Research Technical Report FRITS-TR-2026-08 11 August 2026 Updated 11 August 2026 PRODUCTIVITY
ADOPTION ECONOMICS SKILLS AGENTS MEASUREMENT

# Nine in Ten Firms Report No Productivity Effect From AI — and Use It 1.5 Hours a Week: The Argument Is Not About Whether AI Works

Frits Lyneborg
FRITS AI ApS

ABSTRACT

Nine in ten firms report that AI has changed nothing about their productivity. The same survey of nearly 6,000 executives found that the ones who use AI use it about 1.5 hours a week. Two controlled results sharpen the picture: consultants using AI on tasks outside its capability frontier performed 19 percentage points worse than colleagues using none, and experienced developers were measured 19% slower with AI while believing they were 20% faster. The deciding variable is the operator, not the model, and the sign of the effect flips with the task. Where the specification arrives with the task, AI compresses the skill distribution and novices gain most — 34% for new support agents, nothing for experts. Where the operator must supply the specification, it amplifies instead. Both camps are right about different task classes, and averaging them produces a number describing nobody. We instrument one case from the upper tail — 5,869 commits, 145 working days, 27 languages, one operator — and find its schedule sits so far outside COCOMO II's compression limits that in 2021 the outcome was not purchasable at any headcount.

## 1. The argument is not about whether AI works

Nine in ten firms say AI has changed nothing about their productivity or employment. The same February 2026 survey — nearly 6,000 senior executives across US, UK, German and Australian firms — also asked how much those executives actually use it. **The answer was about 1.5 hours a week** [1 (#ref-1)]. Both facts come from one dataset. Read together they are not a paradox; they are the finding.

Meanwhile a small population of practitioners reports individuals producing the output of dozens of people, and capital is being deployed on that reading, with hyperscaler AI infrastructure spending projected above $500 billion for 2026 alone [4 (#ref-4)]. Both groups hold tools that are, at the level of the model, essentially identical.

The public debate treats this as a dispute about whether AI works. The question has no answer as posed, because it is missing its subject. The same tools produce a 34% gain for one worker and a 19-percentage-point loss for another, and what differs is not the technology.

The variable that decides the outcome is the operator. And the direction of the effect flips depending on whether the task is bounded. Where the task arrives already specified, AI *compresses* the skill distribution and the least skilled gain most. Where the operator must supply the specification, the decomposition and the verification, AI *amplifies* the distribution instead — and the same tool that multiplies a competent operator sends an incompetent one backwards, measurably worse than using no AI at all.

Both camps are reporting honestly, on different task classes. The aggregate statistics meant to settle the argument average the two together.

## 2. Outcomes are bimodal, and the disappointed majority is barely using the tool

The evidence for disappointment is substantial and well-sampled. S&P Global found 42% of over 1,000 enterprises had abandoned most of their AI initiatives, up from 17% a year earlier [3 (#ref-3)]. Pew found 40% of Americans expect AI's effect on society to be

negative against 16% expecting it positive [2 (#ref-2)]. The mood has a name in the trade press — AI fatigue — and a coherent thesis: the technology was marketed aggressively before it was reliably useful.

The 1.5-hours-a-week figure is what that thesis has to survive, and we do not think it does. Nine in ten firms reporting no measurable impact is not a finding about what AI can do. It is a finding about what ninety minutes a week of unpractised use does, which is roughly what anyone would predict. The same survey reports those firms nonetheless expect sizeable gains over the next three years [1 (#ref-1)]. The expectation is intact, the practice is thin, and the distance between them is being logged as a failure of the technology.

The shape underneath is bimodal, which is what makes a single average the wrong instrument. Google's DORA programme surveyed nearly 5,000 technology professionals and found 90% using AI at work, more than 80% believing it had raised their productivity, and — in the same dataset — a *positive* relationship between adoption and delivery throughput alongside a *negative* one with delivery stability [7 (#ref-7)]. Their summary is unambiguous: AI does not fix a team, it amplifies what is already there.

When one group's output rises sharply and another's falls, the mean sits in the empty space between them and describes neither. The widely circulated claim that 95% of enterprise AI pilots produce no measurable return [5 (#ref-5)] makes the point in spite of itself: its own framing is that 5% captured value and 95% did not. Those are two populations, not one disappointing mean. (That report deserves its criticism — preliminary, 52 interviews, counting as failure any pilot short of full production [6 (#ref-6)] — so we use it only for the shape.) One further clue that the variable is competence: in the same report, deployments built with an external specialist succeeded roughly twice as often as internal ones, about 67% against a third [5 (#ref-5)]. The tools are identical in both arms.

## 3. The sign of the skill effect flips with task boundedness

The literature looks contradictory until the studies are sorted by one property: whether the specification arrives with the task, or has to be produced by the operator.

**On bounded tasks, AI compresses the skill distribution.** Brynjolfsson, Li and Raymond studied 5,179 customer-support agents and found productivity up 14% on average — but the average concealed the result that matters. **Novices gained 34%, experts close to nothing** [8 (#ref-8)]. The mechanism they propose is that the model disseminates the tacit best practice of the strongest workers to everyone else. Noy and Zhang found the same shape on 453 professionals doing mid-level writing: time down 40%, quality up 18%, inequality between workers *decreased* [9 (#ref-9)]. In both settings the task arrives pre-specified, and the model supplies the execution that used to separate the strong worker from the weak one.

**On unbounded tasks, AI amplifies the distribution instead.** Dell'Acqua and colleagues ran a field experiment with 758 Boston Consulting Group consultants and found both signs in one study. Inside the model's capability frontier, AI raised output quality by more than 40% and cut task time by more than 25%. Outside it — on tasks that looked equally difficult but that the model handled badly — **consultants using AI were 19 percentage points less likely to produce a correct solution than colleagues working without it** [10 (#ref-10)]. The frontier is jagged and unmarked, and knowing which side a task falls on is not a property of the tool.

METR's randomised trial makes the same point at the expert end. Sixteen experienced open-source developers, in repositories they had contributed to for years, completed 246 real issues with and without AI. They were **19% slower with AI** — while estimating afterwards that it had made them 20% faster, having predicted 24% beforehand [11 (#ref-11)]. On large, ambiguous work, competence with the tool is not conferred by competence at the job.

| | Bounded task | Unbounded task |
|---|---|---|
| Who supplies the specification | The task itself | The operator |

| | Bounded task | Unbounded task |
|---|---|---|
| Who supplies verification | Existing process | The operator |
| Effect on the skill distribution | **Compresses** | **Amplifies** |
| Evidence | +34% novice vs ≈0 expert [8 (#ref-8)]; inequality fell [9 (#ref-9)] | −19 pp outside the frontier [10 (#ref-10)]; −19% for experts [11 (#ref-11)] |
| What the debate concludes from it | "AI democratises expertise" | "AI does not deliver" |

The mechanism is straightforward once the columns sit side by side. A bounded task hands the operator a specification and an acceptance test for free. An unbounded task hands over neither, so the operator must supply the specification, the decomposition into steps the model can complete, the judgement of which side of the frontier each step sits on, the verification, and the recovery path when it fails.

These are the classic engineering capacities, and they have always varied enormously between practitioners — the oldest measurements put the ratio between best and worst professional programmers at roughly 20:1, and even conservative readings of that contested study leave a difference above tenfold [12 (#ref-12)]. AI does not shrink that spread on unbounded work. It multiplies whatever the operator brings. Any study asking "what did AI do for this organisation?" without separating the task classes is averaging two effects that point in opposite directions.

## 4. Why neither camp can see the other

**Neither side's testimony is calibrated.** METR's developers believed they were 20% faster while being 19% slower [11 (#ref-11)]. DORA's respondents report over 80% believing AI raised their productivity, in a dataset simultaneously showing delivery stability getting worse [7 (#ref-7)]. In the other direction, Upwork found 96% of C-suite leaders expecting AI to raise productivity while 77% of employees using it said it had

*added* to their workload — and **47% said they did not know how to achieve the gains their employers expected** [13 (#ref-13)]. That is the competency gap stated plainly by the people inside it.

**The competent case is a tail that averages away.** Adoption is nearly universal, but competence accumulates slowly: Anthropic's Economic Index finds users with six or more months of tenure show a 10% higher success rate, an advantage that persists after controlling for task type [14 (#ref-14)]. A skill that takes months to build, in a population that mostly adopted recently and uses it 1.5 hours a week [1 (#ref-1)], produces exactly the distribution observed — a large majority near the start of the curve and a thin tail far along it. Sample that population and you measure the majority.

**AI implementation is a credence good.** Darby and Karni introduced the term in 1973 for a service whose quality the buyer cannot evaluate before purchase and cannot reliably verify afterwards, because judging the work requires the very expertise being bought. The canonical example is car repair [15 (#ref-15)]. AI implementation sits squarely in the class: a buyer cannot tell in advance whether a provider can genuinely make their workflow ten times more productive, and afterwards cannot separate a good outcome from a lucky one.

Such markets have a known failure mode, inherited from Akerlof's analysis of asymmetric quality information [16 (#ref-16)]. Because any provider can *state* the competence and no buyer can check it, buyers rationally discount every claim by the same factor. The discount falls equally on the genuine and the fraudulent, so high competence cannot command its value.

This reframes the most frustrating part of the phenomenon. When an experienced investor or procurement committee hears a claim of tenfold output and does not believe it, **that is correct behaviour under information asymmetry**, not a failure of imagination. Argument cannot fix it, because the problem is not that the claim is unpersuasive. It is that persuasiveness and truth are weakly correlated in this market, and the buyer knows it.

## 5. A measured case from the upper tail (n = 1)

Population surveys cannot see the tail, so we instrument one directly. This is a single, self-selected observation authored by the operator it describes — an existence proof about the tail's shape and nothing more. Section 8 states the ways it could mislead.

### 5.1 The artefact

GDPRchat is a production European AI assistant with paying users, built and operated by FRITS AI ApS. All figures are counted from the git history and the local agent-session store on 10 August 2026.

| Measurement | Value |
|---|---|
| Period | 1 March – 10 August 2026 (163 calendar days) |
| Days with commits | 145 |
| Commits on `main` | 5,869 |
| Merge commits (integrated feature branches) | 985 |
| Distinct files touched over the history | 9,106 |
| Lines added / deleted in application source | 691,168 / 328,325 |
| HTTP API routes | 283 |
| Database migrations | 177 |
| AI tool modules | 37 |
| Operational scripts | 234 |
| Automated test files | 178 |
| User-facing locales | 27 |
| Recorded AI agent sessions (from June 2026 only) | 709, across 224 distinct checkouts |
| Agent transcript events | 188,034 |

The breadth matters more than any single count. The same operation also produced and runs a data-protection and compliance corpus; an automated programme scoring candidate language models on routing fidelity, per-language non-regression,

performance, cost and values behaviour before any may serve users; a weekly market-watch duty filing change proposals a human approves; a support system that clusters incoming reports, fixes them in parallel isolated environments, deploys, and answers each reporter in their own language; a fifty-article knowledge corpus embedded for retrieval; a published research programme; native iOS and Android releases; and a public marketing site with automated consistency guards. In a conventional organisation these are separate departments.

The production method is a fleet of AI coding agents run concurrently, each in an isolated git worktree so a dozen tasks proceed without colliding; a machine-readable operating constitution every agent reads before acting; blocking automated gates so no agent's work reaches production unverified; and a human contribution of specification, arbitration, prioritisation and judgement, applied at roughly 40 commits per active day.

### 5.2 What this would have taken conventionally: three terms, not one

The natural question is "how many developers would this have taken", and the natural instrument is lines of code. Both are wrong. **Building the tools is the smallest of three terms**, and only the first is measurable by counting anything.

**Term A — building the tools.** Industry measurement puts sustained professional output at roughly 325–750 lines of delivered, tested production code per developer-month [18 (#ref-18), 20 (#ref-20)]. The hand-equivalent code here totals 316,281 lines (application source 239,843, tests 29,779, operational scripts 40,999, SQL migrations 5,660), excluding 90,231 lines of machine-translated locale data as vendor work and counting 16,503 lines of documentation and compliance prose separately under Term C.

A lines-per-developer rate assumes effort scales linearly with size. **COCOMO II** does not: effort = 2.94 × KSLOC^E with E = 1.096, so doubling the size costs 2.14 times the effort [22 (#ref-22)]. AI-generated code is also more verbose than what a disciplined engineer writes, so the second column assumes a careful team would deliver the same functionality in half the lines.

| COCOMO II, nominal ratings | Undiscounted (316.3 KSLOC) | Verbosity-halved (158.1 KSLOC) |
|---|---|---|
| Effort | **1,616 person-months** (135 person-years) | **756 person-months** (63 person-years) |
| Nominal schedule | 38.2 months | 30.0 months |
| Average team over that schedule | 42 people | 25 people |
| People needed to fit it into 5.36 months | **301** | **141** |
| Actual delivery as a fraction of nominal schedule | **14%** | **18%** |

The last row is the result we did not expect, and it is more informative than any headcount. COCOMO II's most aggressive compression rating describes finishing in 75% of the nominal schedule at a 1.43× effort penalty; a published extension reaches 50% compression at 1.51× [22 (#ref-22)]. This delivery sits at **14–18% of nominal**, outside the model's domain by a factor of three to five. The honest statement is not "it would have taken *N* people". It is that **in 2021 this schedule was not purchasable at any headcount.** You could have bought a 25-person team for two and a half years. You could not have bought five months.

COCOMO's calibration draws on an era of heavier process, so it is generous to us, and a strong team with favourable ratings could plausibly halve the effort again, putting a floor near 70 people. **Term A is therefore on the order of 70 to 300 person-equivalents, centred near 140.**

**Term B — the coordination and authorisation never incurred.** Before a line is written, somebody must recognise the need, write a business case, obtain budget, hire, onboard and manage. Then the work carries Brooks's coordination overhead, growing as the square of headcount [20 (#ref-20)] — 25 people is 300 pairwise channels, 42 people is 861, 141 people is 9,870 — and delivery risk: across 1,471 IT projects, Flyvbjerg and Budzier measured an average cost overrun of 27% and a fat tail in which **one in six overran cost by 200%** [21 (#ref-21)].

But the largest component of Term B is a **zero**. Most of the scope in Section 5.1 would never have been authorised. No 2021 budget committee commissions an automated programme scoring candidate language models on their values behaviour before allowing them to answer users; the category did not exist. Term B is not a multiplier on Term A. It is a gate most of Term A never passes.

**Term C — the work the tools then perform, indefinitely.** Term A is one-off. Term C recurs daily at zero marginal headcount, and grows each time another tool is added.

| Standing function | Counted volume | Conventional equivalent |
|---|---|---|
| Localisation, 27 languages | **540,826 translated words**, re-synchronised every release | 180–270 translator-days for the first pass alone, at 2,000–3,000 words per translator-day [23 (#ref-23)] — **1.7–2.5 full-time translators**, plus a manager |
| Documentation, knowledge base, legal and compliance prose | **240,950 words** | 241–482 writer-days at 500–1,000 finished words per day [23 (#ref-23)] — **2.2–4.5 technical writers** |
| Round-the-clock production operations | 47 unattended nightly tasks, 93 operational interfaces, automated backup, monitoring, blue-green deployment | **4.2 FTE minimum** for any single-person 24/7 rota, before holiday and escalation cover |
| Multilingual first-line support | Reports clustered, fixed, deployed, answered in each reporter's own language | Not rated — 27-language coverage has no small-team equivalent |
| Model evaluation and weekly market watch | Every candidate model gated on five dimensions | Not rated — no 2021 analogue exists |

The three rateable rows come to roughly **8 to 12 standing full-time roles**, permanently, excluding the two we could not rate — one of which is the largest, because 27-language human support does not scale down.

**The countable part therefore lands at roughly 80 to 310 full-time roles in 2021 terms, centred near 150**, delivered instead by one operator in 5.36 months. Every number is a floor: Term A excludes the mobile, compliance and research scope that is not source

code, Term C excludes its two largest rows, and Term B is excluded entirely. Take the most pessimistic reading available — Term A's floor of 70, halved again for five years of framework progress, Terms B and C set to zero — and the residual is still around **thirty-five times** one conventional operator, against a disappointed majority reporting no measurable effect at all [1 (#ref-1)].

The decomposition also relocates the competency. The leverage is not writing code quickly. It is seeing a need, deciding it is worth closing, building the tool that closes it, and putting that tool into daily operation — without a business case, a hiring round, or anyone's permission. Term A is a skill. Terms B and C are a *position*, and it is the position that produces the multiple.

### 5.3 Two days to a social platform

The clearest single illustration is not a productivity number. On 24–25 June 2026 the operation built a complete social platform — profiles, posts with sanitised markdown, threaded comments, ranked voting, image uploads, notification and unread-state tracking, native push seams, on-the-fly translation of user content with streaming and caching, an administrative review surface — reaching 54 files and 6,533 lines inside 113 commits across two days.

It was built as an *experiment*, to find out whether that shape of thing belonged in the product at all. Six weeks later it was repurposed into a private one-to-one support-ticket system where every case is visible only to its reporter and the team.

At the industry baseline, 6,533 lines is eight to twenty developer-months, and it undercounts the real change because it excludes the shared components, schema and translations the feature also touched. That is not the point either. The point is what it was *for*. In 2021, building a Reddit-shaped platform to decide whether you want one is not a decision anybody makes; you hold a meeting and decide from opinions, because the experiment costs a team-quarter. When it costs two days, you build it to find out.

## 6. The largest effect is on what gets attempted, and no survey can see it

Competent AI use does three things, measured in inverse proportion to their size. Known work gets done faster — the part everybody measures. The tools built along the way keep working afterwards, without headcount. And **the cost of finding out collapses**, which changes which questions get answered by building rather than arguing. A doubt that would have been settled by the most senior opinion in the room gets settled by evidence instead. The 985 merged feature branches and 328,325 deleted lines are the signature: work built, evaluated, and discarded cheaply.

This is why productivity instruments miss the largest effect. A survey asks what happened to the output of work the organisation had already decided to do. It cannot ask about the work the organisation would now dare to attempt, because in the counterfactual that work does not exist to be asked about.

It also explains an asymmetry in how the two camps talk. The disappointed majority describes AI in terms of tasks: it drafts my emails, it summarises my documents, it did not save me time. The other camp describes it in terms of things that now exist which otherwise would not. Those are reports from opposite sides of the boundedness split, and the second is invisible to any instrument pointed at the first.

## 7. What to do

**Buy on liability, not on evidence.** Dulleck, Kerschbamer and Sutter ran the largest controlled test of credence-goods markets we know of, 936 participants, measuring which institutions restore efficiency when buyers cannot verify quality. The result is sharp: **liability has a crucial effect; verifiability has at best a minor one; reputation has little influence; and competition drives prices down without improving efficiency as long as liability is absent** [17 (#ref-17)]. Certifications, case studies and competitive tendering are therefore weak instruments. What works is making the provider bear the consequence — payment contingent on a delivered, pre-agreed result, on your own real workflow. That also answers §4's impasse: the buyer's rational discount is overcome not by better argument but by an instrument that makes argument unnecessary.

**Make the unit of production a pair, not a person.** The operator supplies decomposition, frontier judgement, verification and recovery; the domain owner supplies the specification and the acceptance test — exactly what an unbounded task is missing, and what an outside implementer cannot invent. That is why "give everyone a chatbot licence" and "hire an AI consultant" both underperform, and why DORA's capability model puts user-centric focus among the capabilities determining whether AI helps at all [19 (#ref-19)].

**Budget the learning curve, and measure the task classes apart.** Six months of tenure is worth a 10% success-rate advantage [14 (#ref-14)], so evaluating AI on a 90-day pilot at 1.5 hours a week measures the bottom of the curve and calls it the ceiling. And because bounded and unbounded work respond with opposite signs, one blended number guarantees an uninformative result.

The competence that decides the outcome is specific, learnable, slow to acquire and unevenly distributed. That is the definition of a skilled trade. Every serious tool in industrial history arrived with a profession attached, and the assumption that this one would not is the assumption that generated the disappointment.

## 8. Limitations

This report is self-published and has not been peer-reviewed.

**The central claim is a synthesis, not a new experiment.** The task-boundedness flip is our reading across five studies that were not designed to test it and that differ in domain, population, model generation and outcome measure. No single study manipulates boundedness directly, and the experiment that would settle it has not to our knowledge been run. "Bounded" and "unbounded" are also a spectrum we have treated as a dichotomy, with no operational measure for placing a given task.

**The case study is n = 1, non-randomised, unblinded, and authored by its own subject.** It has no control condition and no independent quality measure, and its metrics are counts of activity, which correlate with delivered value only loosely. It supports the claim "the upper tail exists and looks like this" and no claim of the form

“AI produces an N× improvement in general”. Survivorship runs through the whole argument: we observe successful high-competence operators, not the ones who believed themselves competent and failed.

**The Section 5.2 estimates are model outputs, and every input is contestable.** COCOMO II is calibrated on an era of heavier process and takes size in lines of code, a discredited proxy; we halve the size for AI verbosity on judgement rather than measurement; and a conventional team would have produced a differently shaped artefact rather than the same one slowly. The schedule result is the most robust part, depending only on the model’s own compression limits and sitting a factor of three to five outside them, but it remains one model’s opinion about a counterfactual nobody ran. Term B is deliberately not quantified, so a reader is entitled to treat an unpriced term as unproven, and the Term C throughputs are industry practice with no authoritative standard [23 (#ref-23)].

**The credence-goods framing is applied theory, and everything here is dated by model generation.** Darby and Karni’s category and the Dulleck–Kerschbamer–Sutter results [15 (#ref-15), 17 (#ref-17)] were established in other markets, and nobody has tested that the liability result transfers. The METR and BCG results were measured on early-2025 and 2023-era models, and a moving capability frontier changes how much of the operator’s judgement is still load-bearing.

## 9. Conclusion

The question “does AI deliver?” has no answer because it is missing its subject. The same tools produce a 34% gain for a novice on a specified task and a 19-percentage-point loss for a professional on an unspecified one.

That leaves two camps talking past each other in a way argument cannot resolve. One reports truthfully from the tail of a distribution that surveys are built not to over-weight. The other reports truthfully from a population that uses the tool about 1.5 hours a week and mostly does not yet know how to get the gains it was promised — 47% say so directly [13 (#ref-13)]. Neither side’s testimony is calibrated, and the competence at issue is a credence good, so the disbelief is rational rather than obtuse.

The deepest part of the gap is not about speed. Where the competence exists, the cost of finding out collapses, and organisations start settling questions by building rather than arguing — an effect that shows up as things attempted rather than tasks completed, and therefore appears in no productivity survey.

The gap that decides this entire question is not between organisations that bought AI and organisations that did not. It is between those that have someone who knows how to use it and those that assumed everyone already did.

Corrections and prior art pointers are welcome: contact (/contact/).

## References


1. Yotzov, I., Barrero, J. M., Bloom, N., Bunn, P., Davis, S. J., Foster, K. M., Jalca, A., Meyer, B. H., Mizen, P., Navarrete, M. A., Smietanka, P., Thwaites, G., Wang, B. Z. — *Firm Data on AI.* NBER Working Paper 34836, February 2026 (rev. March 2026). https://www.nber.org/papers/w34836 (https://www.nber.org/papers/w34836)
2. Pew Research Center — *Key findings about how Americans view artificial intelligence.* 12 March 2026. https://www.pewresearch.org/short-reads/2026/03/12/key-findings-about-how-americans-view-artificial-intelligence/ (https://www.pewresearch.org/short-reads/2026/03/12/key-findings-about-how-americans-view-artificial-intelligence/)
3. S&P Global Market Intelligence — *Voice of the Enterprise: AI & Machine Learning.* 2025 survey of over 1,000 enterprises in North America and Europe. Reported in CIO Dive*, AI project failure rates are on the rise.* https://www.ciodive.com/news/AI-project-fail-data-SPGlobal/742590/ (https://www.ciodive.com/news/AI-project-fail-data-SPGlobal/742590/)
4. Goldman Sachs — *Why AI companies may invest more than $500 billion in 2026.* Goldman Sachs Insights, 2026. https://www.goldmansachs.com/insights/articles/why-ai-companies-may-invest-more-than-500-billion-in-2026 (https://www.goldmansachs.com/insights/articles/why-ai-companies-may-invest-more-than-500-billion-in-2026)

5. MIT Project NANDA — *The GenAI Divide: State of AI in Business 2025.* MIT Media Lab, 2025. Preliminary, not peer-reviewed.
6. Shimoni, A. — *MIT's 95% AI failure rate is wrong.* 2025. https://arnon.dk/mits-95-ai-failure-rate-is-wrong/ (https://arnon.dk/mits-95-ai-failure-rate-is-wrong/)
7. DORA (Google Cloud) — *State of AI-assisted Software Development 2025.* 2025. https://dora.dev/dora-report-2025/ (https://dora.dev/dora-report-2025/)
8. Brynjolfsson, E., Li, D., Raymond, L. — *Generative AI at Work.* Quarterly Journal of Economics 140(2), 2025, pp. 889–942. NBER Working Paper 31161.
9. Noy, S., Zhang, W. — *Experimental evidence on the productivity effects of generative artificial intelligence.* Science 381(6654), 2023, pp. 187–192. doi:10.1126/science.adh2586
10. Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K. R. — *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.* Organization Science, 2025. Harvard Business School Working Paper 24-013.
11. Becker, J., et al. (METR) — *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.* 2025. arXiv:2507.09089.
12. Sackman, H., Erikson, W. J., Grant, E. E. — *Exploratory experimental studies comparing online and offline programming performance.* Communications of the ACM 11(1), 1968, pp. 3–11.
13. Upwork Research Institute / Workplace Intelligence — *From Burnout to Balance: AI-Enhanced Work Models for the Future.* Survey of 2,500 workers and executives, July 2024. https://www.upwork.com/research/ai-enhanced-work-models (https://www.upwork.com/research/ai-enhanced-work-models)
14. Anthropic — *Anthropic Economic Index report: Learning curves.* March 2026. https://www.anthropic.com/research/economic-index-march-2026-report (https://www.anthropic.com/research/economic-index-march-2026-report)
15. Darby, M. R., Karni, E. — *Free Competition and the Optimal Amount of Fraud.* Journal of Law and Economics 16(1), 1973, pp. 67–88.

16. Akerlof, G. A. — *The Market for "Lemons": Quality Uncertainty and the Market Mechanism.* Quarterly Journal of Economics 84(3), 1970, pp. 488–500.
17. Dulleck, U., Kerschbamer, R., Sutter, M. — *The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition.* American Economic Review 101(2), 2011, pp. 526–555.
18. Jones, C. — *Applied Software Measurement.* McGraw-Hill, 3rd ed., 2008; McConnell, S. — *Software Estimation: Demystifying the Black Art.* Microsoft Press, 2006. Ranges as compiled in *How much code can a coder code?*, successfulsoftware.net, 2017. https://successfulsoftware.net/2017/02/10/how-much-code-can-a-coder-code/ (https://successfulsoftware.net/2017/02/10/how-much-code-can-a-coder-code/)
19. DORA (Google Cloud) — *2025 DORA AI Capabilities Model.* 2025. https://dora.dev/ai/capabilities-model/report/ (https://dora.dev/ai/capabilities-model/report/)
20. Brooks, F. P. — *The Mythical Man-Month: Essays on Software Engineering.* Addison-Wesley, 1975 (anniversary ed. 1995).
21. Flyvbjerg, B., Budzier, A. — *Why Your IT Project May Be Riskier Than You Think.* Harvard Business Review 89(9), September 2011, pp. 23–25. Sample of 1,471 IT projects. arXiv:1304.0265
22. Boehm, B. W., et al. — *COCOMO II Model Definition Manual, version 2.1.* USC Center for Software Engineering, 2000. Post-architecture effort and schedule equations and the SCED compression ratings. https://www.rose-hulman.edu/class/csse/csse372/201410/Homework/CII_modelman2000.pdf (https://www.rose-hulman.edu/class/csse/csse372/201410/Homework/CII_modelman2000.pdf) — schedule compression beyond the model's Very Low rating is examined in *Effect of Schedule Compression on Project Effort in COCOMO II Model for Highly Compressed Schedule Ratings*, University of Southampton, https://eprints.soton.ac.uk/266925/ (https://eprints.soton.ac.uk/266925/)
23. Throughput figures are industry practice rather than a formal standard, and the sources state their own caveats. Translation, 2,000–3,000 words per translator-day: Pangeanic, *How many words does a professional translator translate per day?*, https://blog.pangeanic.com/how-many-words-does-a-professional-translator-translate-per-day (https://blog.pangeanic.com/how-many-words-does-a-professional-translator-

translate-per-day) and PacTranz, *Expected translation times by professional translators*, https://www.pactranz.com/translation-times/ (https://www.pactranz.com/translation-times/) — technical writing, 500–1,000 finished words per day: *Productivity in Technical Writing*, thenewtechnicalwriter.wordpress.com, 2018, https://thenewtechnicalwriter.wordpress.com/2018/08/31/productivity-in-technical-writing/ (https://thenewtechnicalwriter.wordpress.com/2018/08/31/productivity-in-technical-writing/)

---

## How to cite

Frits Lyneborg (2026). Nine in Ten Firms Report No Productivity Effect From AI — and Use It 1.5 Hours a Week: The Argument Is Not About Whether AI Works. FRITS AI ApS, Technical Report FRITS-TR-2026-08. https://frits.ai/research/bimodal-ai-outcomes-and-the-operator-skill-gap/ doi:10.5281/zenodo.21887994.

```
@techreport{lyneborg2026nine,
  title       = {Nine in Ten Firms Report No Productivity Effect From AI — and Use It 1.5 Hours
  author      = {Lyneborg, Frits},
  institution = {FRITS AI ApS},
  number      = {FRITS-TR-2026-08},
  year        = {2026},
  month       = {aug},
  doi         = {10.5281/zenodo.21887994},
  url         = {https://frits.ai/research/bimodal-ai-outcomes-and-the-operator-skill-gap/}
}
```