI. The AI-team versus non-AI-team difference is not statistically significant in that specification. The authors describe AI-assisted individuals as comparable to non-AI teams, which is not a formal demonstration of equivalence, and certainly not evidence that established teams can safely halve staffing. Gains came primarily from better initial ideas, not improved selection accuracy, and the trial does not measure product revenue or sustained coordination and rework (Dell'Acqua, Ayoubi, et al., 2026).

4.2 Quality can reverse the result

The consulting experiment published in 2026 tested 758 BCG professionals using a 2023 GPT-4 configuration. Within the tested capability frontier, AI users completed 12.2% more tasks and worked 25.1% faster on average. On a separate business problem deliberately outside that frontier, the

combined AI conditions reduced correctness by approximately 19 percentage points relative to control. These are different task arms, not the same participants succeeding and then failing (Dell'Acqua, McFowland, et al., 2026).

This is a warning against equating fluent output with acceptable output; it does not establish a fixed boundary that today's models cannot cross. An enterprise must test its own acceptance criteria, and human review is not a free, infallible correction layer: reviewers consume capacity and may miss plausible errors.

METR supplies a historical coding counterexample, deliberately outside this paper's main domain. Its early-2025 randomised study found 16 experienced developers took 19% longer on 246 repository tasks when AI was allowed. A February 2026 update calls later data unreliable, likely understating speed-up; a May 2026 convenience survey records perceived gains while questioning counterfactual self-reports (METR, 2025, 2026a, 2026b). The slowdown belongs here only as evidence that perception and measured performance can diverge in a particular setting, not as a current estimate of developer productivity.

4.3 Changed work patterns are not the same as enterprise output

Dillon et al.'s six-month randomised access study covers 7,137 workers in 66 firms during 2023 and 2024. Its November 2025 revision reports a reduction of 1.4 hours a week in measured Outlook sessions during months four to six as the effect of assignment to access. The approximately two-hour estimate is the effect of use induced by that assignment under the study's instrumental-variable assumptions. The researchers did not observe work content or output quality, and detected no meaningful broad changes in measured task quantity or composition; some after-hours activity fell. This is evidence of changed work patterns and potentially valuable relief, not a demonstrated enterprise-output increase or a demonstrated absence of one (Dillon et al., 2025).

Government evaluations offer broad experience but weaker attribution. The Australian whole-of-APS DTA report describes approximate task-time estimates as likely upper bounds under an assumption of daily task performance; editing was almost always required. It is separate from the Treasury trial (Digital Transformation Agency, 2024). The UK's 2024 cross-government deployment involved 20,000 staff; the 2025 report derives its widely repeated 26 minutes a day from self-reported response brackets, not a randomised time study, and explicitly could not identify how saved time was spent. Those findings justify investigation, not multiplying a headline by every licence and calling the result cash (Government Digital Service, 2025).

4.4 A labour-market null is not a productivity null

Humlum and Vestergaard link large Danish adoption surveys to administrative records through December 2024. Their March 2026 working-paper revision estimates near-null effects on earnings and recorded hours at worker and workplace levels, ruling out effects larger than 2% over the studied period, and documents task reorganisation, including AI oversight and integration work (Humlum & Vestergaard, 2026). That result does not measure firm profit, all output or every form of uncompensated effort. Equally, reported time savings do not prove that these administrative nulls conceal large productivity gains. Both interpretations need evidence.

Taken together, the studies support a narrower but useful conclusion: real gains occur, their distribution is uneven, and task effects do not determine downstream value automatically.

5. Existing methods already cover most of the problem

SPACE offers five dimensions of developer productivity, but **SPACE is not a validated business-user productivity score** (Forsgren et al., 2021). From DORA's software-delivery guidance, the transferable idea is to pair flow with failure and recovery, not to import deployment frequency into office work (DORA, 2026).

Task-exposure measures, such as Eloundou et al.'s rubric, and GDPval's initial 2025 evaluation across 44 occupations help shortlist work and assess capability. They do not measure adoption, human-plus-AI organisational performance or ROI, and GDPval's reported speed and cost comparisons explicitly exclude human oversight, iteration and integration (Eloundou et al., 2023; OpenAI, 2025). Total-factor productivity, unlike labour productivity, considers combined inputs, but neither measure attributes changes to AI (U.S. Bureau of Labor Statistics, n.d.-a, n.d.-b).

Cost-of-quality practice already distinguishes prevention, appraisal, internal failure and external failure: source checking is appraisal, correction before acceptance is internal failure, and a recipient's later discovery can reveal external failure (American Society for Quality, n.d.). Benefits realisation already requires baselines, owners, a counterfactual, whole-life costs and evidence that benefits occurred, distinguishes cash-releasing from non-cash benefits, and requires costs, disbenefits and double counting to be addressed (Government Finance Function & HM Treasury, 2025; Government Project Delivery, 2026). Commercial guidance also discusses redeploying capacity (Gartner, 2026). Commercial tools are prior art, not independent validation: Microsoft's Business Impact documentation warns that group differences can reflect other causes and are not tested for statistical significance, and Forrester's commissioned composite model, which projects an ROI range of 132% to 353%, supplies an appraisal structure, not an observed causal average return (Forrester Consulting, 2024; Microsoft, 2026).

Novelty verdict: this review does not establish a theoretical gap requiring a new framework. The practical gap is implementation: keeping task evidence, recipients' effort and benefits accounting connected. The protocol below is an implementation aid within existing practice.

6. A whole-workflow measurement protocol

6.1 Agree the decision and the workflow

Start with a decision that evidence can change, such as expanding access, redesigning a handoff or discontinuing an intervention. Name an operational owner and a benefit owner, and agree which outcomes matter before choosing the easiest telemetry. Record the workflow in a **workflow definition sheet**: trigger, acceptance point, downstream users, case-complexity classes and the period in which failures can emerge. Define the intervention precisely, including model, retrieval sources, training and changed responsibilities. Specify acceptance criteria with recipients before seeing results, and keep separate critical-error gates so that a serious privacy or compliance failure is not averaged away by attractive prose.

6.2 Establish baseline and counterfactual

A pre-intervention baseline, including recipients' checking, is not itself a counterfactual. Prefer random assignment of eligible teams or workflow units, and define business as usual honestly, including existing AI use. Pre-specify the primary outcome, the smallest improvement worth acting on and quality constraints, and set sample size and duration from baseline variation and that threshold, not a standard pilot length. Analyse assignment to access, including non-users, as the principal

intention-to-treat estimate, and report uptake, attrition and uncertainty. Without randomisation, consider a comparable untreated group and difference-in-differences; without a credible comparison, label findings descriptive and reduce the strength of the decision claim.

6.3 Measure quality, effort and flow together

Build an **operational evidence record** around the same work-item boundary in both groups, including verification, correction, handoffs and downstream recipients' rework. Separate active human effort from elapsed time: an agent can run while its operator works elsewhere, and adding overlapping durations double counts labour. Report accepted completions per calendar period and accepted outcomes per total attributable human effort, alongside paid staffing hours and demand. Lower active effort with unchanged output and payroll is possible capacity release, not automatically higher output per paid hour. Use independent, blinded review samples where practicable, and follow the review queue: faster drafting with older pending approvals may simply shift the bottleneck.

6.4 Count full costs and trace capacity

Include licences, compute, integration, training, change management, evaluation, oversight, and operators' and recipients' verification and rework. Record observed incidents separately from modelled expected losses; a short trial with no detected incident cannot establish that the risk is negligible.

Estimate net capacity released after checking and rework, as a range, without summing overlapping task-saving claims. Then record what happened, even if that is no identifiable redeployment, and have a named owner substantiate it with workflow evidence. Non-financial benefits deserve pre-agreed indicators and an owner. An accessibility benefit can warrant investment on its own terms; it need not be monetised into imaginary salary savings to count.

6.5 Reconcile two finance views and decide

Create a **financial reconciliation note** with two clearly separated views. The **cash and commercial view** records attributable spending reductions, incremental contribution margin and actual incremental cash costs, with avoided approved future expenditure separately labelled as cost avoidance. The **resource-efficiency view** records resources required per accepted outcome; salary-valued capacity can describe an opportunity but must not appear as cash released, and the same hours must not be valued twice.

Compare results with the pre-agreed threshold and quality gates, not only statistical significance. A wide interval is uncertainty, not proof of no effect, and a small precise gain can still be uneconomic. Scale only the demonstrated intervention and workflow, and retest after material model or workflow changes.

7. What enterprise leaders should decide differently

CTO: use exposure studies and capability benchmarks to shortlist work, not certify a business case. Require representative quality tests and severe-error stop conditions. Reject a rollout justified only by a capability leaderboard or a vendor-assisted-hours dollar figure.

CIO and operating leaders: own the handoffs. If a pilot helps authors but burdens reviewers, the decision may be to redesign the review process or narrow use. Reject a success claim based on adoption or author satisfaction while the approval queue worsens.

CFO: require the two views to reconcile, and challenge utilisation, demand and error-loss assumptions. Reject a cash-savings target built by multiplying claimed minutes by salaries, but do not reject a worthwhile accessibility or service investment merely because it does not reduce payroll.

Jointly: decide who captures the benefit. Faster work can become higher service, less overtime, margin or greater workload. Communicate those choices before evaluation so staff do not have to guess whether honest reporting will become an individual target.

8. Workers, learning, privacy and measurement incentives

The evidence does not support one rule for novices and experts. An October 2026 working-paper revision by Cruces et al. finds larger assisted gains among lower-education participants, with the higher- versus lower-education gap falling from 0.548 SD without AI to 0.139 SD with AI, and some retained unassisted learning in an immediate follow-up; it is an online business-problem experiment, not evidence of long-term workplace skill retention (Cruces et al., 2026). Evaluate learning rather than assume either inevitable deskilling or automatic upskilling.

Workload intensification deserves direct attention. Ranganathan and Ye's in-progress ethnographic research at one approximately 200-person technology company reports expanded task scope, fewer natural pauses and more simultaneous work. This is qualitative evidence of a possible mechanism, not a causal estimate of prevalence. Conversely, the office-suite RCT records reduced after-hours digital activity. Both outcomes are possible; measure the experience of the actual workforce (Counts, 2026; Dillon et al., 2025; Ranganathan & Ye, 2026).

Consult staff and, where relevant, representatives before instrumentation. Prefer existing work-item records, sampled effort and aggregate reporting over prompt-content capture, keystroke logging or screenshots, and suppress small-group reporting that could identify people. Australia's OAIC notes that the Privacy Act does not specifically cover workplace surveillance and that relevant Australian, state and territory laws matter. This paper is not a legal compliance checklist, and notification alone should not be treated as permission for any proposed monitoring (Office of the Australian Information Commissioner, 2024). Do not create individual productivity league tables from tools designed to assess workflow change.

9. The strongest countercase and limitations

The strongest countercase is that this approach can demand too much measurement for an ordinary tool. Real gains are already established in several settings, new capabilities can make old trials underestimate current benefits, and measurement can cost more than the decision warrants. The productivity J-curve offers a serious explanation for delayed returns, but it is not a guarantee that present costs will become future gains (Brynjolfsson et al., 2021).

The protocol is unsuitable as a standard numerical comparison where acceptance criteria are unstable, work is sparse or non-comparable, outcomes are mainly political or creative judgements, or intrusive instrumentation would be required. Infeasible measurement does not make the work valueless, nor does it justify an unsupported precision claim.

Limitations. The evidence base is selective and cannot estimate economy-wide or all-enterprise effects. Versioned working papers and current tools may have changed since these sources were checked in November 2026; readers should consult the latest versions. No original workflow has been tested, and the three artefacts have no reliability or economic validation. Any implementation requires staff participation and jurisdiction-specific privacy and legal review. Independent multi-

workflow studies showing that individual access routinely produces attributable enterprise returns without material workflow change would weaken this argument; replications linking full workflow outcomes, recipient effort and audited benefits would strengthen it. None of those tests has been performed by the author.

Version decisions and exclusions. For the *Cybernetic Teammate*, this paper uses the final article's separate denominators (791, 776 and 550), not the earlier NBER abstract's single 776. Table 9's exploratory top-decile effect is against solo no-AI control; in Table 9, model 1, team+AI versus team no-AI is not significant (p=.190). Treasury counts are internally inconsistent: the official full report's executive section reports 218 staff, Table 3 (printed p.10) uses 232, and Appendix B (printed p.46) gives 164 post respondents and 100 matched pre/post. This paper does not invent a reconciliation or use headline percentages dependent on those denominators; only the qualitative conclusion is retained, with this disclosure. The preliminary July 2025 *GenAI Divide* report is excluded as population evidence: its interviews, conference-leader responses and public initiatives do not establish a representative denominator or a consistent success/failure measure, so its widely circulated failure-rate headline is omitted (Challapally et al., 2025).

10. A decision checklist, not another score

Before claiming a productivity result, ask:

1. **Which claim are we making?** Adoption, a task effect, a workflow outcome, a financial benefit or another valuable outcome?
2. **Where does the work end?** Has acceptance included recipients' checking, rework and later failures?
3. **What supports attribution?** Is there a credible counterfactual, a stated comparison and uncertainty?
4. **What resources changed?** Include all participants' effort, full costs, quality, queues and sustainable workload.
5. **What happened to the benefit?** Separate released capacity, demonstrated redeployment, service or accessibility improvement, cost avoidance and cash.

AI productivity is not a choice between celebrating every saved minute and dismissing every pilot. The useful discipline is to make each claim no larger than its evidence. Start with the work, follow it beyond the draft, and account for the value actually realised.

Appendix. Three proposed reporting artefacts

Adapted from existing evaluation, cost-of-quality and benefits-management practice. These are proposed, not field-validated. Blank fields are deliberate: the paper supplies no invented pilot results.

A1. Workflow definition sheet

Field	What to record
Decision and ownership	Investment/operating decision, operational owner, benefit owner, recipients and workforce involvement.
Boundary	Trigger, acceptance point, downstream users, reopening/failure window, exclusions and complexity classes.

Field	What to record
Intervention and comparison	Model/version, data/retrieval, access, training, process redesign; business-as-usual tools and concurrent changes.
Outcomes and gates	Primary outcome, smallest worthwhile improvement, recipient-agreed quality criteria, independent review method and severe-error limits.
Evaluation and safeguards	Baseline, allocation/comparison, duration/power rationale, contamination/attrition plan, data minimisation, permitted access and retention.

A2. Operational evidence record

Field	What to record
Claim and provenance	Claim type, evidence design, period/version, sample, case mix, demand and uncertainty.
Production	Accepted completions per period; active effort across all contributors, including rejected work; staffing/paid hours; incomplete/abandoned work.
Quality and downstream work	First-pass acceptance, independent ratings, severe errors, later reopening/corrections and recipients' checking/rework effort.
Flow and resource use	End-to-end cycle time, backlog age, waits/handoffs; full resource cost per accepted outcome, with allocation assumptions.
People and capacity	Training/learning, accessibility and workload indicators; net capacity estimate; destination supported by evidence, or not established.

A3. Financial reconciliation note

Field	What to record
Cash/commercial result	Attributable realised spend reductions and incremental contribution margin; incremental cash costs; period and comparison.
Separate categories	Approved-spend avoidance, forecasts, opportunity-cost capacity and non-financial benefits, each labelled rather than added as cash.
Resource-efficiency result	Labour and other input per accepted outcome, valuation basis, volumes, one-off versus recurring costs and whole-life horizon.
Reconciliation	Overlapping hours/benefits removed; shared costs allocated once; observed losses separated from risk scenarios; uncertainty and sensitivity.
Accountability	Benefit owner, finance review, remaining assumptions, decision and reassessment triggers.

References

Sources were checked in November 2026. Dates in study descriptions refer to the study period or version, not current tool performance. Working papers and commercial sources are labelled.

American Society for Quality. (n.d.). *Cost of quality (COQ)*. Retrieved November 15, 2026, from https://asq.org/quality-resources/cost-of-quality [Evidence note: Prevention, appraisal, internal and external failure categories.]

Australian Centre for Evaluation. (2025). *Evaluation of a trial of generative artificial intelligence (Copilot) in the Department of the Treasury*. Department of the Treasury. https://evaluation.treasury.gov.au/publications/evaluation-generative-artificial-intelligence [Evidence note: Trial conducted in 2024. Official full report:

https://evaluation.treasury.gov.au/sites/evaluation.treasury.gov.au/files/2025-02/evaluation-generative-artificial-intelligence.pdf.]

Bick, A., Blandin, A., & Deming, D. (2025, November 13). The state of generative AI adoption in 2025. *On the Economy*. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/on-the-economy/2025/nov/state-generative-ai-adoption-2025 [Evidence note: 2025 adoption and time-saving analysis, which distinguishes reported savings and modelled aggregate implications.]

Bick, A., Blandin, A., Deming, D., & Schumacher, T. (2026, September 1). What work does generative AI do? *On the Economy*. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/on-the-economy/2026/sep/what-work-does-generative-ai-do [Evidence note: Authors' Federal Reserve summary of RPS task-level research, linking to working paper and data.]

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044 [Evidence note: Author-hosted final PDF: https://danielle.li/assets/docs/GenerativeAIatWork.pdf; published institutional abstract: https://www.gsb.stanford.edu/faculty-research/publications/generative-ai-work.]

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. https://doi.org/10.1257/mac.20180386 [Evidence note: Complementary-investment explanation, not a guarantee of future AI returns. Earlier NBER working paper (2018): https://www.nber.org/papers/w25148.]

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). *The GenAI divide: State of AI in business 2025* [Preliminary report]. MIT NANDA. https://21464181.fs1.hubspotusercontent-na1.net/hubfs/21464181/v0.1_State_of_AI_in_Business_2025_Report.pdf [Evidence note: July 2025 preliminary report, consulted in a mirrored PDF. Consulted only to assess and exclude population-level headline claims; institutional provenance/representative denominator remain unresolved.]

Counts, L. (2026, February 18). *AI promised to free up workers' time. UC Berkeley Haas researchers found the opposite*. Haas News, UC Berkeley Haas. https://newsroom.haas.berkeley.edu/ai-promised-to-free-up-workers-time-uc-berkeley-haas-researchers-found-the-opposite/ [Evidence note: Direct author interview and methodological account of the Ranganathan and Ye research.]

Cruces, G., Fernández Meijide, D., Galiani, S., Gálvez, R. H., & Lombardi, M. (2026). *Does generative AI narrow education-based productivity gaps? Evidence from a randomized experiment* (NBER Working Paper No. 34851). National Bureau of Economic Research. https://doi.org/10.3386/w34851 [Evidence note: Working paper, revised October 2026 (version used). Full text: https://www.nber.org/system/files/working_papers/w34851/w34851.pdf.]

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., & Lakhani, K. R. (2026). The cybernetic teammate: A field experiment on generative AI and teamwork. *Organization Science, 37*(4), 1217–1242. https://doi.org/10.1287/orsc.2025.20702 [Evidence note: Final publisher article rather than the earlier NBER abstract. Publisher page shows Published Online: 12 Jun 2026; issue dated July 2026.]

Dell'Acqua, F., McFowland, E., III, Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science, 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838 [Evidence note: HBS-hosted published PDF: https://www.hbs.edu/ris/Publication%20Files/dell-acqua-et-al-

2026-navigating-the-jagged-technological-frontier_5c589c8c-fbb5-458f-b285-c944746cd717.pdf.]

Digital Transformation Agency. (2024). *Evaluation of the whole-of-government trial of Microsoft 365 Copilot: Full report*. https://www.digital.gov.au/initiatives/copilot-trial/microsoft-365-copilot-evaluation-report-full [Evidence note: Australian whole-of-APS evaluation. Chapters used: productivity, https://www.digital.gov.au/initiatives/copilot-trial/microsoft-365-copilot-evaluation-report-full/productivity; methods appendix, https://www.digital.gov.au/initiatives/copilot-trial/microsoft-365-copilot-evaluation-report-full/appendix-b.]

Dillon, E. W., Jaffe, S., Immorlica, N., & Stanton, C. T. (2025). *Shifting work patterns with generative AI* (NBER Working Paper No. 33795). National Bureau of Economic Research. https://doi.org/10.3386/w33795 [Evidence note: Microsoft-linked research. Working paper, revised November 2025 (version used). Full text: https://www.nber.org/system/files/working_papers/w33795/w33795.pdf.]

DORA. (2026, January 5). *DORA's software delivery performance metrics*. https://dora.dev/guides/dora-metrics/ [Evidence note: Official guidance; Google-associated programme.]

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130 [Evidence note: Task-exposure methodology; technical potential rather than realised adoption or return.]

Forrester Consulting. (2024). *New technology: The projected Total Economic Impact of Microsoft 365 Copilot for SMB*. https://tei.forrester.com/go/microsoft/SMB365Copilot/ [Evidence note: Commissioned by Microsoft, October 2024. Composite, assumption-dependent financial projection.]

Forsgren, N., Storey, M.-A., Maddila, C., Zimmermann, T., Houck, B., & Butler, J. (2021). The SPACE of developer productivity. *Communications of the ACM, 64*(6), 46–53. https://doi.org/10.1145/3453928 [Evidence note: Developer framework, not business-user score validation. Original ACM Queue article: https://doi.org/10.1145/3454122.3454124.]

Gartner. (2026, November 4). *Gartner identifies three steps for CFOs to turn AI time savings into business value* [Press release]. https://www.gartner.com/en/newsroom/press-releases/2026-11-04-gartner-idnetifies-three-steps-for-cfos-to-turn-ai-time-saavings-into-business-value [Evidence note: Commercial analyst guidance, not a field-experiment effect estimate.]

Government Digital Service. (2025). *Microsoft 365 Copilot experiment: Cross-government findings report*. GOV.UK. https://www.gov.uk/government/publications/microsoft-365-copilot-experiment-cross-government-findings-report/microsoft-365-copilot-experiment-cross-government-findings-report-html [Evidence note: Official report on 2024 deployment.]

Government Finance Function & HM Treasury. (2025, November 24). *Government Efficiency Framework (GEF) summary, key definitions and criteria guide*. GOV.UK. https://www.gov.uk/government/publications/the-government-efficiency-framework/government-efficiency-framework-gef-summary-key-definitions-and-criteria-guide [Evidence note: Official cash/non-cash, net-cost and double-counting criteria.]

Government Project Delivery. (2026, July 1). *Chapter 19: Benefits management*. The Teal Book. https://projectdelivery.gov.uk/teal-book/home/part-e-planning-and-control/chapter-19-benefits-

management/ [Evidence note: Official practice guidance on ownership, baseline, counterfactual and whole-life realisation.]

Humlum, A., & Vestergaard, E. (2026). *Still waters, rapid currents: Early labor market transformation under generative AI* (NBER Working Paper No. 33777). National Bureau of Economic Research. https://doi.org/10.3386/w33777 [Evidence note: Working paper first issued May 2025, revised March 2026 (version used); previously circulated under a different title. Versioned full text: https://www.nber.org/system/files/working_papers/w33777/w33777.pdf.]

METR. (2025, July 10). *Measuring the impact of early-2025 AI on experienced open-source developer productivity*. https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/ [Evidence note: Early-2025 experienced-developer RCT. Primary research post with methods and linked data/paper.]

METR. (2026a, February 24). *We are changing our developer productivity experiment design*. https://metr.org/blog/2026-02-24-uplift-update/ [Evidence note: Experiment-design update. Primary research post with methods.]

METR. (2026b, May 11). *Measuring the self-reported impact of early-2026 AI on technical worker productivity*. https://metr.org/blog/2026-05-11-ai-usage-survey/ [Evidence note: Technical-worker self-report survey. Primary research post with methods.]

Microsoft. (2026, August 5). *Copilot Business Impact report*. Microsoft Learn. https://learn.microsoft.com/en-us/viva/insights/advanced/analyst/templates/copilot-business-impact [Evidence note: Vendor documentation warns that group differences are not tested for statistical significance and other causes can explain them. Prior-art source, not causal validation.]

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586 [Evidence note: Preregistered experiment.]

Office of the Australian Information Commissioner. (2024, October 15). *Workplace monitoring and surveillance*. https://www.oaic.gov.au/privacy/your-privacy-rights/surveillance-and-monitoring/workplace-monitoring-and-surveillance [Evidence note: Official general guidance; implementation requires specific advice.]

OpenAI. (2025, September 25). *Measuring the performance of our models on real-world tasks*. https://openai.com/index/gdpval/ [Evidence note: Initial GDPval description, methods, linked paper and limitations. Vendor-produced capability evidence.]

Ranganathan, A., & Ye, X. M. (2026, February 9). AI doesn't reduce work—It intensifies it. *Harvard Business Review*. https://hbr.org/2026/02/ai-doesnt-reduce-work-it-intensifies-it [Evidence note: Authors' report of in-progress research.]

U.S. Bureau of Labor Statistics. (n.d.-a). *Productivity home page*. Retrieved November 15, 2026, from https://www.bls.gov/mfp/ [Evidence note: Official definitions, not AI-effect estimates.]

U.S. Bureau of Labor Statistics. (n.d.-b). *What is productivity?* Retrieved November 15, 2026, from https://www.bls.gov/k12/productivity-101/content/what-is-productivity/home.htm [Evidence note: Official definitions, not AI-effect estimates.]