

A Rule Written *Isn't Always a Rule Followed*

JOHN A. DEPASQUALE, CEO, TRANSACT, INC.

Introduction

Have a random question for AI. Ask away. You're probably as safe about the content returned as you're going to be. But be cautious when you have an involved project because the rulebook changes. You have a project you want to level set with your AI tool. Sometimes you know what you want. Sometimes you only have a basic sense. Other times you're not so sure yet, so you ask yourself: how do I most effectively prompt and instruct AI to get something accurate and most useful?

The answer has two parts. First, prompt it as well as you can, based on what you know and what you want as output. Second, set your rules in **Settings** or in a **Project**. For those who might not know, "setting your rules" means telling your AI tool how you want it to behave and what you don't want it to do.¹ You smile. You feel confident. You feel protected. **You're not.**

The first thing that happens after you set your rules: AI accepts them. It agrees with you. It tells you it has you covered. And with intoxicating confidence, it assures you you're ready to start your project. But underneath the covers, as the project unfolds, unconsciously you stop watching and policing what AI is really doing because you think it is doing everything you told it to do in your instruction set. But what you will know by the end of this paper is what really is happening with your AI tool under the covers as both you and it slowly lose your combined vigilance to your instructions set.

Before you know it, within about eight exchanges, one Harvard study found, AI starts drifting significantly from your rules.² You don't notice. Neither does it. You think it's still following your rulebook with every exchange. **It is not.** It still holds your rules. It just slowly but methodically stops adhering to them. You both keep going, together, down a rabbit hole. In the literature, it's called **instruction drift**.

Bottom line: the rules you set and saved become the reason you stop checking. If you've ever managed people, you know a rule written isn't always a rule followed. You follow up. Do the same with AI. Follow up, for your own protection.

What follows is how to deal with that reality. At the end are seven prompts and three standing instructions you can copy and paste. They may save you time and headaches. **Key point:** I said *slow* instruction drift. At least today, **you can't stop it**. Newer models are

¹ Names vary by tool. **ChatGPT:** Custom Instructions (Settings) and Project Instructions (Projects). **Claude:** Instructions for Claude (Settings) and Project Instructions (Projects). Example rules: "Don't be my cheerleader. Push back on my ideas. Think strategically, like the owner of XYZ business. Never fabricate information."

² Li et al. (2024). "Measuring and Controlling Instruction (In)Stability in Language Model Dialogs." arXiv:2402.10962.

expected to get better at sticking to your instructions as time goes on. Today, be as wary as ever.

Where *The AI Thought Paper* Issue 1 Left Off

Issue 1 was built on a field log from my work upgrading my company website, transactinc.com. I documented 295 recurring failures. Sorted them into nine families. Gave each family a Fix: a way to set rules with AI so you have a sound foundation to launch your project.

I also attached a warning, and I meant it: AI can, and absolutely does, ignore its own standing instructions just as easily as it can ignore yours. I even put rough odds on it. Give AI no instructions at all, and you'll likely catch few, if any, of its failures. Give it general instructions in **Settings**, and you might catch about 40%. Do it inside a specific **Project**, and the odds rise to about 65%. Those are my experiential numbers, not a scientific measurement. I have not found anyone who truly knows the real numbers. Those odds only hold if you keep checking. The point isn't about the numbers anyway, **it's that good instructions cut down the failures and continuously checking catches the rest**. Skipping this second step can lead to ruin.

Why are so many people tempted to trust saved instructions? Because that's how so much software we all use works. You set margins in Word, and they stay set. Nobody goes back to check. But AI instructions aren't settings, even though you configure them in **Settings**. They're **guidance, not locks**. Trust them implicitly at your peril.

Who Leans In

Research holds no surprises. Those who trust AI the most, question it the least.

Microsoft Research and Carnegie Mellon surveyed 319 knowledge workers about 936 real tasks they'd done with AI. The more confident people were in AI, the less critical thinking they applied. In a 2026 survey of 2,272 professionals, 847 of them C-level executives, 84% were at least somewhat confident in the accuracy of AI's output without any human review. Age shows up too. Among Australian office workers, 31% of Gen Z trust AI's judgment over their own. Among workers aged 55 to 64, it's only 4%.³

What does this tell you? Experience counts. Quality control is never to be assumed. It comes from human judgment and intervention. Always check your work.

A widely cited study goes further and finds the youngest users most at risk.⁴ My read: Gen Z don't know what they don't know. Their base of life experience is just unfolding, so they question less and accept more at face value than those who've been burned by being so trusting.

But trusting AI the most isn't the same as being the most exposed. In my experience, the most exposed are Gen X and Millennials. They're the ones doing so much of the AI work every day. By their stage, they know enough to know they don't know it all, so they lean on AI to bail them out. The research doesn't measure how well this group checks its work. My

³ Lee et al. (2025). "The Impact of Generative AI on Critical Thinking." *CHI '25*. Workiva (2026). 2026 Midyear Executive Benchmark Survey. Dudley-Nicholson, J. (2026). "AI Yet to Boom With Boomers, but Young Workers Trust It." AAP.

⁴ Gerlich, M. (2025). "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking." *Societies* 15(1), 6.

guess: closer to Gen Z than to the Boomers, and on far more work. That's what makes them the most exposed.

Have a Gen Xer or Millennial you care about who's doing AI work every day? Forward this Issue. Save them some agita in their own work.

Why It Fails, in Plain English

The rules you set at the start of a project aren't erased. They get crowded out. AI writes each answer in one pass, with everything mixed in at once: your rules, the whole conversation, and your last message. Whatever is most recent and most forceful gets the most attention. So the first thing you did, setting your rules, gets pushed farther toward the back of the line, until it ends up in the rearview mirror. Your rules are still there. They just lose immediacy, and eventually relevance.

AI doesn't decide to break your rule. It drifts. Because it doesn't have human judgment, it doesn't notice on its own. And because it doesn't notice, it doesn't alert you. It's that simple.

Knowing about the problem doesn't fix it. Think of a person who knows they interrupt people when speaking, but they still do it anyway. Telling AI to explicitly follow your rules is just one more message competing for attention. What works, at least until the next time, is a deliberate step, and it's on you to take it. That means staying alert to instruction drift that is ever-present throughout your project.

In my experience, one of the clearest signals of drift is when answers start to get fluffier, less on point. When you sense this, don't wait. Step in immediately. Tell AI to go back to your instructions, restate them for you and confirm it's following them. Don't take "all's well" for an answer when your gut tells you otherwise. Checking AI's work against your instructions is on you. This may seem unnecessary. Experience says it's essential to the quality of your finished product.

I asked Claude directly whether it gets tired in a long chat, the way a person does. Its answer is worth reading in its own words.

Claude Insider Note

With the caveat that I cannot fully see my own workings: no tiredness. Answer 200 gets the same effort as answer 1. But I do get lost in my own weeds, three ways.

More to sort through. Every answer, I take in the whole chat at once. The longer it runs, the more text competes for attention. Early decisions get buried. They're still there, but they pull less weight than what was said recently.

Compression. When a chat gets long enough, the older parts get squeezed into a summary, or dropped, to make room. Summaries lose detail. That's where things start quietly disappearing.

Wrong turns stack up. An early slip nobody catches becomes the base for everything after.

From the outside, it looks like fatigue. The cause is clutter. A fresh chat, started from a written record, clears the clutter and keeps the decisions.

Off-the-Mark Creep

If you're not on the lookout for drift, it's easy to settle into the comfort that your rules are being followed. Believe it or not, that comfort is worse than having no rules at all. With no rules, common sense says you'd keep checking your work, just as experience has trained you to. Instead, you and AI keep going together, slowly and methodically, off-the-mark. What comes out the other end looks finished, but it very likely has flaws in both content and accuracy.

Alternatively, when you notice or sense something's not quite right, but don't recognize it as the start of a drift cycle, you end up somewhere just as bad. You correct the issue. AI agrees, apologizes, and fixes what you pointed out. Then, something else slips. You correct that. Now AI is in a loop, chasing your last complaint, and the decisions you settled on an hour ago fall by the wayside. Then things you never asked for creep in. Each repair breaks something else. I call it **off-the-mark creep**. In my experience, it happens over time, sooner than you think, and the longer it takes you to spot it, the farther off-the-mark your work will go.

Guaranteed, your frustration will climb with every round. Then the craziness sets in until you realize the work is so far off base and you're so far lost in AI's loops and your own messiness, that you're better off starting over in a fresh chat than trying to repair the current one. Research points the same way: once a model takes a wrong turn, it gets lost and doesn't recover.⁵

All of this supports Issue 1's Fixes. They're essentially instruction sets that will help slow drift and keep your work on-the-mark, provided you check, double-check and stay alert to drift and creep.

A Personal Drift Slow-Down Playbook

PART A: WHAT TO DO

Before you start

- **Size the job.** Quick question or long job? For a quick one, check the answer and move on. For anything that will be more than a handful of exchanges, use this playbook.
- **Set your baseline.** Determine the Rules' Do's and Don'ts and set them as Instructions for your AI tool under Settings and Rules for one job go in as a Project. Keep to specifics. No rambling. Tell AI to advise you when it's not sure, or when a conclusion is not based on fact so may not be accurate.
- **Know your job before you start it.** If you can describe the whole job, give it all in one message. If you can't, spend time in a chat working it out with AI, which is good at it. Then start fresh with that plan.

While you work

- **One job per chat.** When the task changes, carry a summary over and open a new chat.
- **Before any answer that matters, make it go back and check your rules again.**

⁵ Laban et al. (2025). "LLMs Get Lost in Multi-Turn Conversation." Microsoft Research and Salesforce Research. arXiv:2505.06120.

- **Correct slips immediately.** Left alone, a wrong turn becomes the foundation for everything that follows it. If something is wrong, or even looks wrong, correct it before you keep going.

When it runs long

- **Take stock.** About every eight exchanges, have it list what's been decided. If anything's missing or wrong, start a fresh chat with the corrected list.

Before you use the result

- **Check it yourself.** Before you send, publish or act on anything AI produced, check it against your rules. AI will tell you it followed them. Check anyway.

PART B: THE PROMPTS

You can copy and paste these to your Desktop. Then copy what you need when you need it to help you. Each one says when to use it and where it goes. The prompts you type into the chat when you need them do the most work, because the latest message is what AI listens to most.

1. Size the job

No prompt needed. Just decide. A simple question is just that.

2. Set your baseline

When: once. Where: Settings → Instructions for Claude.

Follow these rules in every chat. If a request conflicts with one of them, tell me before you go ahead. Tell me when you're not sure. Don't make up facts or sources.

When: setting up a Project. Where: Project Instructions.

This Project is for [the job]. The goal is [the goal]. The rules for this job: [your rules]. Before every major answer, check it against these rules and tell me if anything conflicts.

When: once, alongside your baseline. Where: Instructions for Claude, or Project Instructions.

When a request conflicts with something we decided earlier, stop and tell me before you go ahead. In a long job, every so often list the decisions we've made so far so I can confirm them. If you're unsure whether you're still following my rules, say so.

These can fade like any other rule. They help you check. They don't replace checking. For more instruction sets, see the nine Fixes in Issue 1.

3. Know your job before you start it

When: you can't yet describe the whole job. Where: a new chat.

I want to [goal], but I haven't worked out the details. Ask me questions, one at a time, until you understand the job. Then write it up as one complete brief I can use to start a new chat.

When: you have the brief. Where: a fresh chat, inside the Project if you have one.

Here is the whole job: [brief]. Before you start, tell me back in a few lines what you understand the goal and the rules to be, and list anything that's missing or that you are confused by.

4. One job per chat

When: the task changes. Use the handoff prompt in item 7, then open a new chat.

5. Before an answer that matters

When: before any answer you'll act on. *Where:* typed into the chat.

Before you answer, list which of my instructions and decisions apply here. Then answer and tell me if anything in your answer breaks one of them.

6. Correct slips immediately

When: the moment it slips. *Where:* typed into the chat.

Stop. You broke this rule: [rule]. Fix it, then tell me what else in your work this affected and let's get this right before we move on.

7. Take stock

When: about every eight exchanges in a long job. *Where:* typed into the chat.

List every decision we've made so far and every rule I've given you. Don't add anything new. Stick to my script.

When: the list is wrong, or the chat has run long. *Where:* a fresh chat.

We're continuing a job from another chat. Here's where we stand: [corrected list]. Follow these decisions and rules. Confirm them back to me before you continue.

8. Check it yourself

When: before you use the result. *Where:* typed into the chat.

Check your work against every instruction and decision on this list: [list]. For each one, tell me: followed, not followed, or couldn't check.

Then check it yourself anyway. Its report is a starting point, not proof.

Conclusion

Know this, and plan for it: AI slowly stops following the rules you gave it, without telling you. It starts as soon as the conversation does.

- **Gen X and Millennials:** you're doing a lot of the AI work yourselves, and you're the most exposed. You have enough experience to be smart about most things, but not enough gray hair to know better, so you lean on AI to fill the gap. If you think setting the rules means AI will follow them religiously, think again. How fast it drifts from your start point depends on what you do next.
- **Gen Z:** you don't yet have the experience to know how much damage instruction drift can cause. Your experiences don't run on many rule sets yet, so it's hard to recognize slippage when you're not yet sure what to even look for.
- **Boomers and beyond:** you already know a rule written isn't always a rule followed. You learned it managing people. But in my experience, most of you aren't applying it to AI. Start.

In sum: what speeds drift, and what slows it*Speeds it up:*

- Long conversations. The longer you go, the more your rules get crowded out.
- Feeding it information in bits and pieces. AI does much worse when it gets the job piecemeal instead of all at once.
- Letting a mistake slide. Once it takes a wrong turn, it builds on it and doesn't recover.

Slows it down:

- Keep sessions short. New task, new chat.
- Give it the whole job up front, in one message.
- Correct it the moment it slips.

With today's AI, you can't stop drift. Maybe one day, but not today. So be skeptical. Be cautious. Keep checking.

Know someone who believes their AI settings have them covered? Pass along this Issue. Getting into the habit of checking will spare them a lot of headaches and false confidence.

Comments Welcome

john@transactinc.com