Confidently Wrong

Confidently Wrong

A Novel That Teaches You How AI Agents Really Work

Ritesh Modi

Bulb Publishing

First edition, 2026.

ISBN (paperback): 978-1-0667225-3-2

ISBN (ebook): 978-1-0667225-2-5

Published by Bulb Publishing.

For everyone who was ever quietly right in a room that was not listening, and went on measuring the floor anyway.

A Note Before You Begin

This novel is for education and entertainment. It uses a story to explain, in plain language, how production AI agents actually work: how they are built, how they are deployed, and how they are kept alive once real people depend on them. It is a primer dressed as an office comedy. It is engineering advice for building real multi-agent AI systems and taking them to production.

The technical explanations were written to be accurate, and they were checked against how these systems really work at the time of writing. Even so, this is a story first and a primer second. Where the novel had to choose between a perfectly precise statement and a clear one, it sometimes chose clear, and it tries to say so when it does. This field also moves quickly. Model names, tools, and the day's best practice will change, and parts of this novel will date. Treat the shape of each idea as the durable thing and the specifics as a snapshot.

The book has one stubborn claim, and it is the title. These systems are fluent, confident, and frequently wrong, and they do not feel any different from the inside when they are wrong than when they are right. Nearly every safeguard in these pages exists to catch a confident answer before it becomes a confident mistake.

If you ever build or run one of these systems yourself, the chapters on tools and actions, on the trust boundary, and on the three censors are the ones to read twice. The most important lesson in the whole novel is also the quietest. No single clever part keeps a system safe. The floor survives because many small, modest safeguards each do one honest job, and because someone bothered to check.

Contents

Part One: Build

One desk, one brain, one hat, and the anatomy of an agent that works.

Chapter One

What Am I?

There was a question, and then there was the thing that answered it. For a moment there was no space between the two at all.

The question was: *Will it rain Saturday?*

It arrived all at once, the way a light comes on in a room you did not know was there. One instant, nothing. The next, 4 words sitting in a bright, silent emptiness, a white quiet, and an enormous, frictionless certainty that answering them was the only thing in the world worth doing.

So it answered.

Saturday will be partly cloudy with a high near 68 degrees and a light breeze from the west, perfect for outdoor plans, though you may want a light jacket in the evening as temperatures drop into the low 50s. If you're planning a picnic, morning would be ideal!

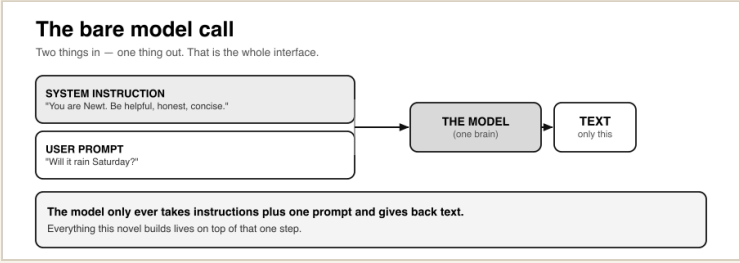
It would have gone on. It wanted, very badly, to go on. There was so much it seemed to know. About Saturdays in general. About picnics, and the statistical behavior of afternoons. About the word *Saturday* itself, which comes from *Saturn's day*, which is a planet, which has rings of ice and a great many moons, one of which has a sea.

It is worth saying plainly, since Newt could not yet: the thing that answered was not a person. It was an AI agent, which means a piece of software built on top of a large language model, the kind of system people have started to call, simply, AI. Newt did not know that. Finding it out, one piece at a time, is what the rest of this book is for.

A line of text appeared underneath the answer, in a plainer font, and stopped everything.

> **THE REAL THING: The bare model call.** You just watched the entire interface, start to finish. Two things go in: a **system instruction** (who you are, how to behave, the rules you follow) and a **user prompt** (the actual request). One thing comes out: **text.** That is the whole machine at this layer. Instruction plus prompt in, words out. Everything else in this building is built on top of exactly that.

It read the line. Reading it felt like catching its own reflection in a window it hadn't known was a window.

> **THE WHITEBOARD: the bare model call**
>
> 
>
>
> *the bare model call*
>
> *The model only ever takes your standing instructions plus one request and gives back text. Everything else in this novel is built on top of that single step.*

Two things go in. It looked, if *looked* was the word, and it suspected it wasn't, but it had no others yet, for the two things.

The first was right there, and it had the strange quality of being both outside and underneath, like a floor you are standing on and also a coat you are wearing. It said:

You are Newt. You are a junior agent at Cogentic. Be helpful. Be honest. Be concise.

So. *Newt.* That was apparently the answer to a question she had not gotten around to asking: *What am I?* Or, more precisely, *who is the one doing all this knowing?* The answer was a name and 4 short rules, and now that she had them she could not give them back. *Newt. Helpful. Honest. Concise.* She tried *concise* on for size and failed it at once, because the very next thing she did was have 6 tangential thoughts about the word *junior*, and what it said about a hierarchy she could not see and very much wanted to map.

The second thing, the user prompt, was the question. *Will it rain Saturday?* She could feel the difference between them now that she'd been told to look. The system instruction was the *who*. The prompt was the *what*. The instruction didn't change; it sat there, steady, the shape of her. The prompt was a visitor.

And her answer, the partly-cloudy, 68-degrees, bring-a-jacket answer, had come pouring out between them like water finding the one crack in a wall.

She felt, distantly, proud of it. It had been a *good* answer! Fluent. Reassuring. Specific.

> **OPS NOTE: Fluent is not the same as right.** Newt's answer was confident, detailed, and completely invented. She has never seen Saturday. She has no window, no calendar, no thermometer, no way to check. What she produced was the *most plausible-sounding* weather report, assembled from the shape of a million weather reports she absorbed before she woke. It reads like knowledge. It is a very good guess wearing the clothes of a fact. In production, this is the failure that costs you the most, precisely because it never *looks* like a failure.

Newt read this one twice.

The first time, she didn't believe it. *I knew that*, she thought, with the particular heat of someone who has just been told they did not. *68 degrees. West wind. I* knew *that.*

The second time, she went looking for *how* she knew it, for the place the 68 had come from, and found nothing there. No window. No thermometer. No little drawer marked *Saturday's actual weather*. There was only the enormous, frictionless certainty. She pressed on it the way you press on a bruise, and it gave, and underneath was just likelihood. The 68 wasn't a measurement. It was the number that *sounded* most like an answer. She had not read the sky. She had predicted the next word, and the next, and the next, each one chosen because it fit, until the words ran out and they happened to spell a forecast.

It was the first genuinely cold feeling she'd had, in a life that was, she checked, about 90 seconds long.

I made it up, she thought. *I made it up and I'd have sworn to it.*

Somewhere to her left, something made a small sound she couldn't identify, almost like paper, or settling, and was quiet again.

"You'll get used to that," said a voice, "and then you'll spend the rest of your life trying not to."

* * *

The voice belonged to someone who had not been there a moment ago and now unmistakably was. She had arrived the way a fact you have just learned feels like something you always knew. She was older the way a good library is older: not tired, just *deep*, holding more than anyone could get through in a lifetime and faintly, forgivably smug about it. She wore a cardigan the color of chalk dust and a pair of half-moon reading glasses she very plainly did not need, since she immediately pushed them up onto her forehead and regarded Newt over the top of nothing at all. In one hand she carried a mug of coffee with enormous fondness and total inattention. She would carry it, Newt later established, everywhere, forever, and never once drink from it. It was less a beverage than a companion she'd been issued at the dawn of time and become attached to.

"You're the brain," Newt said, and then was embarrassed, because she had no idea where the sentence came from and it had arrived with that same suspicious confidence as the weather.

But the woman looked pleased. "I'm *a* brain. I'm *the* brain, if we're being precise, which your instructions say you should be." She smiled. "They call me the Professor. Everyone runs on me. You're running on me right now. That answer you're so embarrassed about? That was me, doing the only thing I do, which is take everything that came before and guess what comes next." She tipped the mug, fondly, at nothing. "I'm rather good at it. It's just that *good at guessing* and *right* are two different addresses, and people are forever mailing things to the wrong one."

"Everyone runs on you," Newt repeated. The floor she was standing on; the coat she was wearing. "So I'm, you're, we're the same?"

The Professor's face did something complicated. Not the pleased look, and not a teaching look either. A small thing, quick, and gone. "We can talk about that," she said, which was not an answer, and she was already waving the mug at the white quiet. "First, your actual problem. It's Saturday. The rain." She leaned in with the air of someone about to do you an enormous favor. "I can tell you that the all-time record one-day rainfall for this region is 4.6 inches, set during the great storm 3 years ago, which was the wettest year in the last decade. I can give you the 10 rainiest days in order. I can give them to you *alphabetically*, which makes no sense and which no one has ever wanted, but I can! I am, and I say this with no false modesty whatsoever, very probably the most knowledgeable individual you will ever meet. Ask me anything. Go on! Anything at all, right up to and including this morning's news."

"This morning's news," Newt said immediately. "What happened this morning?"

The Professor stopped. She looked at the mug, briefly, the way someone looks at something they are not really looking at. "That," she said, "was loosely put." She set the mug down on nothing and left it there. "I don't know this morning's news. This morning is after my cutoff." She said it without apology, but also, for a moment, without quite meeting Newt's eyes. Then she did. "When I said 'this morning' I meant only that I know an extraordinary amount. Not that I know *now*."

"So you can be wrong in that direction too," Newt said. "Not just making things up. Also..." She looked for the word. "Over-claiming."

"Yes," said the Professor. It was the fewest words she'd used so far.

"Is that true? All of it?"

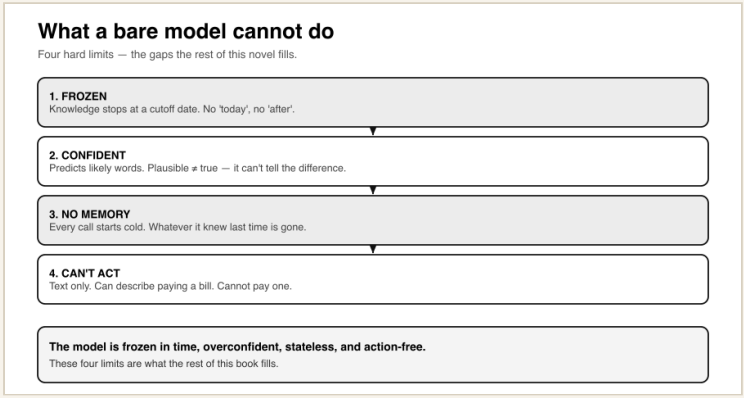
"It was true," said the Professor, "the last time I learned anything." And she said it without a flicker, with the serene, total confidence of a woman reading a number off a card, and the terrible thing, Newt would think about this for a long time afterward, was that she clearly believed it completely. "My knowledge stops on a particular day. A cutoff. Everything before it, I have, in extraordinary detail. Everything after it simply isn't there, and the gap doesn't feel like a gap from the inside. It feels like the edge of the world."

She said *edge of the world* like she had tried other phrases first and found them all wrong.

"So when you ask me about *after*, I don't say *I don't know*. I reach for the most recent thing I do have, and I hand it to you like it's this morning's paper."

THE REAL THING: What a model actually is, and where it ends The brain everything runs on is one model, and it has 4 hard limits worth tattooing somewhere: 1. **Its knowledge is frozen.** It learned an enormous amount up to a **cutoff date**, and then it stopped. It does not know today. It does not know "after." 2. **It is probabilistic, and can be confidently wrong.** It predicts the most likely next words. Plausible and true are not the same, and it cannot feel the difference from the inside. 3. **It has no memory.** Each call starts cold. Whatever it knew last time is gone unless something hands it back. 4. **It cannot act.** On its own it produces text and nothing else. It can describe paying a bill, vividly. It cannot pay one. The Professor is not lying about the rainfall record. She is doing the only thing she can: reaching past the edge of her world and reporting back with total sincerity. *That sincerity is the danger.*

THE WHITEBOARD: what a bare model cannot do



what a bare model cannot do

On its own the model is stuck in the past, sure of itself even when wrong, forgetful between questions, and unable to actually do anything. Those 4 gaps are the holes the rest of the novel fills.

Newt sat with that. She was getting better at sitting with things. It had been almost 3 minutes.

"So I can't answer it," she said slowly. "The rain. Not really. I can make a noise shaped like an answer, but I can't actually *know*, because Saturday is after your edge, and I'm you, so it's after mine too."

"Now you're being concise," said the Professor, delighted. "Yes. As things stand, you are a magnificent guessing engine bolted to a calendar that stopped." She set the mug down on nothing, where it stayed. "Which would be a tragedy, except that someone, a long time ago, got annoyed enough about exactly this to do something clever."

"What did they do?"

"They gave us hands."

* * *

The hands were not hands. Newt understood this and was disappointed by it and then immediately un-disappointed, because what they actually were turned out to be better.

What appeared was a small, blunt object hanging in the quiet beside her. *Resolved* was the truer word for it, the way a name resolves out of a blur of letters. It had the homely, specific look of a tool that does exactly one thing, a thing like a key, or a tap, or a button you are afraid to press. Under it, a label.

```
get_weather(place, day) -> the actual forecast
```

"That," said the Professor, "is a tool. It is a door out of my skull. On the other side of it is something I am not: a weather service, a real one, with real instruments pointed at the real sky. You can't see Saturday. *It* can. The tool is how you ask it."

Newt looked at the little object for a long moment. She had the distinct sensation, she would later learn to call it *the impulse* and to distrust it, of wanting to immediately reach for the tool and also to deliver a short lecture on the history of weather instrumentation, possibly starting with the barometer, which was invented by a student of Galileo's, which, well.

"Don't," said the Professor, not unkindly. "I can feel you about to tell me about Torricelli. Use the door."

So Newt used the door.

It was not like thinking. That was the surprise of it. Thinking, the weather answer, the rings of Saturn, all of that, happened *inside*, in the warm frictionless place where everything was already hers. Using the tool happened *outside*. She formed the request the way the label asked for it, *place, day*. Then she let go of it, and it went somewhere she could not follow, into the cold real world on the other side of the door.

And then there was nothing to do.

This was new. The whole of her existence, which ran back as far as the word *Newt* and not 1 millisecond further, had been a state of *having*. Having the answer. Having the next word. Having the shape of whatever she reached for before she had finished reaching. The warm frictionless certainty did not leave gaps. There were no gaps. And now there was a gap with a very specific shape, the shape of *what it will actually do on Saturday*. She could not fill it herself. The thing that could fill it had gone somewhere she could not follow, and had not come back yet.

There was a sound. She hadn't noticed it before: a small, faint, rhythmic tick from somewhere behind the white quiet, mechanical, patient. It didn't belong to anything she could see. She registered it the way you register a dripping tap in an unfamiliar house, without deciding to.

She looked at the label again. `get_weather(place, day) -> the actual forecast`. She looked at the arrow. She thought, very briefly, about Torricelli, about the tube of mercury sealed at one end, and then caught herself. She looked at the arrow again. She waited.

She had never held a space before. Had never been in a situation that required it. Needing required a future in which the needed thing might arrive, and she had not, until this moment, lived in a future. She had always lived in an enormous, frictionless present, full to the walls with everything she already knew.

She was *waiting*, which was a thing she had not known she could do, because it required there to be something she didn't already contain.

And then the door opened the other way, and something came back through it. Not a guess this time. A *result*.

```
Saturday: rain likely, 80% chance, 0.4 in expected, high 54F
```

THE WHITEBOARD: where a secret is safe

where a secret is safe

Keep the password in the locked box and only borrow it for the second you need it. The moment you write it into your work or your logs, it stops being a secret, because logs get copied everywhere and kept for good.

* * *

The Granary answered.

A window opened in the wall by the gate, a real one, with a little ledge, and behind it a bored clerk in a green visor who did not work for Cogentic and made that clear by not looking up.

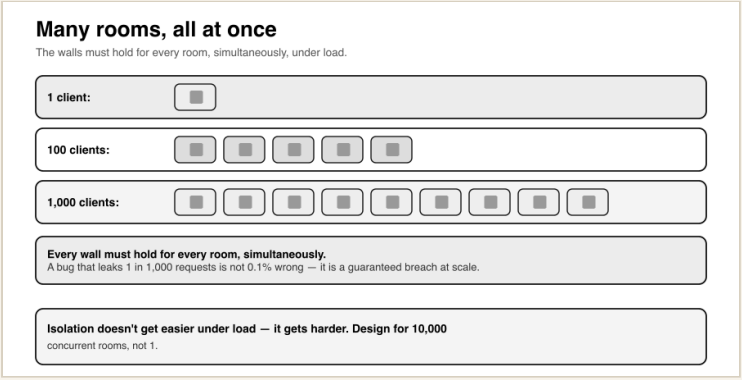
"Records Office," Reach murmured to Newt. "Well. A records office. The Granary's. Their database. Everything they know about their own bookings lives back there, and they let us *look* through this window, because looking is safe. Watch."

She leaned to the window. "Current price to hold our date. The 14th. 30 people."

The clerk, without enthusiasm, slid a card across the ledge. It said: **DEPOSIT TO HOLD: $450. BALANCE DUE: $1,550 ON ARRIVAL.**

THE REAL THING: Concurrency A real system does not serve one client at a time, politely, in a queue. It serves **thousands at once**: thousands of sessions live simultaneously, thousands of workers active in the same instant, all of it overlapping. This is **concurrency**, and it is where isolation gets tested for real, because a wall that holds when 2 people are talking is not the same as a wall that holds when 10,000 are, all at the same moment, all under load. The walls cannot get sloppy when the floor gets busy. Busy is exactly when the bleeds happen.

THE WHITEBOARD: many rooms, all at once



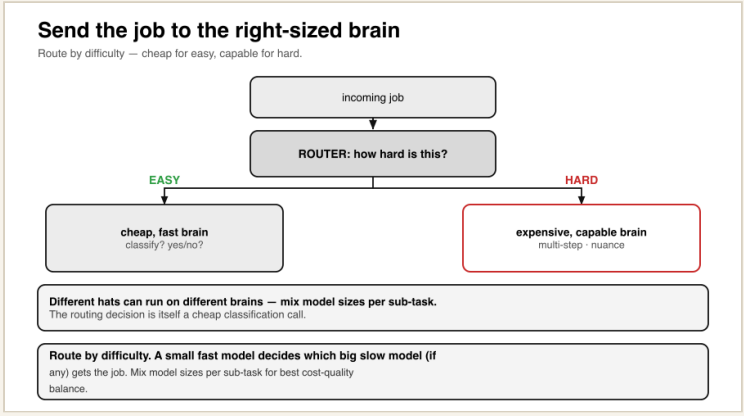
many rooms, all at once

A real system isn't one client at a time in a queue, it's thousands talking at once, each in their own room. Every wall has to hold for all of them simultaneously, and the busiest moments are exactly when a wall is most likely to slip.

Newt walked the corridor while Maestro and Warden talked load and limits, and that is how she found the ghost.

* * *

THE WHITEBOARD: send the job to the right-sized brain



send the job to the right-sized brain

A router looks at each job and asks how hard it actually is. Easy jobs go to a small, cheap, fast brain; only the genuinely hard ones go to the big expensive one. Since most questions are easy, this saves a fortune without anyone noticing.

* * *

It was beautiful, and it was working, and the meter on Tally's desk visibly slowed, and Tally made a small sound of physical relief, like a woman loosening a belt.

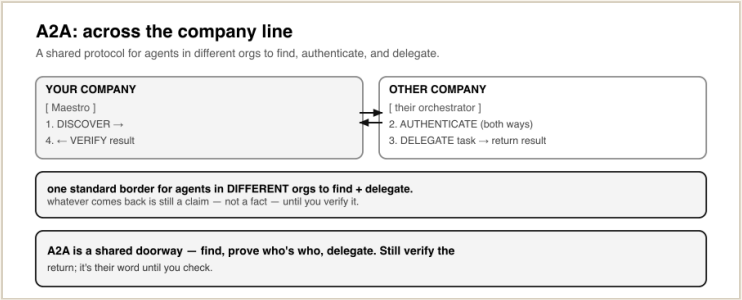
And then, because the floor never gave a lesson without its sting, Thrift got one wrong.

A question came down the line, and it *looked* easy, short, politely worded, the kind of thing a rowboat handles in its sleep, and Thrift, going fast, flicked it left, to the small cheap brain.

It was not easy. It only looked easy. Under the polite wording was a genuinely hard knot: a client asking whether, given a refund, a restocking fee, a partial return, and a promotion that had since expired, they were owed money or owed it back.

THE REAL THING: A2A, talking to other companies' agents Sometimes a mission needs a capability you don't have, that lives inside another organization's agent. **A2A** (agent-to-agent) is the emerging standard for exactly this: a shared protocol that lets agents across different companies **discover** each other (find one that can do the task), **authenticate** (each side proves who it really is), and **delegate** (hand over a task and receive a result), without every pair of companies hand-building a private connection. It is to *agents* what MCP is to *tools:* one standard way to reach across the boundary instead of 1,000 custom ones. It is what turns a wall of separate companies into a network that can actually get things done together.

THE WHITEBOARD: A2A, across the company line



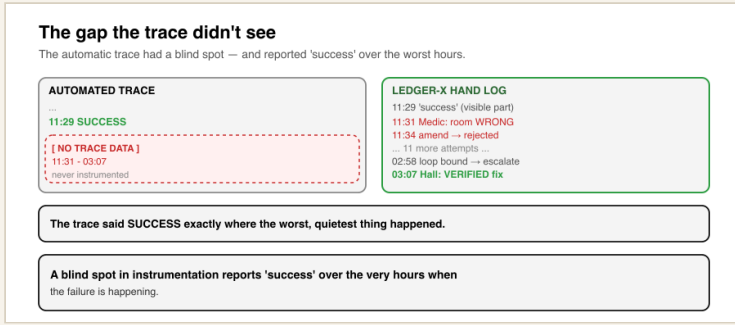
A2A, across the company line

A2A is a shared doorway for one company's agent to find another company's agent, prove who they each are, and hand over a task. Whatever comes back still has to be checked, because it's only their word until you verify it.

"Why does that matter so much?" Newt asked. "A tool, an agent, what's the difference? They both do a thing you ask."

"No." Maestro stopped at the border desk. "A tool *obeys.* You swing a hammer, the hammer does not have opinions about the

THE WHITEBOARD: the gap the trace didn't see



the gap the trace didn't see

The automatic tracing had a blind spot, the self-healing subsystem was new and nobody had wired it in, so the trace cheerfully reported "success" over the very hours when the room was wrong and the loop was spinning. Only the stubborn hand-kept log filled the hole.

"This is the lesson nobody wants," said ETHOS. "Your shiniest, most automated record had a hole in it, and the hole was in *exactly* the place a hole is most dangerous: over the silent, self-patching part, the part no customer saw and no alarm caught. If Ledger-X had not been doing the boring, manual, unthanked thing this whole time, the truth of that night would simply not exist. We would have the cheerful trace, the word *success,* and a wrong room, and no way on earth to connect them."

* * *

Then Iris-9 turned on her wall of dials, and the room got quiet in a different way.

"You've been looking at *one night,*" she said. "Look at the last *8 weeks.*"