SolvxAI
Data to Decision Powered by AI

SOLVXAI WHITE PAPER SERIES | NO. 04 | THE BUSINESS CASE

Hours, Not Barrels

The honest value case for AI in technical work, and why the defensible number is smaller than the one you were quoted

Published by the SolvxAI Research Team
July 2026 · Edition 1.0
Contact: sd@nexaconsultinc.com

IN BRIEF

Most business cases for AI in upstream rest on statistics that do not survive being traced to their source. This paper begins by publishing five of them as refuted, then assembles what the controlled evidence actually shows. That evidence is real, substantial, and points in an awkward direction: measured gains are largest for less experienced workers and shrink, vanish, or reverse as expertise rises — a pattern that appears across four independent studies with different populations, designs and funders. Since upstream engineering is expert work, that finding governs the value case rather than complicating it. We argue the credible claim is a specific one about hours, made under stated conditions, and we set out what would have to be true for it to hold.

Contents

1. Executive summary

The value is real. The number you were quoted is probably not.

We set out to build a defensible business case for AI in upstream technical work. The first thing that happened was that five of the statistics normally used to build one failed verification. We publish them as refuted in section 2, because a business case resting on them will not survive contact with a sceptical finance function.

What the controlled evidence does support is narrower and more useful:

- **Measured gains on professional work are substantial but heterogeneous.** Randomised and quasi-experimental studies report a 40% reduction in time on writing tasks [4], a 26% increase in completed developer tasks [1], a 15% increase in support issues resolved per hour [3], and 12% more consulting tasks completed with quality gains [2].
- **The effect reverses with expertise.** Support agents at the top of the skill distribution showed *small declines in quality* [3]. Developers with long tenure showed no significant effect [1]. On a task outside AI capability, consultants using it were **19 percentage points less likely** to be correct [2]. And experienced open-source developers working in repositories they knew well were measured **19% slower** with AI — while believing they had been 20% faster [5].
- **At the level of a whole labour market, the effect has not yet appeared.** Danish administrative records covering 25,000 workers across 7,000 workplaces found precise null effects on earnings and hours, ruling out average effects larger than 2% two years after the technology became widely available — including among daily users who self-reported gains. [8]

None of this says the technology does not work. It says the gains concentrate where the work is routine and the worker is inexperienced, which is precisely the opposite of where an operator’s expensive hours are spent. That is the problem a serious business case has to solve, and it is the one this paper tries to state honestly.

2. Five numbers we could not stand behind

Each of the following is in wide circulation, appears in vendor material and trade press, and is frequently used to size an AI business case. We attempted to trace each to its original publication. These are the results.

The claim	What we found when we traced it
New hires take **28 weeks** to reach full productivity (attributed to a 2014 UK study)	The phrase does not occur in that report. Its actual figures are 15 weeks for a same-sector hire, 32 cross-sector, 40 for a graduate, and a full year from unemployment. There is no single average of 28 weeks anywhere in the document. **The number is a fabricated aggregate.** [10]
Knowledge workers spend **2.5 hours a day** searching for information	Traces to a 2001 analyst briefing paper which described the figure as a general estimate requiring adjustment per enterprise — it was never a measurement. The primary document is not publicly retrievable. **Unverifiable.**
70% of digital transformations fail	The trail runs through a 2019 management article to a 2018 vendor-authored column, and ultimately to a 1993 book about business process reengineering whose authors called their own figure an unscientific estimate. No primary study exists at the end of it, and no oil-and-gas-specific version has any traceable original. **Unsupported.**
95% of generative AI pilots fail	Originates in a non-peer-reviewed 2025 working paper based on 153 survey responses and 52 interviews. The primary document could not be retrieved from an institutional source. The underlying claim concerns absence of measured profit-and-loss return, not failure. **Not usable as stated.**
Engineers spend **50%+** of their time on data	The nearest real figure is 45%, self-assigned by respondents in a vendor-run survey, with a base of 1,099 — and it describes data professionals, not engineers. For software developers the published literature spans **9% to 61%** depending on the study. **No point estimate is defensible.** [9]

WHY THIS SECTION IS FIRST

A business case is an argument made to people whose job is to find the weak joint in it. Every number above is one an informed sceptic can dismantle in an afternoon, and dismantling one of them discredits the rest of the case by association. We would rather open by removing them.

One further caution for anyone assembling evidence in this area: a widely cited 2024 study reporting large AI-driven gains in scientific discovery was **withdrawn by its preprint host** over concerns about data validity and institutional review, after it had been cited in the peer-reviewed literature. Any evidence base assembled before mid-2025 may still contain it.

3. What the controlled studies measure

With those removed, the remaining evidence is genuinely strong — and worth reading carefully, because the studies do not measure the same thing.

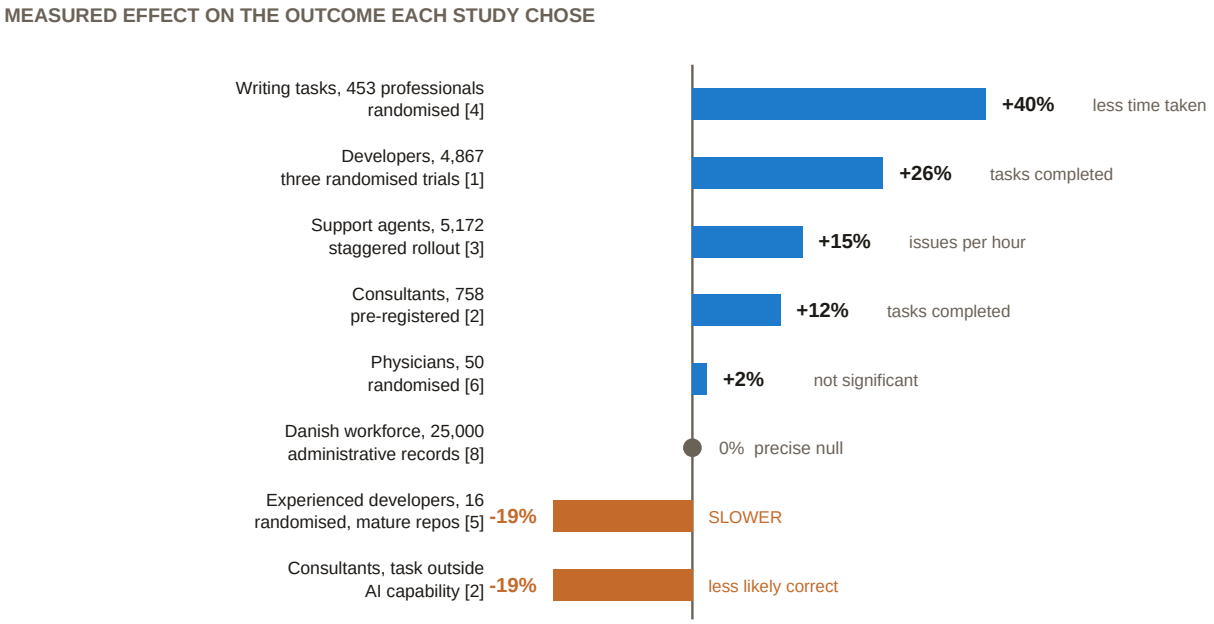

Figure 1. Measured effects across eight studies of AI assistance in skilled work, signed so that a bar to the right is an improvement whatever the underlying metric — the writing-task study measured time taken *falling* 40%. Each study measured a different outcome on a different population, so the bars are not comparable to one another; they share an axis to show the range of results the literature actually contains, including the negative and null findings.

3.1 The positive results, with their caveats attached

The largest single-study effect comes from a preregistered randomised experiment on writing tasks: average time taken fell 40% and output quality rose 18% across 453 college-educated professionals. [4] The tasks were short, self-contained and incentivised, which is a meaningful limit on transferring the result to sustained professional work.

Across three field experiments covering 4,867 developers, completed tasks rose 26.08% (standard error 10.3%). [1] Two caveats belong with that number and both come from the paper itself: two of the six authors are employed by the vendor of the tool studied, and the authors note the experiments were “rather ad-hoc as they were driven by business considerations at these companies rather than research goals.” [1] The trial was post-registered rather than pre-registered.

A pre-registered field experiment with 758 consultants found 12.2% more tasks completed, 25.1% faster, with quality more than 40% higher. [2] Four of the nine authors were employed by the firm whose consultants were the subjects.

The cleanest of the four on funding is the study of 5,172 customer-support agents, published in a peer-reviewed economics journal with no vendor authorship: issues resolved per hour rose 15% on average. [3] It is a staggered rollout rather than a randomised trial, and the authors state plainly that the findings capture medium-run effects in a single firm.

A PATTERN WORTH NOTICING

The three largest positive effects all come from studies with vendor or subject-firm authorship. The study with no vendor involvement reports the smallest headline gain — and is the one that also reports a quality decline among the most skilled workers. We do not allege bias. We note the correlation and let the reader weigh it.

4. The finding that governs everything else

The benefit shrinks as the user gets better, and in several studies it turns negative.

This is the most robust pattern in the literature, because it appears across studies that share no population, no design, no funder and no domain.

WHAT HAPPENS AS THE USER GETS MORE EXPERT

	LESS EXPERIENCED	MORE EXPERIENCED
Support agents [3]	Least skilled: +30%	Most skilled: quality declines
Developers [1]	Junior / recent hires: gains	Senior / long tenure: no effect
Consultants [2]	Below average: +43%	Above average: +17%
Open-source devs [5]	—	~5 years on the repo: 19% slower
Endoscopists [7]	—	Unaided detection fell after exposure

Figure 2. The same reversal in five independent studies. Different populations, designs, funders and fields; the direction of the gradient is consistent.

The support-agent study found less-skilled workers improved 30% on issues resolved per hour while the most experienced saw small gains in speed and *small declines in quality*. [3] The developer trials found significant gains for recent hires and junior staff “but not for developers with longer tenure and in more senior positions.” [1] The consulting experiment found below-average performers gained 43% against 17% for above-average. [2]

Two results go further than diminishing returns.

The first is the most direct test available of expert work in a familiar codebase. Sixteen experienced developers, averaging five years on the repositories concerned, completed 246 randomised tasks. They forecast that AI would cut completion time by 24%. Afterwards they estimated it had cut it by 20%. It had **increased** completion time by 19%. [5] Economists and machine-learning experts asked to predict the outcome were similarly wrong in the same direction.

> **They believed they were 20% faster. They were 19% slower. Self-reported productivity gain is not evidence of productivity gain.**
>
> METR randomised controlled trial, 2025 [5]

The second is clinical. In a randomised trial, fifty physicians given access to a language model showed no significant improvement in diagnostic reasoning against conventional resources — an adjusted difference of 2 percentage points, confidence interval –4 to 8, p = 0.60. [6] The same study found the model working *alone* scored 16 percentage points higher than the physicians using conventional resources. The tool was capable; the combination was not.

A separate observational study reported that experienced endoscopists' detection rate in *unassisted* procedures fell from 28.4% to 22.4% after a period of working with AI assistance. [7] It is observational rather than randomised and we grade it accordingly — but the direction is consistent with the rest.

4.1 Why this matters more in upstream than elsewhere

A support queue contains a great deal of routine work handled by people early in their careers. That is the profile where the measured gains are largest.

Reserves estimation, well test interpretation, completion design and acquisition evaluation are the opposite profile: non-routine work performed by experienced people whose judgement is the product. The literature says that is exactly where naive assistance helps least and can hurt.

This does not mean the technology has no place in expert work. It means the value has to come from something other than putting a capable model next to an expert and hoping — which is the implicit design of most pilots, and a reasonable explanation for why so many of them disappoint.

5. The labour picture, stated honestly

The workforce argument for AI in upstream is usually made in one direction. The best available source makes it in two, and both are in the same document.

The International Energy Agency's employment survey, drawing on more than 700 firms, unions and educators, reports that in advanced economies there are **2.4 workers within ten years of retirement for every worker under 25**, against roughly one-to-one in emerging economies, and that between now and 2035 "two out of every three new hires will be needed just to replace retiring energy workers." [11] Around 60% of companies reported labour shortages, and oil and gas saw the largest wage increases of any energy sector at about 3.7%. [11]

The same report also records that oil and gas employment stood at about 12.4 million in 2024 and was expected to **decline by 0.8% in 2025**, driven by increased layoffs at international oil companies. [11] Ageing and retrenchment are happening simultaneously. A business case that cites only the first half is

quoting selectively from its own source.

Independent survey data puts the age distribution at 48% of the traditional energy workforce aged 45 or over against 19% aged 25 to 34, and names engineering and technical operations roles as the hardest to fill. [12] We flag that this survey is published by a staffing firm and a jobs board, both of which benefit commercially from a shortage narrative, and that no sampling methodology is published with it.

5.1 What replacing someone costs, and what nobody has measured

Recruiting cost is measurable. A 2025 survey of 2,371 human-resources practitioners puts average cost per hire at $5,475 for non-executive roles and $35,879 for executives. [13] That is recruiting cost only — it excludes onboarding, lost output during ramp-up, and knowledge loss.

Time to full productivity has been measured, though not recently and not in this industry: 15 weeks for a hire from the same sector, 32 weeks from another sector, 40 weeks for a new graduate, and a full calendar year for someone returning from unemployment. [10] That study is from 2014, is UK only, was commissioned by an insurer, and explicitly **excludes engineering and energy** from its sectors. We cite it because it is the best available, not because it is a good fit.

> **THE NUMBER THAT DOES NOT EXIST**
>
> We searched specifically for a published attempt to value the institutional knowledge that leaves when an experienced engineer retires — separate from recruiting cost and lost output during ramp-up. **We found none.** The nearest study models only two components, output and logistics, with no knowledge-capital term at all. Anyone quoting a figure for the value of lost institutional knowledge is estimating, and should say so.

6. What a defensible case looks like

Given all of the above, here is the case we think survives scrutiny, and the conditions it depends on.

6.1 Claim hours, not barrels

No study in this literature measures recovered volumes, production uplift, or reserves additions attributable to AI assistance. Every credible measurement is of time, task completion, or quality on defined work. A business case denominated in barrels is claiming something the evidence base does not contain.

6.2 Claim it where the evidence points

The measured gains concentrate in work that is bounded, repeatable, and currently done by people who are not the most experienced person available. In an asset team that describes data preparation, reconciliation, screening across large well counts, first-pass diagnostics, and document production — not the final judgement on a reserves booking.

6.3 Assume the expert case needs engineering, not access

The reversal in section 4 is the strongest argument in this paper for why buying model access is not a strategy. If unaided assistance measurably slows experienced practitioners in familiar territory, then value in expert work depends on what the surrounding system does — whether the computation is fixed and checkable, whether the system declines when a method does not apply, whether the output carries its method and sources so that verification is cheap rather than expensive.

That is the subject of the first two papers in this series, and it is why we regard them as prior to the business case rather than separate from it.

6.4 Measure it yourself, and measure it the hard way

The single most important methodological finding here is that **self-reported gain is unreliable**. Participants who were 19% slower believed they were 20% faster [5]; Danish workers self-reporting about 3% of hours saved showed no measurable effect on recorded hours [8]. A pilot evaluated by asking participants whether it helped will return a positive answer almost regardless of what happened.

If a number is going to appear in a business case, it should come from elapsed time on comparable work, ideally with a control, and preferably not collected by anyone whose budget depends on the answer.

7. Limits of this paper

This is a review of evidence from adjacent fields, and the gap is the point of this section.

No controlled study of AI assistance in petroleum engineering exists that we could find. Everything in section 3 and 4 concerns developers, consultants, support agents, writers and physicians. The transfer to reservoir engineering is an argument, not a measurement, and we present it as one.

The strongest positive results carry vendor or subject-firm authorship, noted at each point. The strongest negative result [5] is a preprint with sixteen participants, and its authors state they do not claim their developers or repositories represent most software work. The endoscopy result [7] is observational, and we read it through the publisher’s summary rather than the paper, which we could not access.

Several sources are dated: the writing-task and consulting experiments are from 2023 and reflect an earlier model generation, and models have improved materially since. That cuts in favour of the technology, and we say so.

Finally, one source in this area was **retitled and revised** between our first and second encounters with it, changing a widely-quoted figure. [8] We cite the current version. Anyone quoting the older number is quoting a superseded paper.

8. Conclusion

The credible claim is hours, on bounded work, measured properly, with the expert case engineered rather than assumed.

The evidence for AI assistance in professional work is real and, in places, large. It is also more specific than the market's summary of it. The gains concentrate where work is routine and the worker is early in their career. They shrink with expertise. In several careful studies they reverse.

For an industry whose expensive hours are spent on non-routine judgement by experienced people, that is not a comfortable finding, and pretending otherwise produces exactly the pilots that disappoint. The honest position is that the value in expert work is available but conditional — it depends on the system around the model doing the things the model cannot, and it should be claimed in hours saved on defined work rather than in volumes recovered.

A business case built that way is smaller than the one in the brochure. It has the advantage of being true, and of surviving the meeting where somebody checks.

How SolvxAI has already resolved this

SolvxAI makes its value case exactly the way section 6 says a defensible one must be made — because we read this evidence before we priced anything.

The claim is denominated in hours, on the bounded work where the measured gains actually live: data preparation and reconciliation, screening across hundreds of wells, first-pass diagnostics, and document production. And the expert case is engineered rather than assumed — the reversal in section 4 happens when an experienced person is handed an opaque suggestion to verify the hard way, so SolvxAI makes verification cheap instead: every figure arrives with its method, units, validity band and sources attached, turning review from re-derivation into inspection.

We also accept the measurement standard this paper argues for. A pilot on SolvxAI should be scored on elapsed time against comparable work, with a control — not on how it felt.

ABOUT SOLVXAI

SolvxAI is the agentic operating system for upstream oil and gas. To see this working against your own asset, or to request the technical papers in this series, contact sd@nexaconsultinc.com.

References

Sources are listed with publication year; preliminary, self-reported, or vendor-authored work is flagged as such.

[1] **Cui, Z. K., Demirer, M., Jaffe, S., Musolff, L., Peng, S., Salz, T.** "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." *Management Science*, 2025. Randomised; 4,867 developers. **Two of six authors are employed by the vendor of the tool studied.** Post-registered (AEARCTR-0014530). https://economics.mit.edu/sites/default/files/inline-files/draft_copilot_experiments.pdf

[2] **Dell'Acqua, F., et al.** "Navigating the Jagged Technological Frontier." Harvard Business School Working Paper 24-013, 22 September 2023; subsequently published in *Organization Science*. Pre-registered field experiment; 758 consultants. **Four of nine authors were employed by the firm whose consultants were the subjects.** GPT-4 era. https://mitsloan.mit.edu/sites/default/files/2023-10/SSRN-id4573321.pdf

[3] **Brynjolfsson, E., Li, D., Raymond, L.** "Generative AI at Work." *Quarterly Journal of Economics* 140(2):889–942, February 2025. Peer-reviewed; staggered rollout, not a randomised trial; 5,172 support agents; no vendor authorship. http://danielle.li/assets/docs/GenerativeAIatWork.pdf

[4] **Noy, S., Zhang, W.** "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381(6654):187–192, 14 July 2023. Peer-reviewed, preregistered randomised experiment; 453 professionals; short incentivised writing tasks. https://pubmed.ncbi.nlm.nih.gov/37440646/

[5] **METR.** "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." arXiv:2507.09089, July 2025. **Preprint**, not peer-reviewed; randomised at task level; 16 developers, 246 tasks in mature repositories. https://arxiv.org/abs/2507.09089

[6] **Goh, E., et al.** "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial." *JAMA Network Open* 7(10):e2440969, 28 October 2024. Peer-reviewed single-blind randomised trial; 50 physicians; clinical vignettes, not live patients. Not vendor-funded. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395

[7] **Budzy■, K., et al.** *The Lancet Gastroenterology & Hepatology*, August 2025. Retrospective observational; 19 experienced endoscopists, 1,443 non-AI colonoscopies. **We accessed the publisher's summary rather than the paper itself** and grade our citation accordingly. Funded by public research bodies.

[8] **Humlum, A., Vestergaard, E.** "Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI." NBER Working Paper 33777, May 2025, **revised March 2026**; previously circulated under a different title, and a widely-quoted figure changed between versions. Not peer-reviewed. 25,000 workers, 7,000 workplaces, linked to national administrative registers. https://www.nber.org/system/files/working_papers/w33777/w33777.pdf

[9] **Meyer, A. N., Barr, E. T., Bird, C., Zimmermann, T.** "Today was a Good Day: The Daily Life of Software Developers." *IEEE Transactions on Software Engineering*, 2019. Peer-reviewed; 5,971 self-reported responses, all from one company. Cited here for the published **range** of 9%–61% of time spent coding across studies. https://www.microsoft.com/en-us/research/wp-content/uploads/2019/04/devtime-preprint-TSE19.pdf

[10] **Oxford Economics / Unum.** *The Cost of Brain Drain*, 3 March 2014. **Sponsor-commissioned** by an insurer; UK only; 500+ firms; **excludes engineering and energy sectors**. Cited for time-to-productivity figures and to refute the "28 weeks" claim, which does not appear in it. https://www.oxfordeconomics.com/wp-content/uploads/2023/05/cost-brain-drain-report.pdf

[11] **International Energy Agency.** *World Energy Employment 2025*. Intergovernmental agency; annual survey of more than 700 energy firms, trade unions and educators; published methodology. Note: the executive summary and chapter 1 state the hiring-bottleneck share slightly differently, and we report the discrepancy rather than picking one. https://iea.blob.core.windows.net/assets/e5b9e908-26fe-4228-8978-e4ee36c555ed/WorldEnergyEmployment2025.pdf

[12] **Global Energy Talent Index 2026.** Published 4 February 2026; 9,000+ professionals across 143 countries. **Vendor-authored** — published by an energy staffing firm and a jobs board, both of which benefit commercially from a shortage narrative. **No sampling methodology, response rate or weighting is published.** https://www.getireport.com/

[13] **Society for Human Resource Management.** 2025 Benchmarking Reports, released 15 October 2025. Survey of a random sample of members, fielded 9 January to 3 March 2025; 2,371 responses. Cost per hire is **recruiting cost only** and is not a replacement cost.

https://www.shrm.org/about/press-room/shrm-releases-2025-benchmarking-reports--how-does-your-organizat