The claim

Somewhere between the vendor demonstration and next year's budget, a claim changes species. In the demo it was a capability; in the budget it is money — a productivity line, a headcount assumption, a margin promise. The decision is what evidence a board requires before that change of species is allowed.

Our position is strict: a productivity claim is only real when it survives translation into a profit-and-loss line — headcount, cycle time, error cost, revenue per employee — against a named baseline, with a named owner and a review date. Demos, benchmarks and vendor case studies are not that evidence: they show a task can be done, not that a cost line here will move. A board that budgets on capability evidence is not being rigorous with better tools; it is being imprecise with better vocabulary.

The evidence

Real evidence exists, and it is strong at the task level: a staggered field rollout across 5,179 customer-support agents, a randomised experiment on 453 professionals in Science, a controlled experiment on 95 developers. The averages hide the finding that matters — the gains concentrate among novices and fall to near nothing among the most experienced.

> EVIDENCE
>
> Experiments on generative AI report large task-level gains: 14 per cent more issues resolved per hour among 5,179 support agents (34 per cent for novices); 40 per cent less time and 18 per cent higher quality on writing tasks among 453 professionals; 55.8 per cent faster completion of a defined coding task among 95 developers.
>
> Source: Brynjolfsson, Li & Raymond, Quarterly Journal of Economics 140(2), 2025; Noy & Zhang, Science 381, 2023; Peng et al., arXiv:2302.06590, 2023 (working paper) · Geography: United States and remote study populations · Method: staggered field rollout and randomised experiments · SEG-4 · Caveat: task-level, population-specific, short-horizon; the Copilot study is not peer-reviewed; none measures a firm's P&L.

The fine print is the point. Each study measures output per unit of time on the treated task, in a specific population, over a short horizon. None measures what a board budgets on: payroll, end-to-end cycle time, error cost, or what the saved minutes became.

The task numbers cannot be pasted into a firm-level plan; the reason is arithmetic. Acemoglu's task-based analysis concludes that the implied gain in total factor productivity over ten years is modest. The structure, not the disputed inputs, is what transfers: firm-level gain equals task-level gain, times the share of work the task represents, times the share of cost that work represents. Every term is a fraction. A 40 per cent saving on a task that is 5 per cent of the work of 20 per cent of staff is not a productivity programme; it is a rounding error wearing one.

EVIDENCE

A task-based macroeconomic analysis aggregating measured AI task-level savings across the share of exposed work estimates total TFP gains over ten years of roughly 0.5 to 0.7 per cent — modest, and possibly overstated if hard-to-learn tasks yield less than the easy tasks studied.

Source: Acemoglu, "The Simple Macroeconomics of AI", Economic Policy 40(121), 2025 (NBER WP 32487, 2024) · Geography: United States, economy-level model · Method: task-based model calibrated on experimental estimates · SEG-4 · Caveat: model-based projection, contested inputs; other economists estimate larger effects; the aggregation arithmetic, not the point estimate, is what transfers.

Nor will official statistics referee the claim. Labour productivity is a sector-level residual, published in scheduled vintages and revised as source data arrive. The statistics answer a different question, later, for a different unit of analysis; they cannot attribute a gain to your deployment.

What none of these sources supplies — because it is firm judgment, not measurement — is the budget test: before an AI productivity claim enters a plan, it must name its P&L line, its dated baseline, its owner and its review date. A demo proves capability; a vendor case study proves a customer was willing to be quoted. Neither proves that your baseline moved. A claim that cannot name its line and baseline is procurement sentiment, and should be budgeted as a cost with an experiment attached, not as a saving.

Evidence cards

CLM-GENAI-TASK-EXPERIMENTS · SEG-4. Claim: experiments measure large task-level generative-AI gains — 14 per cent average throughput among 5,179 support agents (34 per cent for novices, minimal for experienced workers); 40 per cent less time and 18 per cent higher quality on writing tasks among 453 professionals; 55.8 per cent faster on a defined coding task among 95 developers. Context: Brynjolfsson, Li & Raymond, QJE 140(2), 2025; Noy & Zhang, Science 381, 2023; Peng et al., arXiv 2302.06590, 2023. Method: staggered field rollout; randomised experiments. Contradictory evidence: task-level, short-horizon, population-specific; the Copilot study is an unrefereed working paper. Causal confidence: experimental, in-setting. Transferability: bounded to comparable tasks and populations. Review date: 2026-08-02.

CLM-TASK-TO-FIRM-GAP · SEG-4. Claim: aggregating measured task-level savings across the exposed share of work yields modest economy-level effects — roughly 0.5 to 0.7 per cent total TFP gain over ten years — because firm-level gain is task gain multiplied by task share and cost share. Context: Acemoglu, Economic Policy 40(121), 2025; NBER WP 32487. Method: task-based model calibrated on experimental estimates. Contradictory evidence: inputs contested; other estimates run larger; projections are not measurements. Causal confidence: none claimed — model arithmetic. Transferability: the aggregation structure transfers; the point estimate does not. Review date: 2026-08-02.

CLM-PRODUCTIVITY-STATS-VINTAGE · SEG-5. Claim: official labour-productivity statistics are sector-level residuals published in scheduled vintages and revised as source data arrive — the BLS publishes a preliminary quarterly estimate and two scheduled revisions — and cannot attribute gains to a specific firm's deployment. Context: US Bureau of Labor Statistics productivity-programme documentation. Method: official statistical methodology. Contradictory evidence: none on the fact; the limitation is by design. Causal confidence: none

claimed — institutional fact. Transferability: analogous vintage-and-revision structures in other national statistical systems. Review date: 2026-08-02.

CLM-FIRM-PNL-TRANSLATION · SEG-1 (firm position). Claim: an AI productivity claim may enter a budget only when translated into a named P&L line against a named, dated baseline, with an owner, a review date and a kill threshold; demos, benchmarks and vendor case studies are capability evidence, not budget evidence. Context: firm operating practice in the corridor. Method: interpretation, not measurement; no client deployment statistics cited. Contradictory evidence: for small, cheap experiments, full P&L instrumentation may cost more than the experiment; the test is proportionate to the money at stake. Causal confidence: none claimed. Transferability: bounded — strongest where claims are large relative to the cost base and baselines are buildable. Review date: 2026-08-02.

The limits

Every imported productivity claim crosses the corridor with the study population's wage structure attached, and does not survive intact. The experiments price saved time in US wages. The same throughput gain in an Abidjan back office is the same operational fact and a different financial one: the saved hour is priced in a payroll several multiples lower, while the tool is priced in dollars or euros and does not scale down. The break-even moves, sometimes past reach; connectivity and power enter the cost line in ways the study firms never metered. The harder corridor problem is that the baseline often does not exist: cycle times unmeasured, error costs unpriced, throughput known by feel. An operating company that cannot state its cost per processed file cannot know whether AI improved it, but it can still be invoiced monthly. The costs arrive denominated and dated; the benefits arrive as narrative until someone builds the line.

Sources and limitations

Sources and limitations. The task-level findings rest on Brynjolfsson, Li & Raymond (QJE, 2025), Noy & Zhang (Science, 2023) and Peng et al. (arXiv working paper, 2023, flagged as unrefereed); the aggregation argument on Acemoglu (Economic Policy, 2025), whose contested estimates are used for their structure, not their precision; the statistical caveat on BLS documentation. All were verified against official and academic sources on 2026-08-02; no consultancy estimates are cited. The P&L translation test is the firm's position, graded as interpretation: not a validated instrument, with no client productivity figures published because none has passed our evidence-release gate. The corridor wage arithmetic is reasoning, not measured client data. This note is not valid as a forecast of any firm's savings, nor as a claim that AI productivity gains are illusory — only that they must be translated, baselined and owned before they are budgeted.

What to do on Monday

1. The CFO: refuse any AI line in the plan that cannot name its P&L line, its dated baseline and its owner in one sentence. Fund it as a measured experiment or not at all.

2. The programme sponsor: name the study population the use case resembles, and run the wage arithmetic — the saving priced in our payroll, the tool in the vendor's currency. Gains concentrate among novices on instrumented, high-volume tasks; the cited evidence is not about senior judgment work.

3. The business owner: agree a date, a number and a kill threshold before deployment. If the line has not moved by the review date, the default is stop.

4. The financial controller: reconcile the gross claim against the full cost line. Subscriptions, integration, connectivity, review time, rework. A benefit tracked gross of its costs is an advertisement.