Raj Jobalia

I CAN'T TYPE

When the Model Becomes the First Draft of the Mind

SEPTEMBER 2026

Dedicated to the ugly first draft.

This essay is written by someone who uses these tools every day and does not intend to stop. It is a complaint about sequence, not a case for abstinence.

Every empirical claim is sourced. Every figure that is not built from a published number says so on its face. Nothing here diagnoses anybody.

A RESEARCH ESSAY
COGNITIVE OFFLOADING IN THE FRONTIER-MODEL ERA
September 2026

I can't type a LinkedIn post.

Not literally. My hands work. What has stopped working is the part that comes before the hands: the moment where I decide what I think, choose the first word, misspell it, and fix it. That moment used to be the whole job. Now it is a prompt, and the prompt is shorter than the thought.

I notice it most on things that don't matter. A birthday message. A two-line reply. A post nobody will remember by Thursday. I open the box, and instead of a sentence I get a kind of static — and then, without quite deciding to, I open a model. The draft comes back cleaner than anything I would have written. I change two words so it sounds like me. I press post. The whole transaction takes ninety seconds and leaves no trace in my memory at all.

Mathematics still feels like mine. I can still estimate, still catch a wrong order of magnitude, still do the arithmetic in my head while someone reaches for their phone. I think that is not a fact about mathematics. I think it is a fact about when I learned it — before there was anything to reach for.

Meanwhile the models keep getting better. Every few months another one lands, and the gap between what I can do alone and what I can do with help gets wider in the direction that flatters the help. The industry measures that gap obsessively, from one side. This essay is about the other side.

So: is this real, or is it just what change feels like from the inside? I went looking for the evidence. Some of it is worse than I expected. Most of it does not exist.

Let me tell you what I found.

Contents

I. The Confession

The problem is not access to intelligence. It is the disappearance of the moment before access. A spell-checker corrects a token after you type it. A generative model removes the need to originate the token at all.

> The tool that does the work for you does not merely save you the work. It saves you the practice.
>
> AFTER A REMARK OF DONALD NORMAN'S

I USED TO HAVE an idea and then meet the resistance of language. I had to choose the opening line, remember the spelling, decide what I actually meant, and tolerate an ugly first draft on the way to a better one. That resistance was annoying. It was also the entire mechanism by which I found out what I thought.

Now the idea enters a model before it enters a sentence. The model proposes the frame, supplies the transitions, and removes the embarrassment. The artifact improves. The authorship experience thins out.[1]

This matters because of *where* in the sequence the tool sits. Consider the five stages that turn an intention into a published sentence: intent, retrieval, composition, revision, publication. A spell-checker acts at stage four. It waits until I have produced something and then corrects it, which means stages two and three still happened. A generative model acts at stage one. It can absorb the two stages in the middle entirely — and those are precisely the stages that rehearse vocabulary, spelling, and the order of an argument.

[1] This is not a complaint about laziness. Laziness would be a choice. This is closer to a reflex — the hand reaching for the tool before the mind has finished forming the problem.

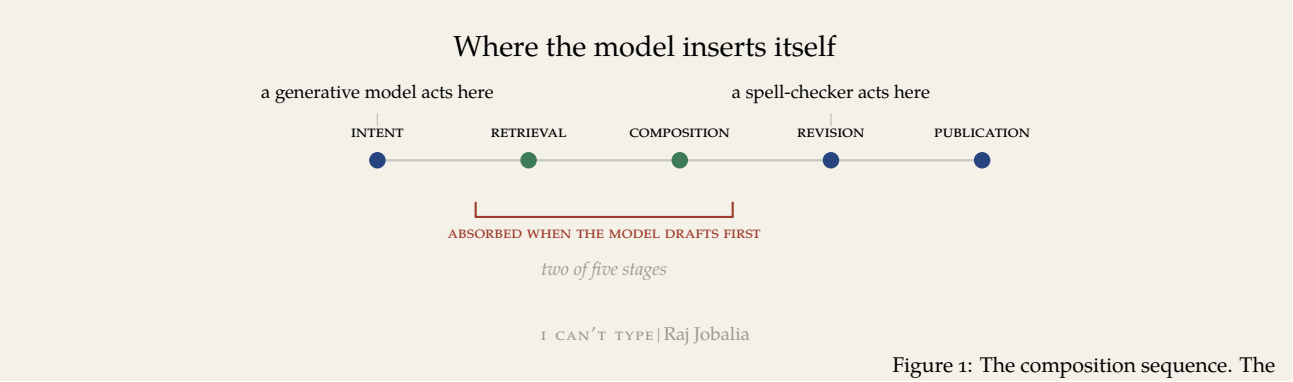

Figure 1: The composition sequence. The difference between the two tools is not capability but position: one waits for output, the other precedes it. *Conceptual process model.*

That is the whole hypothesis, and it is narrower than "these machines make you stupid." It is: *when generation precedes retrieval, retrieval stops being practised.* Narrow claims have the advantage of being checkable, so the rest of this essay tries to check it.

II. From Offloading to Delegation

Cognitive offloading is ordinary and mostly good. What is new is not that we offload, but how much of the reasoning loop the tool is able to absorb — and that a language model is the first one that can propose the problem representation itself.

Cognitive offloading is the use of physical action or an external tool to reduce the mental demand of a task.[2] Notes preserve memory. Calculators execute arithmetic. Search engines retrieve stored information. None of this is a failure of intelligence; all of it is ordinary intelligent system design, and people who refuse to do it are not smarter, only slower.

[2] Risko & Gilbert (2016), *Cognitive Offloading*, *Trends in Cognitive Sciences* 20(9).

There is also a long literature saying the trade is real. Sparrow and colleagues found that expecting future access to information lowered recall of the information itself while improving recall of where to find it.[3] We did not lose memory; we reallocated it to an index.

[3] Sparrow, Liu & Wegner (2011), *Google Effects on Memory*, *Science* 333(6043). Fisher, Goddu & Keil (2015) found relatedly that search inflates people's estimates of their own internal knowledge.

So the reassuring analogy is available and it is not stupid: people said the same thing about calculators, and about writing itself. The question is whether this tool belongs in that sequence or breaks it.

I think it breaks it, and the reason is in the last column of that chart. A calculator requires a formed equation; you must model the problem before it is of any use. Search requires a formed query and returns material you must still synthesise. A language model accepts a mood, a fragment, or an inarticulate discomfort — and

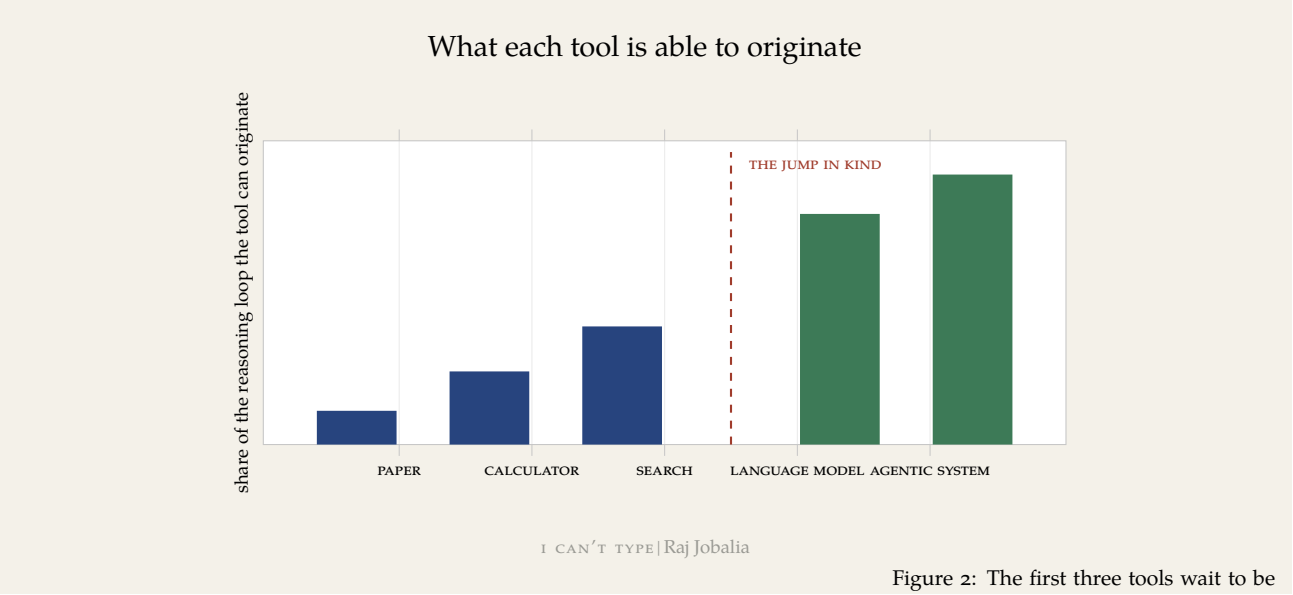

Figure 2: The first three tools wait to be told what the problem is. The last two will tell you. *Conceptual — bar heights are illustrative and were not measured.*

supplies structure. That accessibility is genuinely wonderful. It is also what allows the tool to arrive *before* the user has represented the problem at all.

Watch what happens to the verb. With a calculator, the human still models. With search, the human still writes. With a chat model, the human selects. With an agent, the human approves. Every step is defensible on its own and every step was adopted because it was useful. The cumulative effect is that origination — the part that builds the internal model — has quietly moved off the human's side of the table.

The concern is sequence, not purity. Offloading after you have formed a view is efficient and always was. Offloading in order to avoid forming one is a different act that happens to look identical from the outside, and increasingly identical from the inside too.

III. The 2026 Model Layer

By 2026 the model is not a website you visit. It is a layer running underneath the places where writing already happens. That is why the complaint feels inescapable: you can quit a product, but you cannot easily quit a layer.

When I say the models keep getting better, I am not really describing any single model. I am describing the number of surfaces on which a competent draft is one keystroke away. Chat. Search. The document. The code editor. The operating system. The reply box I am typing into right now.

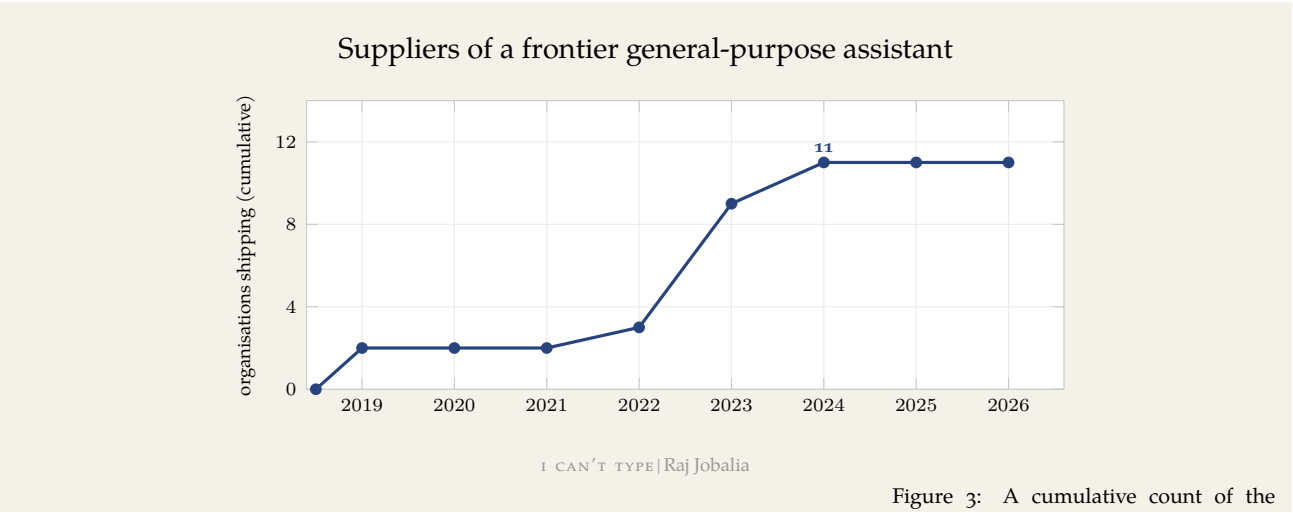

Figure 3: A cumulative count of the eleven organisations listed opposite. The curve flattens not because progress stopped but because the market saturated: by 2026 the assistant is ambient. *Illustrative census at year granularity — representative labs, not a release database or a market-share claim.*

The suppliers, in rough order of arrival: OpenAI (ChatGPT), Google DeepMind (Gemini), Meta (Llama), Microsoft (Copilot), Amazon (Nova and Bedrock), Cohere, Mistral, xAI (Grok), Alibaba (Qwen), DeepSeek, and Apple. These are named as representative product context; the names and version numbers change

far faster than the cognitive contract underneath them, which has been stable since about 2023: fluent text, on demand, before you have finished thinking. Eleven organisations is not the interesting number. The interesting number is that in 2019 you had to go somewhere to find one, and by 2026 you have to go somewhere to avoid one.

What nobody publishes is the split that would actually matter: how often the layer is helping a person do something, and how often it is simply doing the thing instead. I do not know which of those two my birthday message was. I suspect I have stopped being able to tell.

HERE IS THE ASYMMETRY that organises everything that follows. Capability is measured continuously, publicly, and expensively, because somebody profits from the number. The other axis — what happens to the person's ability to do the thing alone — is measured almost never, because nobody does.

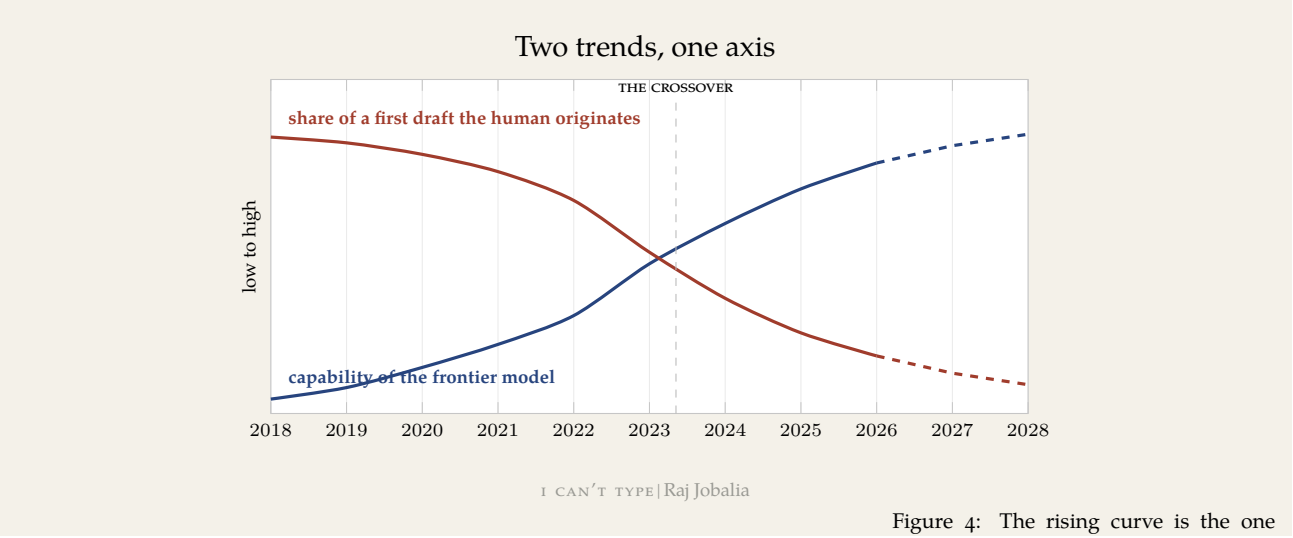

Figure 4: The rising curve is the one the industry reports every quarter. The falling curve is the one this essay is about, and there is no instrument for it. Where they cross is where a person's default stops being *draft it* and becomes *prompt it*. *Schematic — both curves are illustrative and contain no measured values.*

I want to be exact about what that figure is and is not. It is not data. Neither curve carries a number, and the dashed segments are extrapolation rather than forecast. It is a picture of an argument, and the argument is that these two things are usually discussed in separate rooms when they are plainly the same subject.

IV. The Productivity Bargain

A serious argument about dependence has to begin by admitting why dependence is attractive. The gains are real, they are large, and they fall unevenly — most heavily toward the people who knew least to begin with.

> Any account of these tools and cognition that cannot explain why so many people adopted these tools voluntarily, and kept them, is not an account worth having.
>
> THE CONCESSION THIS CHAPTER MAKES

Start with the good news, because it is genuinely good. In a preregistered experiment, 453 college-educated professionals wrote mid-level business documents. Access to ChatGPT cut average completion time by 40% and raised blind-graded output quality by 18%.[4] In a field deployment covering 5,172 customer-support agents, assistance raised issues resolved per hour by 15% on average.[5]

[4] Noy & Zhang (2023), *Science* 381(6654), 187–192.

[5] Brynjolfsson, Li & Raymond (2025), *Quarterly Journal of Economics* 140(2), 889–942.

Note the shape of panel B. The largest gains went to the least experienced workers; the most experienced saw small gains and, in some comparisons, small quality declines. This is the first appearance of a pattern that will not go away: *assistance substitutes most powerfully for skill the user does not yet have.* That is wonderful if you are new. It is a more complicated bargain if the skill it is substituting for is one you used to possess.

Then the edge. Dell'Acqua and colleagues studied 758 consultants on realistic tasks. Inside the model's capability frontier, users

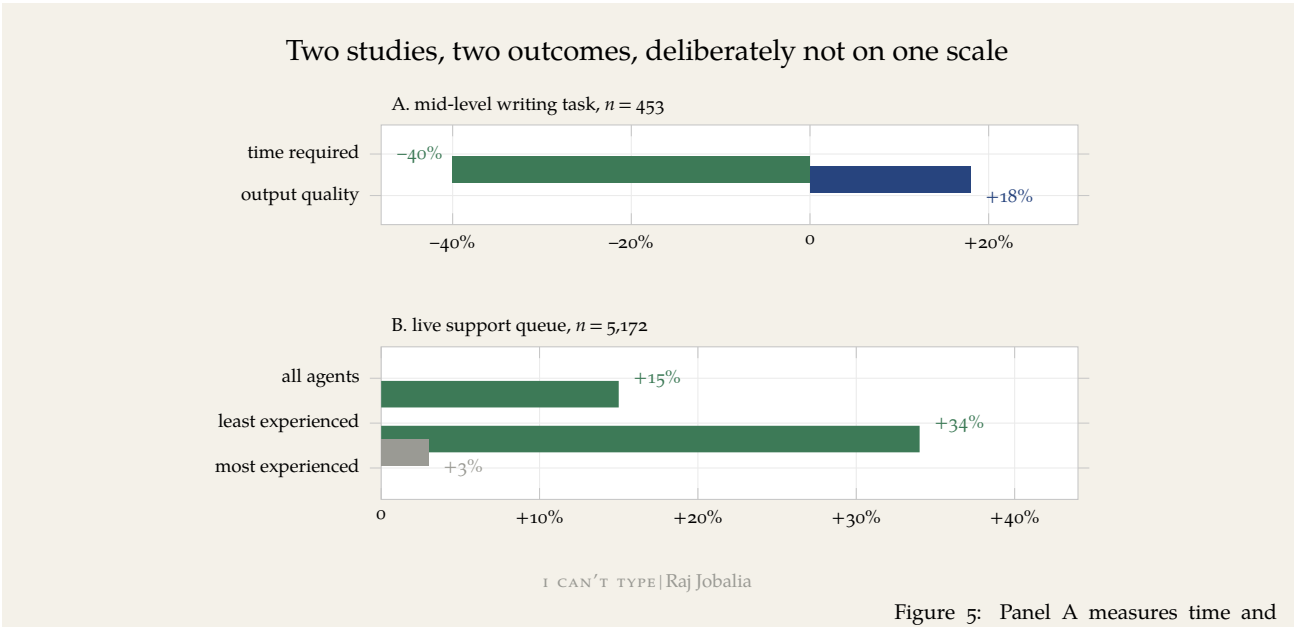

Figure 5: Panel A measures time and rated quality on a writing task; panel B measures throughput in a live support queue. They are different outcomes and the bars must not be compared across panels. Tier values in B are approximate readings of the reported gradient. *Reported effect sizes.*

completed 12.2% more work, moved 25.1% faster, and produced output rated more than 40% higher in quality. On a task deliberately chosen to sit just outside that frontier, the same users were 19 percentage points *less* likely to reach the correct answer.[6]

[6] Dell'Acqua et al. (2026), *Organization Science*.

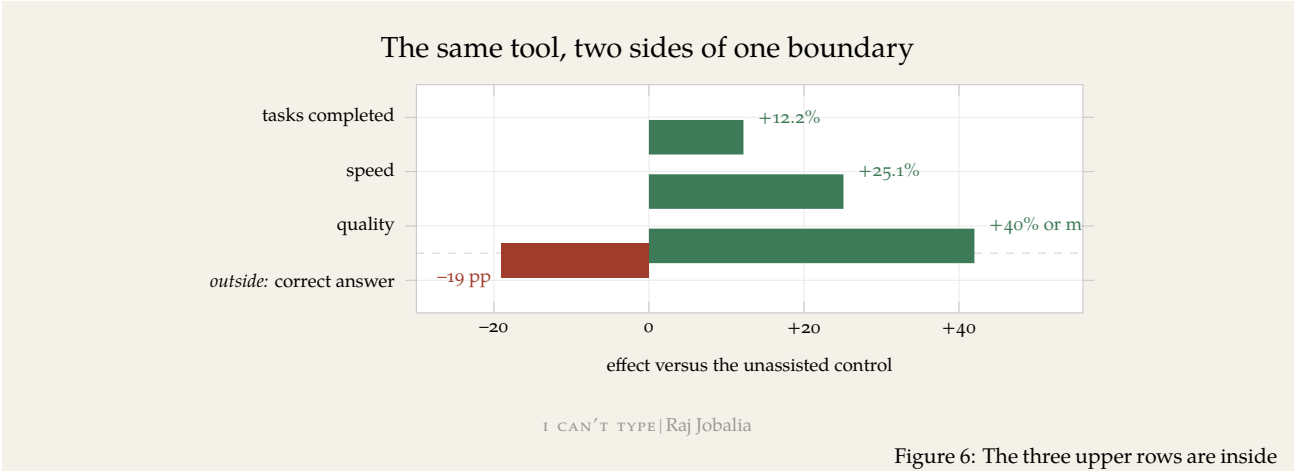

Figure 6: The three upper rows are inside the model's capability frontier; the bottom row is the task chosen to sit outside it. It is not a smaller version of the rows above — it points the other way. *Reported treatment effects; the final row is percentage points, not percent.*

If the model were merely weaker outside its frontier, its users would have scored like the control group. They scored worse. The tool did not fail to help; it displaced a judgement the consultant would otherwise have made. And nothing in the interface distinguishes the two cases — the prose is equally fluent on both sides of the line. Fluency is not evidence. It is the absence of a visible seam.

THEN THE PART THAT UNSETTLED ME MOST. METR ran a randomised controlled trial with 16 experienced open-source developers on 246 tasks in repositories they knew well. Beforehand, they expected the tool to make them 24% faster. Afterwards — *having done the work* — they believed it had made them 20% faster. Measured completion time had increased by 19%.[7]

[7] Becker, Rush, Barnes & Rein (2025), arXiv:2507.09089. A preprint; small sample, expert developers, early-2025 tools.

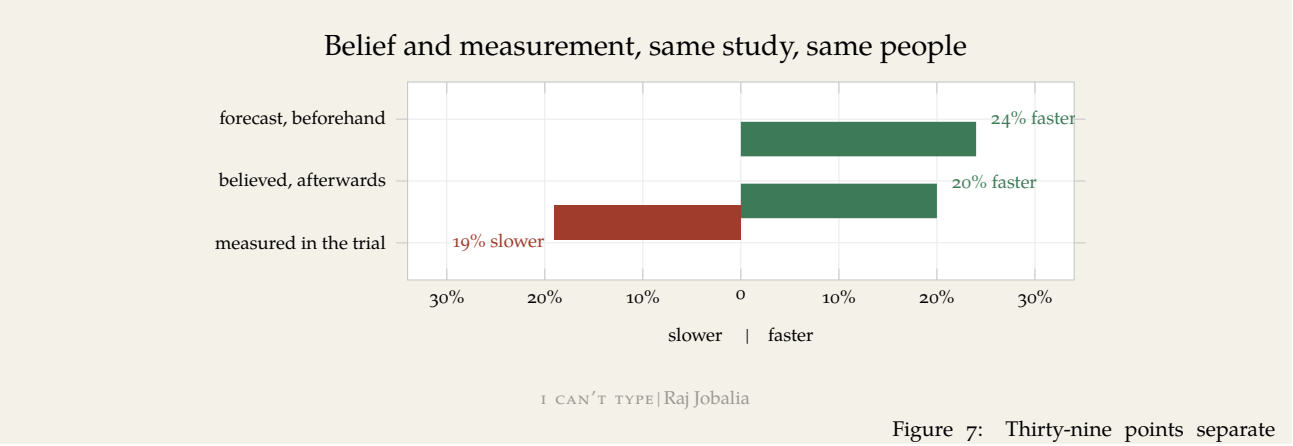

Figure 7: Thirty-nine points separate what these developers believed from what the stopwatch recorded — and the belief did not move after they had done the work. *Reported values; positive is faster.*

Here is why that result matters for a document like this one. If experienced engineers cannot detect a nineteen-percent slowdown in their own working day, then I am not a reliable instrument either. My sense that I have got worse is exactly the kind of introspective report this study just falsified in the opposite direction. That cuts against my complaint as hard as it cuts for it, and I would rather say so than pretend otherwise.

V. Memory, Learning, and What Happens Afterwards

Two studies ask a question almost nobody else asks: not how good was the work, but what happened to the person who did it. Their answers are the reason this essay exists.

The first is small, contested, and impossible to forget. A 2025 MIT-affiliated preprint assigned 54 participants to write essays with a language model, with a search engine, or with nothing at all. Minutes after submitting, each was asked to quote a single sentence from the essay they had just written. Fifteen of eighteen search users managed it. Sixteen of eighteen brain-only writers managed it. In the language-model group, the number who managed it was zero.[8]

[8] Kosmyna et al. (2025), arXiv:2506.08872.

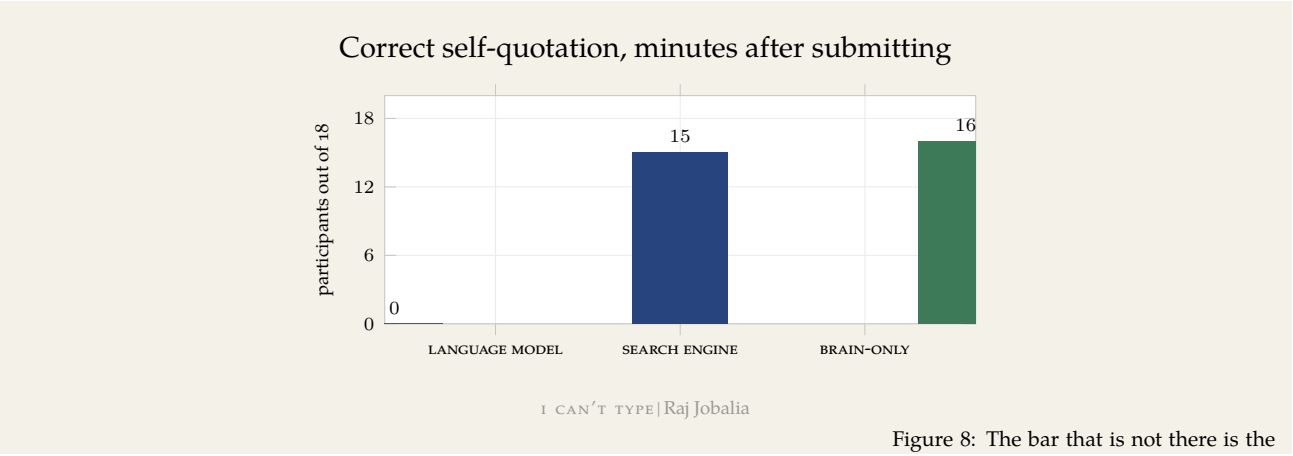

Figure 8: The bar that is not there is the finding. Every group produced an essay. Only one group could not produce a sentence from it. *Session-one counts; preprint under review, with a published rebuttal.*

A 2026 methodological comment argues that study is underpowered, hard to reproduce, inconsistently reported, and vulnerable to over-interpretation of its EEG comparisons. It is a signal worth following, not a finding to build policy on.

They all had an essay. One group did not have a memory of writing it. Whatever the EEG claims in that paper are worth — and the rebuttal suggests not much — the quotation count is a simple behavioural fact, and it is the closest thing in the literature to the specific thing I notice about my own posts.

THE SECOND STUDY IS LARGER, BETTER DESIGNED, and more hopeful. In a field experiment with nearly a thousand high-school mathematics students, access to a GPT-4-based tutor raised performance sharply while the tool was available: 48% above control for a standard chat interface, and 127% above control for a safeguarded tutor built to hint rather than answer. Then the researchers took the tools away and tested everyone unaided. The students who had used the unrestricted interface scored 17% *worse* than students who had never had access at all. The safeguarded tutor left no measurable deficit.[9]

[9] Bastani et al. (2025), *PNAS* 122.

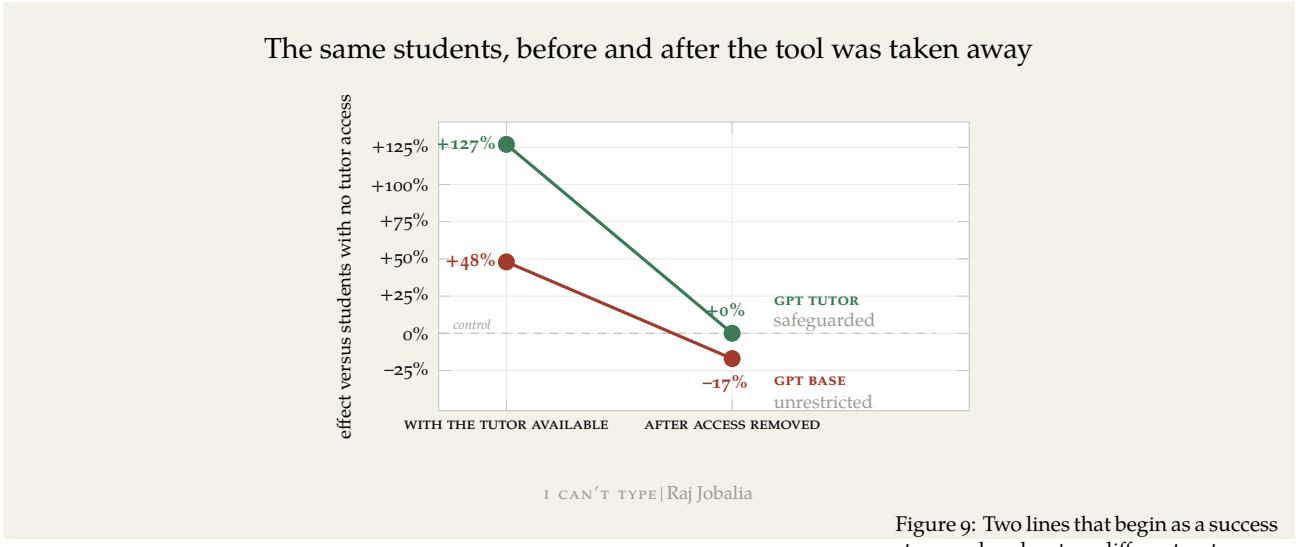

Figure 9: Two lines that begin as a success story and end as two different outcomes. The safeguarded tutor gave the *larger* boost during practice and left no deficit; the unrestricted one gave a smaller boost and ended below the students who never touched it. *Reported treatment effects.*

I find this the most important result in the essay, and also the most encouraging, because of what caused it. The harm was not caused by the model being too capable. Both arms used the same underlying model. The harm was caused by the interface being willing to answer. Wrap the identical model in something that demands an attempt, hints instead of solving, and checks understanding, and you get the best result on both measures at once.

That is a product decision. Not a property of the technology, not an inevitable consequence of intelligence — a decision somebody

made about a text box.

VI. The Measurement Gap

Every study in this essay measures what happened to the work. Exactly one also measures what happened to the worker. Plotted together, the shape of the literature becomes visible: a crowded horizontal band and a nearly empty vertical one.

Put every finding on two axes at once. Horizontally: what did assistance do to the output? Vertically: what did it do to the person's unassisted ability afterwards? Almost every study in this essay has a horizontal position. Almost none has a vertical one, because almost none asked.

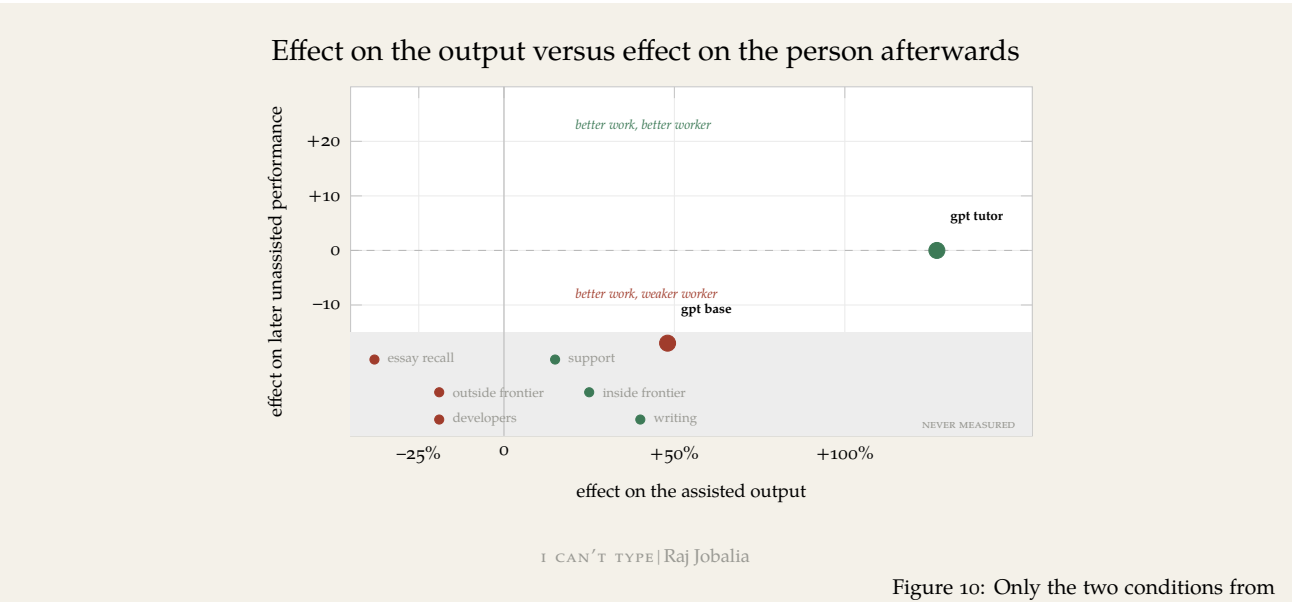

Figure 10: Only the two conditions from the schools experiment sit in the plane at all. Every other point is parked in the grey band because its study never asked the second question. *Horizontal positions use each study's own outcome measure and are not comparable with one another.*

Every grey dot is a research programme that could have answered the question and did not. That is a claim about the literature,

not about the technology — it remains entirely possible that assistance leaves independent skill untouched in most settings. What can be said with confidence is that we have mostly not looked. And in the one place we did look properly, the two axes disagreed.

VII. Language, and Why Mathematics Feels Different

Spelling is the symptom; retrieval is the training signal. The reason arithmetic still feels like mine and sentences do not has nothing to do with the subjects and everything to do with when I learned them.

There is no strong longitudinal evidence that ordinary language-model use causes durable spelling decline in adults. Nobody has run that study. My complaint is therefore a hypothesis about practice, not a diagnosis — but it is a testable one, because spelling is a retrieval task, and retrieval is precisely the thing offloading reliably reduces.

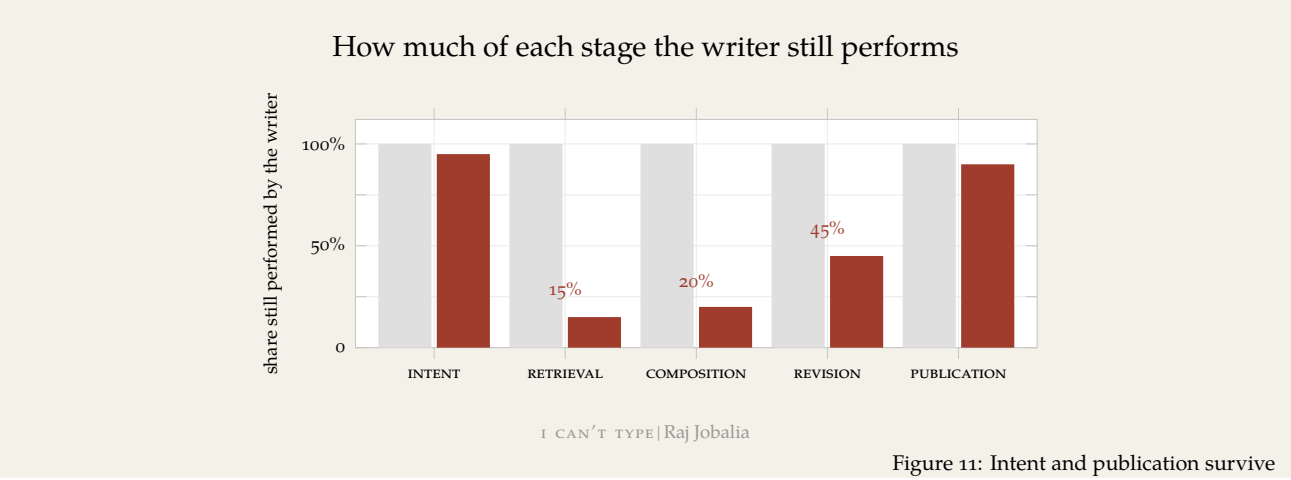

Figure 11: Intent and publication survive intact — the writer still wants to say something, and still presses post. What collapses are the two middle stages, which are the only ones that rehearse vocabulary, spelling, and sentence order. Grey is unaided writing; red is what remains when a model drafts first. *Conceptual — proportions are illustrative and were not measured.*

This gives the complaint a shape that can be checked. If the mechanism really is reduced retrieval practice, the deficit should be *specific*: worse at producing a word cold, unchanged at recognising one already on the screen. It should also be reversible, because

retrieval practice is the standard remedy for exactly this kind of decay.

WHICH BRINGS ME TO ARITHMETIC. I said at the start that mathematics still feels like mine. I now think the protective factor was never the subject. It was that I ran the skill to automaticity, unaided, for years, before any tool was permitted — and a routine consolidated that thoroughly does not evaporate when a calculator appears. It rests. Writing, for me, never got that treatment in the same way, because the assistance arrived while the habit was still forming.

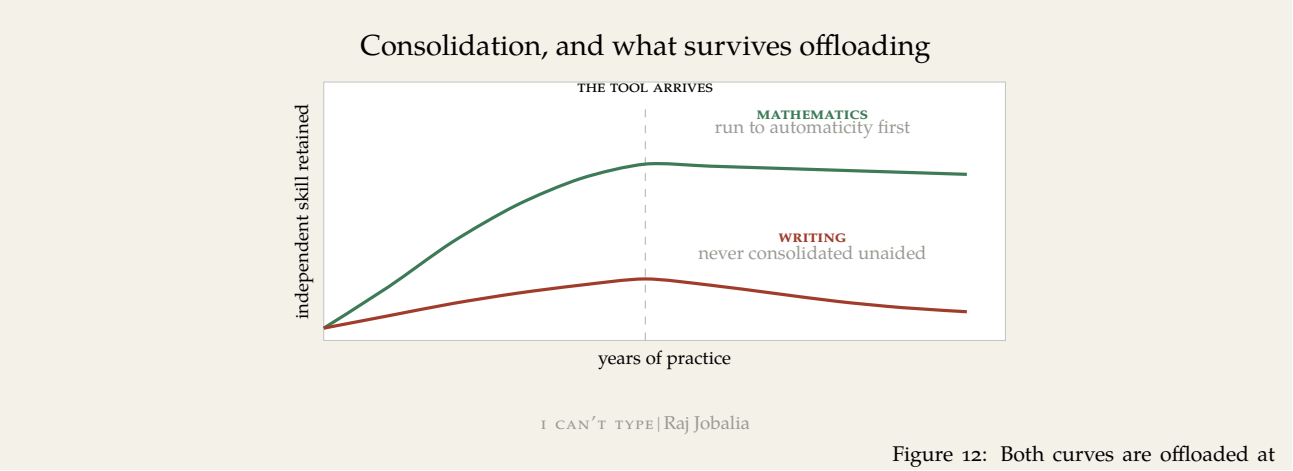

Figure 12: Both curves are offloaded at the same moment and to the same degree. Only the height at which they are interrupted differs — and that height was set years earlier. *Schematic; not a comparative study of the two subjects.*

If that is right, the policy is not *never use tools*. It is *build the internal model before renting the external one for every step*. And it means the person most at risk is not me. It is the writer who never had a pre-model baseline at all, and who therefore has no way of noticing what is missing.

VIII. A Model of Cognitive Agency

Treat a task as a bundle of operations rather than a binary choice between using a model and not using it. Agency survives when the person keeps enough of the loop to form intent, notice error, and learn from the result.

Five operations turn a problem into a finished thing. frame: what is the real question? retrieve: what do I already know? generate: what are the options? verify: what is true and complete? commit: what do I endorse? Only one of those five is a text problem, and it is the one the model is best at.

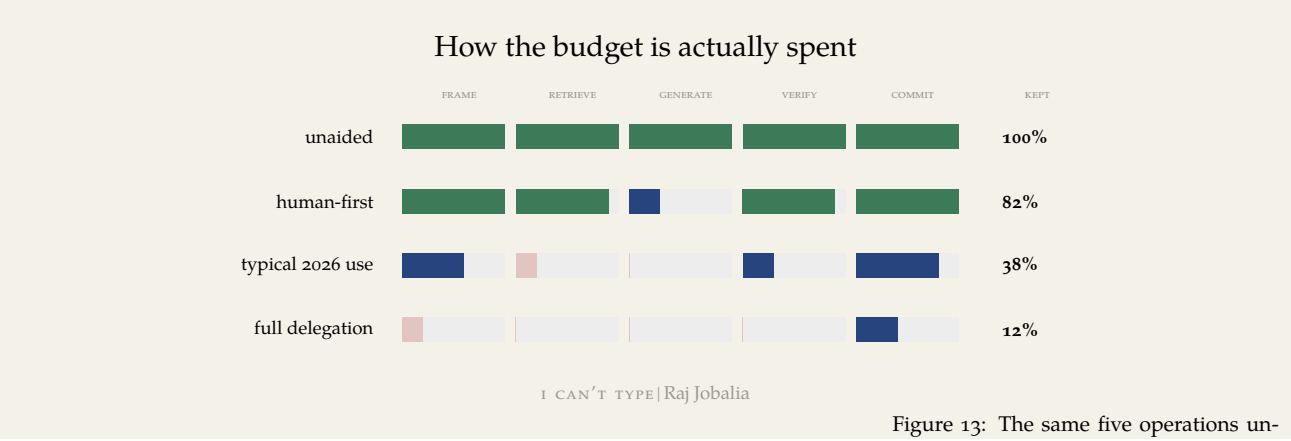

Figure 13: The same five operations under four working styles. Human-first delegates one operation; typical use delegates three and a half. *Conceptual allocation — not measured.*

The value of this frame is that it replaces an unanswerable question — *is the model making me worse?* — with an answerable one: *which of the five did I still do myself, and did I choose that?* Someone who frames the problem, retrieves what they know, verifies the output and commits to it has delegated one operation out of five. Someone who prompts, skims and posts has delegated four and kept only the one that carries the consequences.

Nothing here says the second person is lazy. It says the second person has stopped generating the evidence they would need in order to notice a change.

IX. The Human-First Protocol

One load-bearing rule and eleven pieces of scaffolding. The rule is that the model does not go first.

This is a proposal, not a finding. It is assembled from the mechanisms the evidence actually supports: attempt before assistance, retrieval before retrieval-substitutes, and explicit verification where fluency is most persuasive.

Before the model. Write the claim in one sentence. List three things you already know. Attempt the hard step for ten minutes. Name what would change your mind.

With the model. Ask for alternatives, not authority. Request sources and counterarguments. Make it expose its assumptions. Log which of its moves you accepted.

After the model. Close it and restate the answer cold. Check every material claim at source. Rewrite the opening in your own voice. Pick one skill to practise unassisted.[10]

And a reset, for anyone who suspects the baseline has already gone. Four weeks, one change per week. week 1, notice: log every prompt, delay the first one by ten minutes, write 150 words unaided daily. week 2, retrieve: before search or chat, list what you already know. week 3, verify: use the tools normally, but verify three consequential claims and record one model error a day. week 4, create: publish one piece from a human first draft, using the model only for critique.

[10] Two of these twelve do most of the work. "Attempt the hard step for ten minutes" is the attempt gate that separated the two lines in the schools experiment. "Close it and restate the answer cold" is the retrieval test the essay writers could not pass.

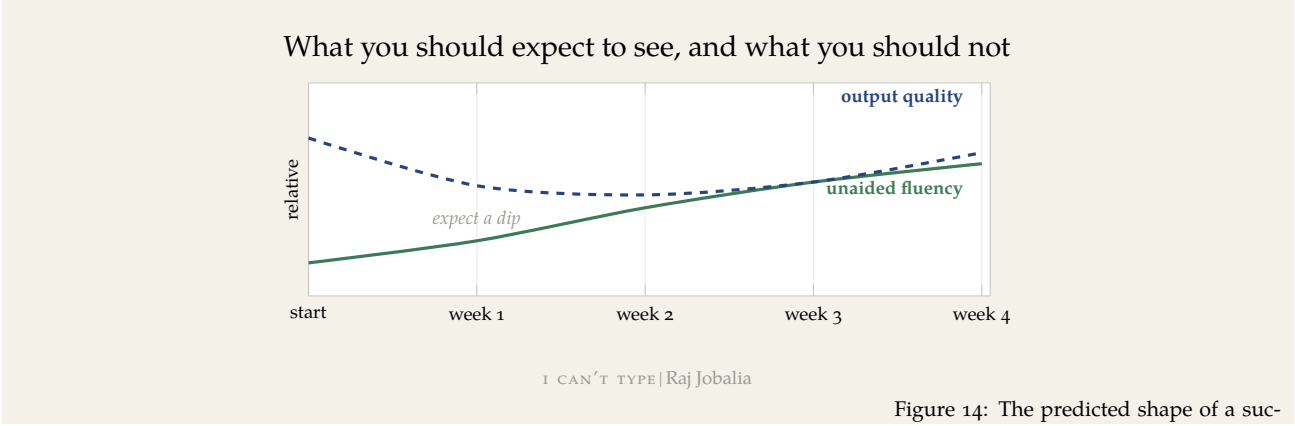

Figure 14: The predicted shape of a successful reset. Output quality is expected to fall in week one and recover by week four; the dip is the point, because it measures how much of the fluency was rented. If it never recovers, the reset has told you something true and you should stop. *Hypothesised trajectory — a prediction to be tested, not a result.*

Measure four things weekly: minutes before the first prompt; claims still remembered after twenty-four hours; whether it still sounds like you, one to five; and model errors you caught yourself.

X. Limits, Design, and Parting Thoughts

What this essay cannot prove — and what follows anyway.

Six limits belong in the body of the argument rather than in an appendix. There is *no longitudinal spelling study*; my complaint is credible and internally consistent but it is not population-level evidence. *Different tasks give different outcomes*, and several results here point in opposite directions; there is no single "assistance effect". The *systems move fast*, and the 2025 developer trial is already a study of old tools. Some findings rest on *selection and self-report* rather than measured ability. Two key studies are *preprints*, one with a published rebuttal. And *dependence is often rational* — external cognition is usually efficient, and the concern is unchosen dependence, not dependence itself.

What survives all six is modest in form and large in consequence: the second axis is barely measured, and where it has been measured, it disagreed with the first. That is enough to justify changing how you work. It is not enough to justify a diagnosis.

If I could change one thing about the tools themselves, it would not be their intelligence. It would be their scorecard. Assistants are optimised for task completion and user satisfaction, and they are very good at both. There is no row for whether the user could explain, verify, or repeat the work without them. The schools experiment is the proof that such a row is achievable: attempt gates, uncertainty shown by default, recall checkpoints at the end of a session, provenance for every claim, a critique-only

mode, and — above all — reporting assisted output and later independent performance as two separate numbers rather than one.

So, THE PARTING THOUGHT. Frontier models can make a professional writer faster, a novice support agent better, a consultant more productive, and a student more successful during practice. The same systems can make an expert developer slower without his noticing, leave a student worse off the moment the tool is withdrawn, and move critical thinking to a verification step that may never actually happen. That is not a contradiction. It is the jagged bargain of cognitive tools.

So if you cannot type the LinkedIn post, do not begin by asking the model to rescue the post. Begin with the sentence you are ashamed to write. Misspell something. State the claim badly. Discover what you think by watching yourself fail to say it. Then invite the model in — as critic, adversary, researcher, or editor, but not as author.

The complaint that opened this essay was that the models keep getting better and I keep getting worse. Only the first half of that sentence has ever been measured.

The first draft is not waste.
It is where the mind leaves evidence of itself.

References

Bastani, H., et al. (2025). Generative AI without guardrails can harm learning: evidence from high school mathematics. *PNAS* 122. doi:10.1073/pnas.2422633122

Becker, J., Rush, N., Barnes, E. A., & Rein, D. B. (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv:2507.09089

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *Quarterly Journal of Economics* 140(2), 889–942. doi:10.1093/qje/qjae044

Dell'Acqua, F., et al. (2026). Navigating the Jagged Technological Frontier. *Organization Science*. doi:10.1287/orsc.2025.21838

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for Explanations. *Journal of Experimental Psychology: General* 144(3), 674–687. doi:10.1037/xge0000070

Kosmyna, N., et al. (2025). Your Brain on ChatGPT. arXiv:2506.08872. Preprint under review.

Lee, H.-P., et al. (2025). The Impact of Generative AI on Critical Thinking. *CHI 2025*. doi:10.1145/3706598.3713778

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

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences* 20(9), 676–688. doi:10.1016/j.tics.2016.07.002

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory. *Science* 333(6043), 776–778. doi:10.1126/science.1207745

Stankovic, M. S., et al. (2026). Comment on: Your Brain on ChatGPT. arXiv:2601.00856

Ward, A. F., et al. (2017). Brain Drain. *Journal of the Association for Consumer Research* 2(2), 140–154. doi:10.1086/691462

A narrative synthesis, not a systematic review. Peer-reviewed work was given priority; preprints are labelled and, where a published rebuttal exists, paired with it. All figures are original. No chart has been reproduced from a source publication. Accessed 1 September 2026.