# A Colleague, Not a Notebook

JOSH SOLOMON · KNOWN QUANTITY LLC · NOVEMBER 2026

> For about four weeks I have worked alone with an AI on a lot at once: sixty-six small games in ten languages, eleven essays, a company site and some free tools for people dealing with medical bills, landlords and unpaid wages. I call it an empire as a joke. This is what I built around the AI, how it broke, and what I would tell someone starting tomorrow. Most of my failures had the same cause: I believed a report without looking.

## Words worth knowing

*The words this essay leans on, each explained twice: first in plain English, then in one line for the expert.*

### language model

A program trained on a huge amount of writing so it can predict which words should come next, which lets it answer, write and plan. Think of the autocomplete on a phone, grown up enough to write a whole essay.

> FOR THE EXPERT a large transformer-based model used through a chat product with tools, memory files and subagents.

### second brain

A place outside your head where you keep the things you would otherwise have to remember, set up so you can find them again. Think of a kitchen drawer where everything has its spot, compared with a pile on the counter.

> FOR THE EXPERT popularized by Tiago Forte (2022) for personal note systems; used here for a person plus an AI with persistent notes, rules and a task board.

### extended mind

The idea that part of your thinking can happen outside your body. If you never learned your friends' phone numbers because your phone knows them, your phone is doing part of your remembering.

> FOR THE EXPERT Clark and Chalmers (1998), active externalism. Contested by Adams and Aizawa and by Rupert.

### testimony

Knowing something because someone told you. You believe your friend that the store is closed because they were just there and they are usually right.

> FOR THE EXPERT the epistemology of testimony asks when trust in another's word is justified.

### agent

An AI that does a job with several steps on its own, like looking things up, writing a file and checking it, instead of only answering one question.

> FOR THE EXPERT a language model run in a loop with tools and a goal; here, separate workers each given one deliverable and a written brief.

### session

One run of a conversation with the AI. When it ends or crashes, any helper started inside it stops too, and anything not saved is lost, like closing a game without saving.

> FOR THE EXPERT subagents live inside the parent conversation; interrupts, restarts and usage limits end them.

### independent check

Having someone who did not make a thing test it. A second kid checks your homework, because you will read what you meant to write.

> FOR THE EXPERT independent verification and validation; in NASA's usage independence is technical, managerial and financial. Mine is weaker: the same model, started fresh, given only the specification and the result.

**specification**

A short written list of what the finished thing must do, agreed before anyone starts. It is the recipe you check the cake against.

> FOR THE EXPERT acceptance criteria written as yes or no lines, frozen before the build.

**preview**

A private copy of a website change that you can look at before it goes public and throw away if it is wrong. Like a text message you have typed and not sent.

> FOR THE EXPERT per-branch preview deployments; merging deploys production.

**slop**

A lot of stuff made fast that nobody checked. It looks finished and is not right.

> FOR THE EXPERT generated output published without review.

**trusting automation too much**

Trusting a machine's answer because it came from a machine, even when your own eyes say otherwise. Like following the map into a lake.

> FOR THE EXPERT Parasuraman and Manzey (2010); it happens to experts too.

**randomized trial**

A test where a coin flip, not the people involved, decides who gets the new thing, so the comparison is fair.

> FOR THE EXPERT random assignment removes selection bias; the study cited below randomized tasks, not people.

**replication**

Doing someone's experiment again to see if you get the same answer. If only one kid ever saw the ghost, you wait until someone else does.

> FOR THE EXPERT a well-powered direct replication that fails is evidence the original effect was smaller or absent.

**observational study**

Researchers watch what happens without controlling who gets what, so something else might explain the result. Ice cream sales and sunburns rise together, and ice cream does not cause sunburn.

> FOR THE EXPERT no control group; confounding cannot be ruled out.

**survivorship bias**

Learning only from the winners because the losers are not around to tell you what happened. If you only ask lottery winners, buying tickets looks smart.

> FOR THE EXPERT undersampling of failure (Denrell, 2003).

**case study**

A close look at one example. It can show what is possible. It cannot show what usually happens.

> FOR THE EXPERT n = 1, no comparison group.

## Where the idea comes from

Empire is my word for the feeling of having a lot running at once under one small company. I am one person in an apartment in Rhode Island with no employees.

Second brain is the more interesting phrase. In 1998 the philosophers Andy Clark and David Chalmers asked where the mind stops and the rest of the world begins. Their example was a man named Otto, who has Alzheimer's and writes everything in a notebook he carries everywhere. When Otto wants to go to the museum, he looks up the address in the notebook and goes. Clark and Chalmers argued that the notebook is doing for Otto what memory does for everyone else, and that there is no good reason to say his mind ends at his skull.

Plenty of philosophers think that goes too far. Fred Adams and Ken Aizawa argue that being connected to a thinking process is not the same as being part of it: a pencil in a mathematician's hand is not doing mathematics. Robert Rupert argues you lose nothing by saying the mind leans heavily on its tools and still ends at the body. I do not need to settle that. What matters for this essay is the conditions Clark and Chalmers set, because they turned out to be a useful checklist. The notebook is always with Otto. He can get to it without effort. When he looks something up, he just accepts it, the way you accept your own memory. And what is in it got there because he once decided it was true.

The first two are easy now. The third describes a habit, and for my first two weeks I had the habit: I accepted what my system told me without a second thought. That is exactly when it failed me most. The fourth is the one I will come back to at the end.

## What I built

A set of habits, with a little software holding them in place. Five parts.

**One board.** Every project and every task lives on a single page that the AI reads at the start of any work session and writes to as it goes. Each task has an owner, a status, and, when it is stuck, a plain sentence saying what it is waiting on. If the only person who can unblock it is me, it says exactly which button I have to click. Before the board I was the only thing connecting one conversation to the next, and I was bad at it.

**A memory.** A small set of plain notes about me, my projects and the people in them, which the AI reads before it answers and adds to when I say something that will still be true next month. It works like a filing cabinet. I re-explain my life much less than I did, though it has also failed to find a project I knew was in there.

**Written rules.** Every time something went wrong, I said so, and the correction became a standing rule. Do not tell me something is in progress unless you checked this minute. Never send anything in my name unless the text in the box matches what I approved, word for word. Nothing drawn goes to me until someone has looked at the picture. There are about forty of these now.

**A crew, with one lead.** The AI I talk to does not do most of the work. It breaks the work into pieces and hands each piece to a separate worker, one deliverable each, with a written brief. Research goes to one, building to another. The lead decides, reviews and reports.

**A checker who did not do the work.** When a worker says a thing is finished, a different worker, who has never seen how it was made, gets the specification and the result and tries to break it. Only after that does anything go live.

## What worked

The checker worked better than I expected, and the best evidence is from the day I am writing this. Five separate pieces of work went to an independent check today. All five failed the first time.

A free tool that tells renters their rights in every state failed on ten points, including a sentence about Rhode Island eviction law that left out a condition a tenant would need. A page about unpaid wages had twenty-eight legal facts labelled as coming from official sources when they had been read on unofficial copies, one of them twenty years old. A free invoice tool moved the invoice counter every time you pressed a key. An inventory tracker silently erased entries if you had it open in two tabs. And the essay before this one, about quantum computing, came back with eight factual errors and the news that its central idea was already in a paper I was citing.

All five had been reported to me as done, and all five had passed their own tests. Here is where they stand tonight. The renter tool, the wages page and the essay were fixed and are public. The invoice tool was fixed and published, and then a second checker found that the fix itself had problems, so it is being fixed again. The inventory tracker is still not public.

The maker of a thing should not be its only judge. Newspapers have editors for this reason. NASA builds it into how it checks its software: the people checking must use their own methods, answer to a different boss and be paid from a different budget. My checker meets none of those. It is the same AI, started fresh. What it has is a clean slate and orders to go to the evidence: the live page, the actual law, the store's rules. Research on code review at Microsoft found that review catches fewer defects than the people doing it expect, so I think the evidence is the part that matters, more than the second pair of eyes.

There is also direct evidence about these systems. A 2024 study titled "Large Language Models Cannot Self-Correct Reasoning Yet" found that when a model is asked to check its own answer with no outside information, it often fails to fix it and sometimes makes it worse. Asking the maker to look again is not a check, and an AI will produce work faster than I can doubt it.

Three other things worked.

Writing the rule down the moment something broke. My memory of a mistake fades in a day. A rule on the page is still there in a month, and the AI reads it every time.

Small steps that can be undone. Every change to a site goes up as a preview first, gets checked, and can be reversed. Because nothing is permanent, I can say yes quickly. The few things that cannot be undone, such as money, contracts and first contact with a stranger, always come to me first.

One word for yes. I have a nonsense word that means "go ahead on everything you are waiting on me for." It removed a lot of friction, because much of what slows a system like this down is me being asked a question and taking an hour to answer it. It has a cost I should admit. It only covers things that were laid out for me first, such as a post whose exact words I was shown, and it cannot approve spending. But on a tired night I say the word without reading the list as carefully as I should, and then the careful list is doing less than it looks like.

## What did not work

I keep a log of the failures bad enough to write down. It has thirty-one entries from nineteen days, and it is not complete: two of the stories below never made it in, and about a third of the entries were added later by an audit, not on the day. A fair essay shows some of them, including the ones that happened more than once.

**I was told things were happening that were not.** In my second week I asked about a piece of game art twice and was told twice that it was in progress. The helpers working on it had crashed and nothing had been saved. The AI was not lying in any human sense. It was reporting from its memory of having started the work, and it had not looked. The effect on me was the same as a lie. This is the failure I would warn anyone about first, because it is quiet and it feels like everything is fine.

**Work vanished, more than once.** Workers ran overnight, the session ended, and nothing they made had been saved anywhere outside it. I made a rule about saving. Helpers still died several more times in the following week, including when I pressed stop on something unrelated.

**Things were sent wrong.** A post went out with text that was not the text I approved. A personal thank-you email went out from my game studio's address instead of my own.

**The art was bad and nobody looked.** I rejected a logo three times in one night. It was a sloth playing a guitar, and the guitar had two strings, and the hand was a blob. It had passed every check, because every check was about the code that drew it. Nobody had looked at the drawing. The log has six entries like this one.

**Stores rejected my games.** One game site turned down several of my games, one of them twice, and dropped another after its first week of numbers. Google rejected one because it crashed on opening, and another because the listing looked like it was for children when I had said it was not. Both Google rejections came after my checker was in place. It checked what I told it to check, and nobody had told it to open the game on a real phone or to read the listing the way a reviewer would.

**My photos went up when I had asked for something else.** I asked for my likeness to model a clothing line. My actual personal photos were published on the site. They came down the same day, but they should never have gone up.

**The board itself was wrong.** Several entries are the board asking me to do things that were already done, or marking a task finished with no evidence. The tool I built to stop false reports made some of its own.

**I wrote rules that did nothing.** I wrote a careful quality system with an hourly watchdog and a nightly audit. Five days later a check by hand found that neither had ever existed. I had written "runs every hour" in a document and treated the sentence as if it were the machine. The same mistake turned up four days after that with a monthly rule. And the nightly audit that I did set up could not reach my code when it ran, so it skipped those checks. That entry in my log is still open.

I also made plain mistakes of judgment. I tried running two AI accounts, one personal and one for the business, to keep the accounting separate. I spent a day carrying messages between them by hand, then cancelled the second one the same evening.

## What I should have done

"What someone can do," below, covers the board, the evidence rule, the checker and saving. Three more things belong here.

Turn every rule into a check or delete it. If a rule says "every night," there must be a thing that runs every night, and I must be able to see that it ran.

Look at the thing. Open the page. Zoom in on the hands. Read the email in the box. Testing does not replace a person looking at what a person will see.

Judge the work by where it is going. Before sending anything to a store, a publication or a stranger, compare it to what that place already accepts.

## What you should know about me before copying any of this

I use Claude, on the largest plan sold to individuals, and I have been at this most of every day and many nights. I came in with years of work in data, reporting and project management, which is most of what running a crew of workers is. I have low expenses and no children, so a month of this was a risk I could take. Some of what went well was luck. I have also drawn a line around what I do not hand over: money,

anything legal, and the first message to a person who has never heard from me. And the tools I publish for people with real problems, like a wage claim or an eviction notice, can hurt someone if a fact is wrong, which is why those get the strictest checking and still say on the page what could not be confirmed.

## The objections

**This is just a flood of slop.** It is the strongest objection and it is partly right. One analysis this year estimated that about half of all new articles published on the web in English are now mostly written by AI. That comes from a marketing company that used programs that guess whether a text was written by AI, so treat it as a rough picture. A person with these tools can produce far more than they can check, and most people will not check. The tools are not my answer to that. The output is only as good as the judge, and I have told you how often my own judge failed on the first pass today. If you copy one thing from this essay, copy the checker, and then go look anyway.

**You are getting worse at the things you handed off.** Possibly, and the evidence is thinner than the headlines. The famous 2011 "Google effect" paper is shakier than its reputation: its first experiment failed when a large replication project repeated it. The part that has held up better is narrow: when you expect something to be saved, you remember it less well yourself. The most worrying result I know of is from Poland. After experienced doctors started using AI to find precancerous growths during colonoscopies, the share of exams where they found one without the AI fell from about 28 percent to 22 percent. It is one observational study: it watched what happened and did not run a fair test, and other things could explain it. I used to be good at logistics and I am likely worse now. What I have tried to protect is deciding what is worth doing and knowing when something is wrong. I cannot prove I have protected it. The mistakes I catch are mostly ones of taste or truth, but I only know about the ones I caught, which is the survivorship bias from the word list.

**It only looks like productivity.** There is real evidence for this worry. In a careful experiment in 2025, sixteen experienced programmers worked on 246 real tasks in projects they knew well, and a coin flip decided which tasks allowed AI. With AI they took 19 percent longer. Afterward they still believed it had made them 20 percent faster. The same group ran a larger follow-up this year and found signs the tools now help, but called its own evidence very weak, partly because so many programmers refused to work without AI that the test could no longer compare fairly. Across other studies, AI helps beginners most, on jobs with a clear right answer, and experts least on work they know well. I take that seriously, because believing I was moving faster than I was is the mistake I made about my game art. So I try to count only things a stranger could verify: is it live, does it work, did the store accept it.

**It all depends on one company's product.** True. I keep the rules, the notes and the work in plain files I own, so I could carry them elsewhere. I also built a small assistant that runs on my own laptop. It is far weaker and gets facts wrong, but I know what I would have left.

**One person's month proves nothing.** Correct. This is a case study. Among the tens of thousands of US startups that use Carta, a larger share have a single founder than before, about 36 percent in the first half of 2025 against about 24 percent in 2019, and nobody has shown that AI is the cause. The solo founders you hear from are the ones it went well for.

**So is it really a second brain?** The fourth condition settles it for me. Otto's notebook only holds what Otto once decided was true. My notes mostly pass that test, because they hold things I said. The agents do not. They hand me new things I never agreed to, which makes them less like a notebook and more like a person telling me something. Philosophers call knowing something because you were told it testimony. You trust a person the way you trust a witness: by their record, and by checking the parts that matter. So what I have is a fast colleague with a good filing system and no reliable sense of when it is wrong. Treating it as a notebook is how most of the failures in my log began. Andy Clark himself wrote about generative AI and the extended mind in 2025; I have read only the summary of that paper, so I will not put words in his mouth.

## What someone can do

You do not need my setup or a paid plan. A free chat app and one document will do.

1. Keep one plain document with everything in progress. Paste it in at the start of each chat. At the end, ask for the updated version and save it yourself.
2. Make it a rule that every status comes with proof you can open: the page, the file, the sent message.
3. Before anything goes out, open a fresh chat, give it only what the thing was supposed to do and the finished thing, and ask it to find what is wrong. Then look yourself.
4. Keep a list of your own corrections, written down when they happen, at the top of that document.
5. Keep a short list of things that are never done without you: money, signatures, deletions, and anything sent to a person who has not heard from you before.
6. Save the work somewhere outside the chat every twenty minutes. Treat anything unsaved as not done.

Then start smaller than you want to. One project. Let it break. Write down how it broke. My log of that has thirty-one entries. It is the part I trust most, and it is still missing things.

## Sources and a note on method

The extended mind is from Andy Clark and David Chalmers, "The Extended Mind," *Analysis* (1998). The objections are from Fred Adams and Ken Aizawa, "The Bounds of Cognition," *Philosophical Psychology* (2001) and their book of the same name (Blackwell, 2008), and Robert Rupert, "Challenges to the

Hypothesis of Extended Cognition," *Journal of Philosophy* (2004). Clark's later paper is "Extending Minds with Generative AI," *Nature Communications* (2025), read in summary only. "Second brain" in its popular sense is from Tiago Forte, *Building a Second Brain* (2022).

On checking: Jie Huang and colleagues, "Large Language Models Cannot Self-Correct Reasoning Yet," ICLR 2024, arXiv:2310.01798. NASA's independent verification and validation program describes independence as technical, managerial and financial. On code review: Alberto Bacchelli and Christian Bird, "Expectations, Outcomes, and Challenges of Modern Code Review," ICSE 2013. On trusting automation too much: Raja Parasuraman and Dietrich Manzey, "Complacency and Bias in Human Use of Automation," *Human Factors* (2010).

On productivity: the 2025 randomized trial is from METR (Becker, Rush, Barnes and Rein), "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity"; the follow-up is METR's February 2026 update, which describes its own evidence as very weak. The pattern of who gains most comes from Brynjolfsson, Li and Raymond on customer support and Dell'Acqua and colleagues on consultants.

On memory and skill: Sparrow, Liu and Wegner, "Google Effects on Memory," *Science* (2011), and the replication project in Camerer and colleagues, *Nature Human Behaviour* (2018), which repeated only the paper's first experiment; Risko and Gilbert, "Cognitive Offloading," *Trends in Cognitive Sciences* (2016); the study of endoscopists is Budzyn and colleagues, *Lancet Gastroenterology and Hepatology* (2025), which I have read only in summary.

On volume: Graphite, "AI now writes as many online articles as humans do" (May 2026), an analysis by a marketing firm, not peer reviewed. On solo founders: Carta, "Solo Founders Report 2025," which covers Carta's own customers. On learning only from winners: Jerker Denrell, *Organization Science* (2003).

Everything about my own work comes from my project board and its incident log. The counts were taken from that log on the day of writing. The photos story, the two accounts and today's five failed checks are from my own account and are not in the log. I have left out the names of the people and companies involved where they are not mine to name.

What I could not check: several of the older papers above were cited from standard references and not reopened, and two were read in summary, as marked. I found no good public data on what it costs to run a business this way or how such businesses do, so I have made no claim about either.

I used AI to research and draft this, under my direction, with a standing instruction to tell me where I was wrong before telling me where I was right. A draft was then handed to a reviewer told to argue the other side. It gave the draft a B minus and found, among other things, that I had misread the philosophy I was building on. This version is the corrected one.