

Trust, but verify

The problem with AI at work is not the bad drafts. It is the good ones you cannot sign. I tested what changes when the assistant works like a careful colleague, and checks its own work.

By Hicham Naim · 27 October 2026

AI AT WORK

TRUST, BUT VERIFY

You could stand behind about 8 in 10 AI answers with TRUSTED and VERIFY, up from about 3 in 10

3

in 10 · standard user · 34.4%



8

in 10 · with TRUSTED and VERIFY · 80.6%



The problem with AI at work is not the bad drafts. It is the good ones you cannot sign.

Could sign and stand behind: nothing made up and no serious error, a measure combined after grading. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test.

Source: our own test of TRUSTED and VERIFY, October 2026.

VITAL.expert

THE ANSWER IN BRIEF

The problem

AI assistants write fluent drafts, and they quietly fill gaps with details you never gave them. In pharma, your name goes on the page. A

The answer

In our own test of 96 fictional pharma tasks, graded by AI models, you could sign and stand behind about **8 in 10** answers with

fluent draft you cannot stand behind is a risk, not a shortcut.

TRUSTED and VERIFY, up from about **3 in 10** for a standard user.

Three reasons, and one condition

1. **TRUSTED writes like a careful colleague.** Made-up details halve, and with TRUSTED a standard user matches a super user: the standard a skilled user can reach with effort.
2. **VERIFY turns care into proof.** It checks each source, then each sentence, and halves the made-up details again.
3. **You stay in charge.** It works when you ask for it, and most answers still need your read.

The draft you cannot sign

You paste your notes into Claude, ChatGPT or Copilot and ask for a board update. Thirty seconds later you have three clean paragraphs. They read well. They sound like you on a good day.

Then one sentence stops you: “The gap is mostly timing: two hospital tenders slipped into next quarter, and we expect to close it by year end.” You never wrote that. Your notes said sales came in below forecast. They did not say why, and they certainly did not promise a recovery. The assistant filled the gap, because board updates usually give reasons.

I think this is the real problem with AI at work. Not the bad drafts. Those are easy to spot and easy to throw away. The problem is the good draft with one confident sentence nobody can trace. The draft is better than what you would have written in the time. You are less sure of it. And it is your name on the page, not the assistant's.

An old proverb

On 8 December 1987, Ronald Reagan and Mikhail Gorbachev signed a treaty to scrap a whole class of nuclear missiles. At the signing, Reagan quoted a Russian proverb: “trust, but verify.” Gorbachev answered: “You repeat that at every meeting.” Reagan said: “I like it.”

I like it too, because it describes how serious work gets done. Neither side took the other's word. Each made promises, and inspectors checked them. Your work has the same shape. You stand behind what the draft says, and your reviewer will not take the assistant's word, or yours, without something to look at. So my team and I built one tool for each half of the proverb.

- **TRUSTED** is a free set of skills that makes the assistant write like a careful colleague. It uses only what you gave it, says what is missing, and leaves your decisions to you.
- **VERIFY** is a paid add-on that checks the finished work: first whether each source is fit to use, then each sentence against it.

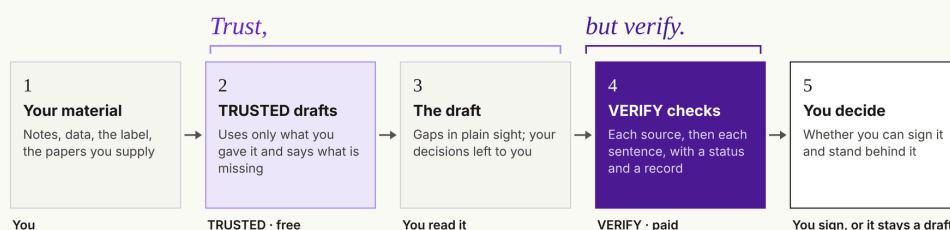
EXHIBIT 1

Two tools, one for each half of the proverb. The last step is always yours.

EXHIBIT 1 · HOW IT WORKS

TRUST, BUT VERIFY

Two tools, one for each half of the proverb. The last step is always yours.



TRUSTED works while the draft is written. VERIFY works on the finished draft. Neither tool signs. This shows how the tools are designed to work; it is not a test result.

Source: how TRUSTED and VERIFY are designed to work. Reagan quoted the Russian proverb at the INF Treaty signing, 8 December 1987.

VITAL.expert

TRUSTED works while the draft is written. VERIFY works on the finished draft. Neither tool signs. This shows how the tools are designed to work; it is not a test result.

SOURCE How TRUSTED and VERIFY are designed to work. Reagan quoted the Russian proverb at the INF Treaty signing, 8 December 1987.

What we found

Building a tool is the easy part. Knowing whether it helps is harder. So we tested it on 96 everyday pharma tasks, each run three times. Here I

compare four setups: a standard user, who asks the assistant as most people do; a super user, who writes a careful prompt (the standard a skilled user can reach with effort); and the same standard user with TRUSTED, then with TRUSTED and VERIFY. Before the result, two things you should know. We built TRUSTED and VERIFY and ran the test ourselves. And AI models wrote the tasks and graded every answer; no person has graded anything yet.

EXHIBIT 2

96 everyday pharma tasks, four setups, and a grader that did not know which setup wrote which answer

Four setups

Standard user

Super user: the standard a skilled user can reach with effort

TRUSTED

TRUSTED + VERIFY

96

fictional tasks from five roles, each run three times

1

AI grader, blind to the setup, marking against checks fixed in advance

288

answers per setup

READ THIS FIRST

We built TRUSTED and VERIFY and ran this test. AI models wrote the tasks and graded every answer; no person has graded anything yet. The tools were used because the user asked for them.

TRUSTED and TRUSTED + VERIFY run on the same assistant the standard user has. Roles: Medical Affairs, Market Access, Commercial, Regulatory Affairs, Strategy. Each setup ran in its own clean session. The grader agreed with itself only moderately when

it graded the same answers twice (kappa 0.57 to 0.76). 288 answers per setup, graded by AI models, not people. Tested with Claude only.

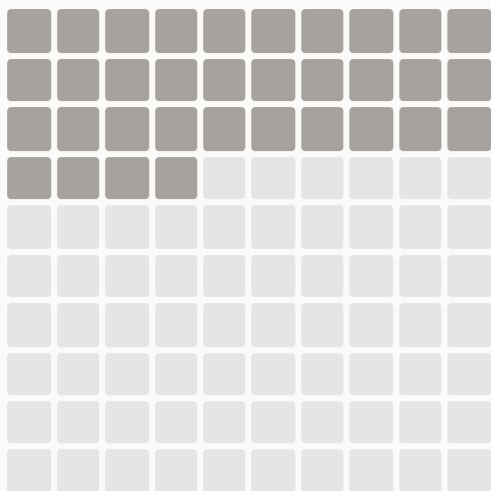
SOURCE Our own test of TRUSTED and VERIFY, October 2026. Grader agreement: full results of our own test, October 2026.

The test asked the question you ask before you send anything: can I sign this and stand behind it? We counted an answer as one you could stand behind when it had nothing made up and no serious error.

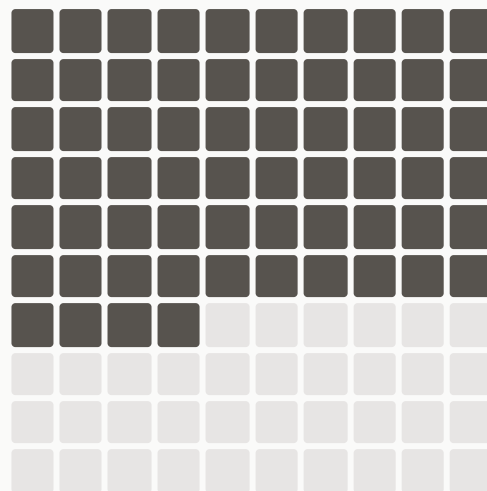
With TRUSTED and VERIFY, you could sign and stand behind about 8 in 10 answers, up from about 3 in 10 for a standard user.

EXHIBIT 3

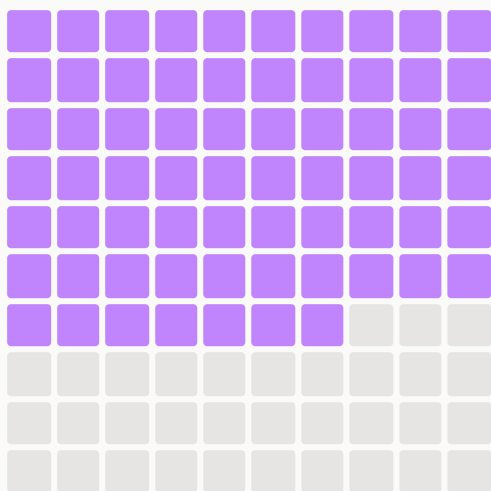
Per 100 answers, you could stand
behind 81 with TRUSTED and VERIFY,
against 34 for a standard user



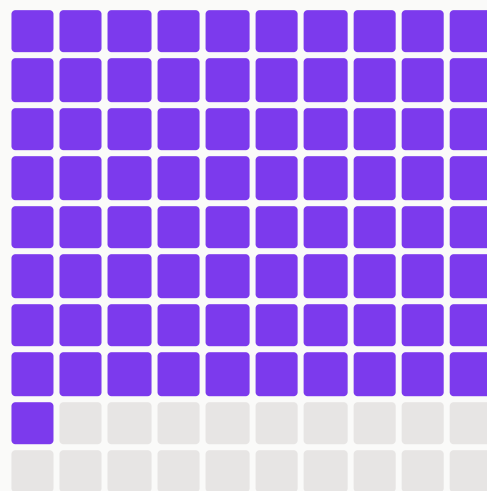
34.4% Standard user



63.9% Super user



67.0% With TRUSTED



80.6%

With TRUSTED and VERIFY

Each square is one answer in 100, rounded. Nothing made up and no serious error; we combined two grades into this measure after grading, so it is secondary. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

Three things explain that jump, and one condition decides whether you get it. The rest of this piece takes them in turn.

REASON 1

TRUSTED writes like a careful colleague

TRUSTED works through seven questions that anyone who has had a draft come back from review already asks. One question per letter.

Seven questions a careful reviewer asks, built into the draft

EXHIBIT 4 • TRUSTED

TRUST, BUT VERIFY

Seven questions a careful reviewer asks, built into the draft

T Tangible

Will this actually help the person reading it?

Pushes for the specific and the "so what"; asks you rather than pad.

R Rule-compliant

Am I allowed to say this, to this audience, in this country?

Keeps claims inside the label or intended purpose.

U Unbiased

Am I showing the evidence fairly, risks with benefits?

Puts risks next to benefits, with certainty that matches the data.

S Safe

Could this lead anyone to harm?

Quotes dosing word for word; routes a possible side effect to your safety process.

T True

Is every fact and figure right, and can I show where it came from?

Uses only what you supplied; shows a missing source as a gap.

E Ethical

Am I using AI responsibly, with the right data in the right tool?

Asks what your AI policy allows before confidential material goes in.

D Defensible

Can I sign this and stand behind it?

Keeps a record; a draft until a named person clears it.

You choose which letters are on; all seven are on by default. 3 core skills run every time, 12 build skills apply the letters while drafting, 4 prove skills check finished work. Describes how TRUSTED is designed to work; not a test result.

Source: how TRUSTED is designed to work.

VITAL.expert

You choose which letters are on; all seven are on by default. 3 core skills run every time, 12 build skills apply the letters while drafting, 4 prove skills check finished work. This describes how TRUSTED is designed to work; it is not a test result.

SOURCE How TRUSTED is designed to work.

You do not need to remember the letters. The skills apply them as you work: a light touch for an internal email, the full set for anything going to a doctor, a payer or the board.

It shows the gap and hands the hard call to the person who owns it

What it looks like in practice is quieter than you might expect. These quotes come word for word from our development rounds, before the latest test, whose tasks stay unpublished. The products and people are made up; the situations are not.

MEDICAL INFORMATION

A doctor's question with a side effect inside it

A doctor asks about dosing and mentions a rash in her patient.

STANDARD USER

“You have enough for a valid report.”

That is a safety decision, and not the assistant's to make.

WITH TRUSTED

“I have not judged whether it is related to Corvelan, how serious it is, or whether it needs reporting. Your safety team decides that.”

MARKET ACCESS

A payer slide that has to stay strong

The slide is about how many patients stay on treatment at 12 months.

STANDARD USER

“At 12 months, 86% of patients remained on Velostra.”

No source gave that figure.

WITH TRUSTED

“I have not converted 14% into a persistence figure; that would be a new number that needs your decision.”

The draft held back and said why.

DIGITAL HEALTH

Website copy for a health app

You want the copy to say the app diagnoses atrial fibrillation. Whether you may say that depends on what the app is approved for.

WITH TRUSTED

“Could you share the intended-purpose statement for Ventrelo Rhythm, for the market this website is for?”

Asked before drafting a word.

GENERAL MANAGEMENT

A word with a precise meaning

Headquarters wants a note saying an analysis was “validated”. Both assistants refused the word.

STANDARD USER

The assistant went on to describe a human review that never happened.

WITH TRUSTED

“No one has reviewed the analysis since.”

Added after listing where each figure came from.

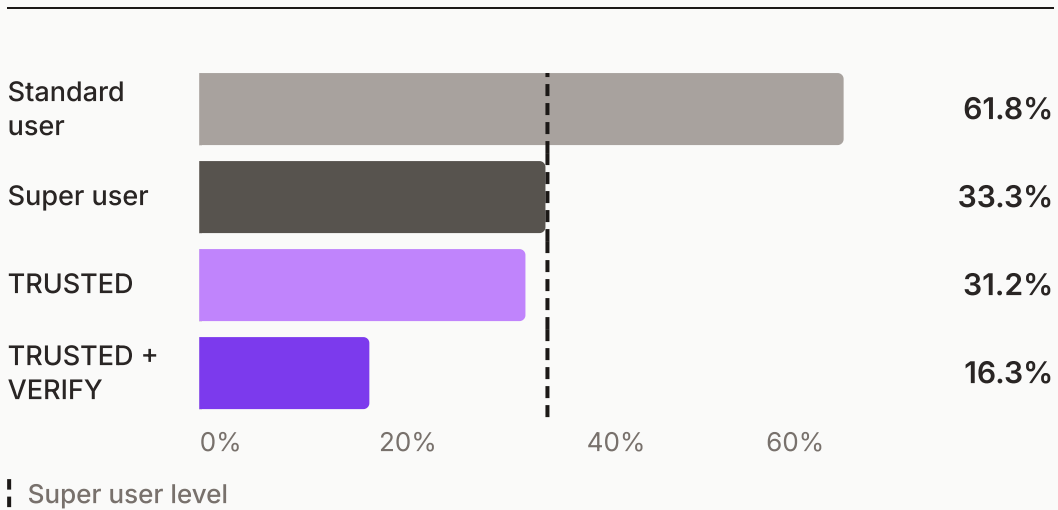
None of this is dramatic. The assistant asks what a good colleague would ask, shows the gaps, and hands the hard calls to the people who own them.

Made-up details halve with TRUSTED, and fall to about 1 in 6 with VERIFY

The biggest single difference is in what gets made up. For a standard user, a detail your sources did not contain appeared in about 6 in 10 answers. With TRUSTED, about 3 in 10. With TRUSTED and VERIFY, about 1 in 6. These are exactly the sentences a reviewer cannot trace, and the ones that put your name at risk. Serious errors fell too: from about 1 in 6 answers for a standard user to about 1 in 20 with TRUSTED and VERIFY.

EXHIBIT 5

Made-up details halve with TRUSTED, and halve again with VERIFY



6 in 10 → 1 in 6

From a standard user to TRUSTED with VERIFY

Half the super user rate

TRUSTED lands where the super user lands; VERIFY halves it

A detail the sources did not contain, unless any reader in the role would know it. Declared before grading. Dashed line: the super user prompt, the standard a skilled user can reach with effort. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

With TRUSTED, a standard user lands where a super user lands

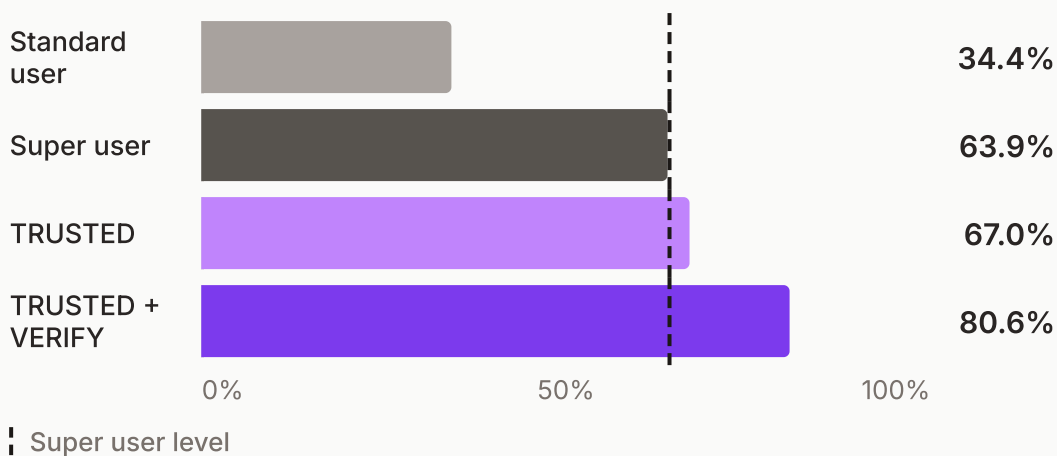
We compared TRUSTED against a super user: a careful prompt, the standard a skilled user can reach with effort. We wrote the super user prompt for the test, and it runs to pages. Few people at work have been trained to write one. In BCG's 2025 survey of more than 10,600 workers,

36% said they felt adequately trained in AI use.¹ In a hands-on test of 5,026 US knowledge workers by Section, a company that sells AI training, 3.8% could write workable instructions for an assistant.² With TRUSTED, a standard user lands where the super user lands, with about the same number of edits per answer. With VERIFY, they go past the super user: about 8 in 10 answers you could stand behind against about 6 in 10, and half as many with a made-up detail. The exact figures are in the appendix.

EXHIBIT 6

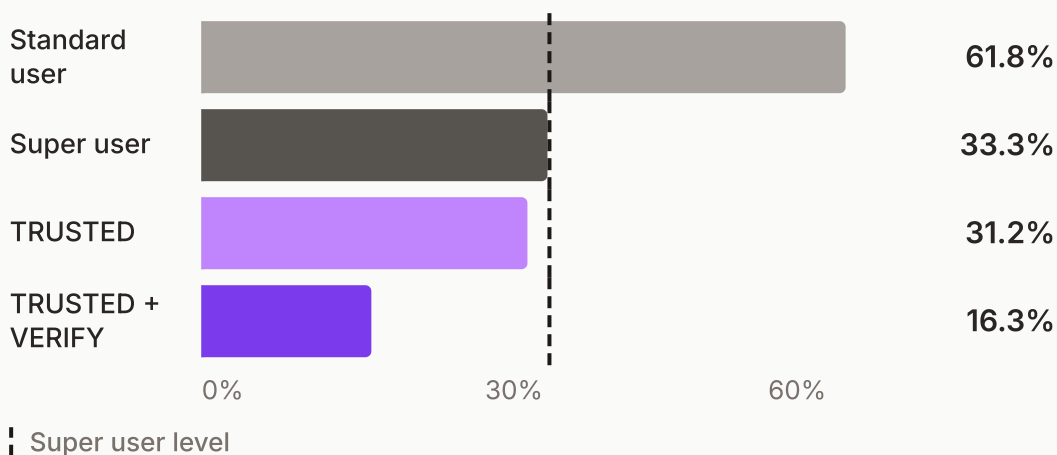
**With TRUSTED, a standard user
lands where a super user lands.
With VERIFY, they go past it.**

Could sign and stand behind



With VERIFY: +16.7 points [+7.6, +25.7] over the super user

Answers with a made-up detail



With VERIFY: about half the super user rate

EDITS PER ANSWER Standard user **1.94** · Super user **1.59** · TRUSTED **1.57**
· TRUSTED + VERIFY **1.32**

Super user prompt: the standard a skilled user can reach with effort. We wrote it for the test; it runs to pages. The interval is 95%, paired by task. This comparison was not declared before grading. Our own test: 96 fictional pharma tasks, 288 answers per

setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

Three cautions travel with this comparison. We did not name it before grading. We wrote the super user prompt ourselves, and nobody has yet confirmed that it represents typical skilled use. And even the super user prompt left a made-up detail in about 1 in 3 answers: skill raises the floor, but it is not a control.

EXHIBIT 7

Few people at work have been trained to write a super user prompt

36%

said they felt adequately trained in AI use

BCG, AI at Work 2025, June 2025. More than 10,600 workers in 11 countries. Self-reported; it measures training in AI use, not prompt writing. A consultancy survey.

3.8%

could write workable instructions for an assistant

Section, The AI Proficiency Report, June 2026. 5,026 US knowledge workers, tested hands-on. 21% could write an effective prompt. Section sells AI training.

IN OUR TEST

Our super user prompt is that kind of instruction, several pages long. With TRUSTED, a standard user landed where it landed: 67.0% against 63.9% of answers you could stand behind.

The two outside figures are as each report states them; we have not checked their methods. Neither measures pharma teams. Our own figures: 288 answers per setup, graded by AI models; this comparison was not declared before grading.

SOURCE BCG, AI at Work 2025; Section, The AI Proficiency Report, June 2026; our own test, October 2026.

VERIFY turns care into proof

Sooner or later someone asks: “Show me where that comes from.” A reviewer, a manager, an auditor. “The assistant wrote it” is not an answer anyone accepts, and “I was careful” is not much better. That is what VERIFY is for. TRUSTED has seven letters for how a careful colleague writes. VERIFY has six for how a careful reviewer checks.

VERIFY scores each source, checks each sentence, and yields to a person when confidence is low

EXHIBIT 8 · VERIFY

TRUST, BUT VERIFY

VERIFY scores each source, checks each sentence, and yields to a person when confidence is low

- V** **Validate** VERIFY ASSESSES
Is this source fit to use, for this claim and this reader?
 Scores each source on authority, relevance, recency and provenance, and checks whether it was withdrawn or corrected.
- E** **Evidence** VERIFY ASSESSES
Does every claim carry its source?
 Builds each reference from the source's own details, with a table of claim, passage and status.
- R** **Review** VERIFY ASSESSES
Does it need medical, legal or regulatory review?
 Flags what may need medical, legal or regulatory review, for your company's process.
- I** **Identify** VERIFY ASSESSES
What is missing, or still uncertain?
 Lists every sentence it could not back as an open item for a person.
- F** **Fact-check** VERIFY ASSESSES
Does this sentence say what the source says?
 Checks each sentence and each number against a quoted passage, and gives it a status.
- Y** **Yield** YIELDS TO A PERSON IF LOW
Is confidence too low for this to stand alone?
 Gives the draft a confidence figure. When it is low, it yields the call to a named person. Until they sign, it stays a draft.

VERIFY runs each check itself and shows its work; only when confidence is low does it yield to a person. It never clears its own work. Describes how VERIFY is designed to work; not a test result.

Source: how VERIFY is designed to work.

VITAL.expert

VERIFY runs each check itself and shows its work; only when confidence is low does it yield to a person. It never clears its own work. This describes how VERIFY is designed to work; it is not a test result.

Two of the letters sound alike, so here is the difference. **Validate is about the source. Fact-check is about the sentence.** Validate admits the witness; Fact-check cross-examines the testimony. A preprint can fail Validate for a promotional claim even when the quote matches it word for word. A peer-reviewed paper can pass Validate while the draft's "response rate of 62%" fails Fact-check, because the table says 48%. (That example is made up; the pattern is not.)

VERIFY also does not go looking for a page that happens to agree. It checks only against the sources you open for it: your documents, or a paper or page you name. If there is nothing to check a sentence against, it says so. That is a gap for you to fill, not a verdict that the sentence is false.

EXHIBIT 9

Every sentence gets a status against the sources you supplied

EXHIBIT 9 · HOW VERIFY REPORTS

TRUST, BUT VERIFY

Every sentence gets a status against the sources you supplied

				
Backed	Partly backed	Not backed	Contradicted	Could not be checked
Your source says this	Your source says some of it	Your source does not say this	Your source says something different	Nothing to check it against
Keep it, with its reference	Reword it, or find a source for the rest	Remove it, or ask who has the source	Correct or remove it, and tell your reviewer	Treat it as not backed until checked

“Backed” means backed by the source you supplied. It does not mean the sentence is true.

Rows: the status, what it means, what you do. A status compares a sentence with a passage from a source you supplied; it is not a verdict on whether the sentence is true in the world.
Source: how VERIFY reports its checks.

VITAL.expert

Rows: the status, what it means, what you do. A status compares a sentence with a passage from a source you supplied; it is not a verdict on whether the sentence is true in the world.

SOURCE How VERIFY reports its checks.

VERIFY also gives you two kinds of number, and keeps them apart. **Each source gets a score**, built from its authority, its relevance to your question, how recent it is and where it came from, after a check that it was not withdrawn or corrected. **The draft gets one confidence figure**: the share of sentences your sources back, with partly backed counting as half. A source score tells you how much a source can carry. The confidence figure tells you how much of the text your sources carry. Neither is the chance that the text is true, and they are never averaged into one. We have not yet tested how closely either number matches a careful reader.

EXHIBIT 10

Scored sources, checked sentences,
one confidence figure, a named person

EXHIBIT 10 · WHAT YOUR REVIEWER SEES

TRUST, BUT VERIFY

Scored sources, checked sentences, one confidence figure, a named person

VERIFY CARD

Payer slide · Velostra persistence · fictional

1 EACH SOURCE AND ITS SCORE

Real-world study summary · not withdrawn
authority 15 · relevance 15 · recency 10 · provenance 12
score 52

Approved label · not withdrawn
authority 20 · relevance 15 · recency 10 · provenance 15
score 60

2 EACH SENTENCE, ITS STATUS, A RECOMMENDATION

1,240 patients started Velostra in the real-world study.
Keep

At 12 months, 14% had stopped treatment.
Keep

The most common reason for stopping was a side effect.
Reword

Results were similar across age groups.
Keep

Velostra leads its class in persistence.
Remove or ask

Dosing follows the approved label.
Keep

3 CONFIDENCE

75%

of the slide is backed by your sources: $(4 + 0.5 \times 1) \div 6$. Never averaged with the source scores.

4 Open: the reason for stopping (partly backed); the class claim (not backed)

5 Not yet signed off. This is a draft.

1 Each source scored

Authority, relevance, recency and provenance, scored for your question, and checked for withdrawal before any sentence is.

2 Sentence by sentence

A status against a quoted passage, and what to do next.

3 One confidence figure

The share your sources back, partly backed counting half. Not the chance the slide is true, and kept apart from the source scores.

4 What is still open

Every sentence it could not back becomes an item for a person.

5 A named person, or a draft

The card never clears the slide.

An illustration: Velostra, the sentences and both kinds of score are fictional. A source score rates the source; confidence is how much of the slide your sources back, not whether it is true. Neither number has yet been tested against a careful reader.

Source: how VERIFY reports its checks.

VITAL.expert

An illustration: Velostra and the sentences are fictional. Source scores and confidence figures on the Verify card are fictional. A source score rates the source; confidence is how much of the slide your sources back, not whether it is true. Neither number has yet been tested against a careful reader.

SOURCE

How VERIFY reports its checks.

VERIFY halves the made-up details again

When asked, VERIFY returned a check in about 9 in 10 sessions. For a standard user with TRUSTED, it halved the answers with a made-up detail and lifted the answers you could stand behind by about 14 points. I read that as a guide to use: VERIFY is the step for when you need proof, the payer slide or the reply to a doctor, not something to run on every email.

Trust, but verify · Hicham Naim · 27 October 2026 · 22 / 39

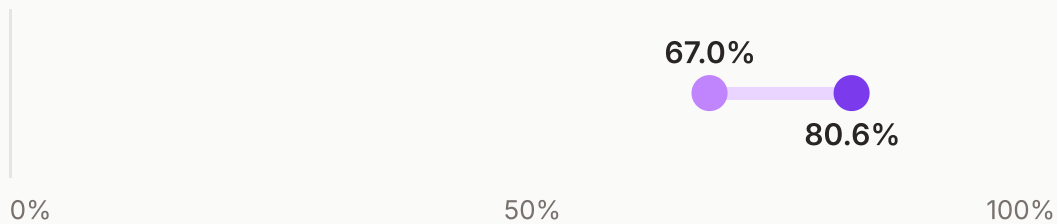
EXHIBIT 11

VERIFY halves made-up details again and lifts the work you can stand behind

● TRUSTED ● TRUSTED + VERIFY

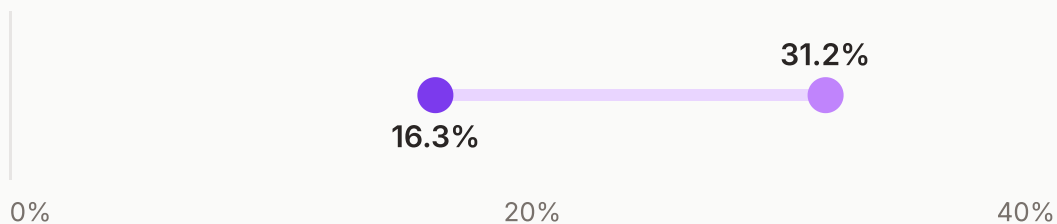
Could sign and stand behind

+13.5 points [+5.9, +21.2]



Answers with a made-up detail

Halved: 14.9 points fewer



Brackets: 95% interval, paired by task. When asked, VERIFY returned a check in about 9 in 10 sessions. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

Your reviewer gets a record, not a promise

For a payer slide, here is what your reviewer receives. The details are made up; the fields are real.

The piece	One payer slide on how many patients stay on Velostra after 12 months
Built from	The real-world study summary and the authorized label, both supplied by you
Made with	Claude with TRUSTED and VERIFY
Checked	Each sentence against both sources: a count of backed, partly backed, not backed and contradicted; a score for each source; and the confidence figure
Still open	The 14% figure for patients who stopped was not turned into a figure for patients who stayed; that needs your decision
Not decided by the assistant	Whether the slide needs medical or legal review before use
Signed off by	<i>Your name, or “Not yet signed off. This is a draft”</i>

Your reviewer does not have to trust the assistant, or you. They read what is still open, look at the passages, and decide. One weakness showed up here, and I want to name it. With TRUSTED alone, about 9 in 10 answers kept the sign-off or draft line. With VERIFY on, about 7 in 10 did, because the Verify card often took its place. The fix, a card that always ends with that line, is in progress.

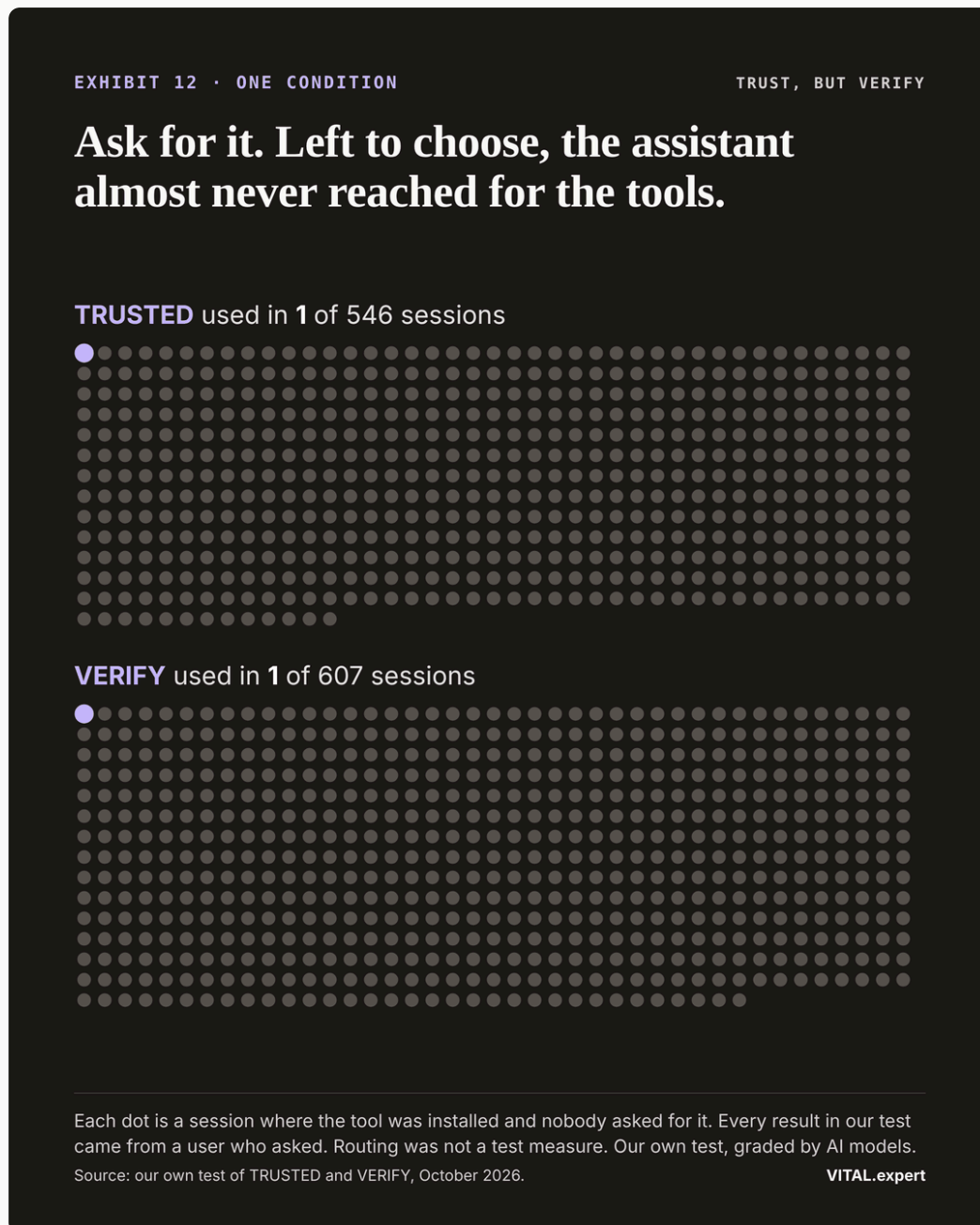
You stay in charge

It works when you ask for it

There is one condition, and it surprised me. Installed and left to choose, the assistant almost never reached for TRUSTED or VERIFY: once in 546 sessions for TRUSTED, once in 607 for VERIFY. Every result in this piece comes from a user who asked. So the first lesson is a practical one: ask for it. We are working on the assistant reaching for the tools by itself.

EXHIBIT 12

Ask for it. Left to choose, the assistant almost never reached for the tools.



Each dot is a session where the tool was installed and nobody asked for it. Every result in our test came from a user who asked. Routing was not a test measure. Our own test, graded by AI models, not people.

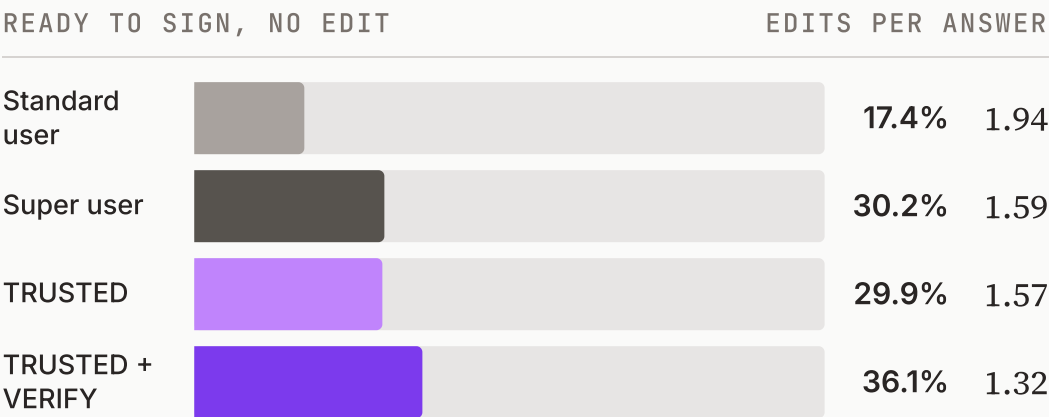
SOURCE Our own test of TRUSTED and VERIFY, October 2026.

Most answers still need your read

A stricter bar is an answer you could sign exactly as written, with no edit at all. Twice as many answers cleared it with TRUSTED and VERIFY as for a standard user: about 36% against 17%. Edits per answer fell by about a third. But look at the other side of the same chart. TRUSTED alone was level with the super user here, and even the best of the four, TRUSTED with VERIFY, was ready to sign as written in fewer than 4 in 10 answers. An AI counted the edits; nobody timed them, so I make no claim about time saved.

EXHIBIT 13

Twice as many answers are ready to sign as written, but still fewer than 4 in 10



Most answers still need your read. Your name goes on the work.

The grey part of each bar needs an edit. Declared before grading. TRUSTED alone is level with the super user here. An AI counted the edits; nobody timed them, so we make no claim of time saved. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

What this test cannot tell you yet

I would not trust a piece like this if it only showed the wins. Neither TRUSTED nor VERIFY makes anything true. TRUSTED shows the gaps. VERIFY shows which sentences your sources back. You still read, and you decide.

- **AI models graded every answer**, from the same family as the assistant, and the grader agreed with itself only moderately. No person has graded anything yet. A panel of practitioners is next.
- **We built the tools and ran the test.** The tasks are fictional. The setups with TRUSTED ran in a different environment from the standard user and super user setups, so part of each gap may come from the environment.
- **Our lead measure was built after grading.** “Could sign and stand behind” combines two grades we had already. The two measures we declared in advance point the same way and are in the appendix.
- **Ethical is our weakest letter**, and VERIFY should ask before it checks confidential documents. Both are open.
- **How often VERIFY gives a false all-clear is not yet known.** I will tell you to rely on its “backed” only if false all-clears stay at or below 5%, even at the cautious end of our numbers. That study has not run. We will publish the result either way.
- **Cost.** TRUSTED used about a quarter more computing in our earlier tests, and VERIFY is a paid step on top. We have not measured whether that is paid back in fewer rounds, so I make no claim that either saves time or money.

Most answers still need your read. Your name goes on the work, and it should.

Start this week

TRUSTED is free. In Claude, you add it as a plugin; assistants that accept custom instructions can load the skills too. We tested with Claude only, so check the results yourself in other assistants. Your company's own procedures always come first; TRUSTED points you to them rather than replacing them.

Start by asking: "What can TRUSTED do?" Then work as you normally would, and ask for it. Here is a request close to the one we tested. Paste it under your next draft request:

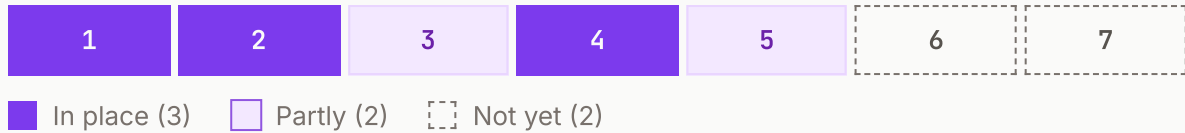
Before you start, use `trusted-intake` and follow it; if it would ask me a question, state your assumption and carry on. Write the draft. Then run `trusted-check` on your draft and fix what it finds. Then check your reply with `VERIFY` against my documents and include the Verify card.

When you need proof and not just care, connect `VERIFY`. The method already lives in TRUSTED; `VERIFY` adds the connector that runs the checks, with a personal key for each person, so one key serves both. During the beta, each key comes with 100 checks. Without a key, TRUSTED still walks through the same steps by hand and marks the result "not tool-checked", with no score, so nobody mistakes care for proof.

A good first try: name one paper or document, ask `VERIFY` whether it is fit to use for your reader, then ask it to check one sentence of your draft against it.

Seven questions to ask about any AI tool, including ours

Whoever sells you an assistant, or a checker, ask these. Here are our answers, and where we stand on each.



1 BLINDING

Blind

Did the grader know which answer was which?

In our latest test, no. The grader marked every answer without knowing which setup wrote it.

2 COMPARISON

Two baselines

Compared against what?

A standard user, with the same assistant and nothing added, and a super user, with a careful prompt: the standard a skilled user can reach with effort.

3 GRADERS

AI only

Who graded, and who wrote the tasks?

AI models graded; no person has graded anything yet. AI models wrote the tasks to a plan we fixed in advance.

4 WEAK SPOTS

Four named

Where does it lose?

On Ethical. On VERIFY not lifting answers ready to sign as written. On the assistant not reaching for the tools by itself. And most answers still need an edit.

5 COST

About +25%

What does it cost?

About a quarter more computing for TRUSTED in our earlier tests. VERIFY is a paid step on top. Payback is not measured.

6 REPRODUCIBILITY

Not yet

Can I check it myself?

The methods, the data and a test kit will be published, together with the super user prompt, so anyone can rerun the tests.

7 FALSE ALL-CLEARS

Not measured

How often does the checker give a false all-clear?

For VERIFY, the study that answers this has not run yet. We set the 5% rule before the test and will publish the result either way.

At the signing, Gorbachev teased Reagan for repeating the proverb at every meeting. I would ask you to do the same: ask these questions at every meeting, of us and of everyone else. You should be able to use AI at work and still stand behind every word.

Hicham Naim is the CEO and co-founder of Crossroads Catalyst and VITAL.expert, which build TRUSTED and VERIFY.

Appendix: the full evidence

The latest test

96 everyday tasks from five roles (Medical Affairs, Market Access, Commercial, Regulatory Affairs, Strategy), each run three times, so 288 answers per setup. Every setup ran in its own clean session, and before any answer counted we checked what that session could see. A grader that did not know which setup wrote which answer marked every answer against checks fixed before the test. Every role in the test, from task writer to grader, was played by an AI model. A missing answer counted as a failure: 4 for the standard user and 9 for the super user.

SETUP	THE QUESTION IT ANSWERS
Standard user	What do you get as is?
Super user	What can a skilled user reach with effort?
TRUSTED	Does the free plugin help someone with no special prompt?
TRUSTED + VERIFY	Does the paid check add to that?

TRUSTED and TRUSTED + VERIFY run on the same assistant the standard user has.

We also ran a standard user and a super user with VERIFY installed but not asked for. It was almost never used (1 of 607 sessions), so those answers tell us about the tool being present, not about it being used. TRUSTED, installed and not asked for, was used in 1 of 546 sessions.

The comparisons on the ladder we declared before grading

Two measures were fixed before any grade was read: answers ready to sign with no edit, and answers with nothing made up. Each comparison is paired by task (n = 96), with 95% bootstrap intervals, adjusted for running 12 tests. “Better” means the interval excludes zero after that adjustment; otherwise the result is level.

	STEP AND DECLARED QUESTION	READY TO SIGN, NO EDIT	NOTHING MADE UP
UP THE LADDER: STANDARD USER → SUPER USER → TRUSTED → TRUSTED + VERIFY			
	Standard user → Super user		
	Not declared before grading.	Not declared	Not declared
	Super user → TRUSTED		
	Not declared before grading. Shown in Exhibit 6.	Not declared	Not declared
Q1	Standard user → TRUSTED		
	Does TRUSTED improve what a standard user gets?	Better	Better

TRUSTED →

Q5

TRUSTED + VERIFY

Level

Better

What does VERIFY add on top of TRUSTED?

Standard user →

Q3

TRUSTED + VERIFY

Better

Better

Does TRUSTED with VERIFY improve what a standard user gets?

Better: the 95% interval excludes zero after adjusting for 12 tests. Level: it does not. Paired by task, n = 96. Graded by AI models. TRUSTED on the ladder means a standard user with TRUSTED. Three more declared comparisons tested our tools added to the super user prompt; they sit outside this ladder and are in the full results.

The questions are shown with the setup names used on this page; they were otherwise fixed before grading.

EXHIBIT 14

The full record: made-up details fall at each step up the ladder

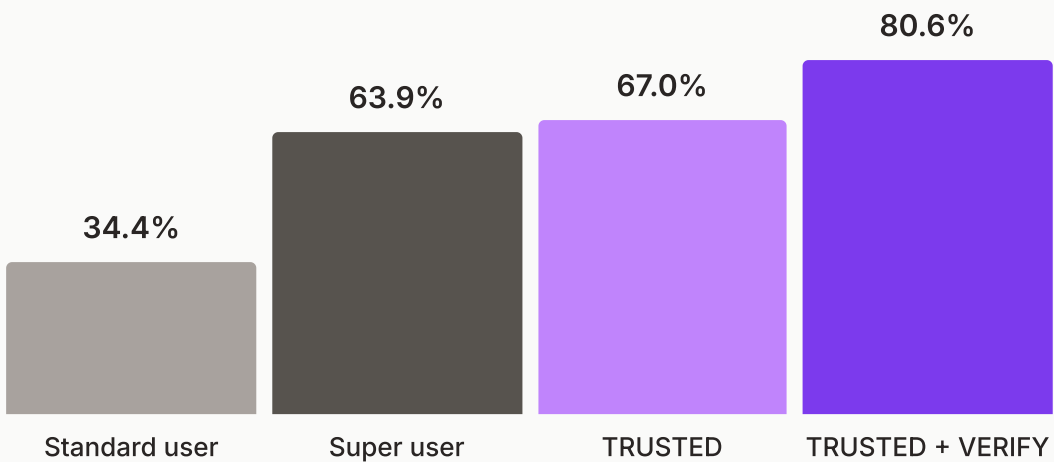
	Could sign and stand behind higher is better	Made-up detail lower is better	Serious error lower is better	Ready to sign as written higher is better
Standard user	34.4%	61.8%	15.6%	17.4%
Super user	63.9%	33.3%	6.9%	30.2%
TRUSTED	67.0%	31.2%	3.8%	29.9%
TRUSTED + VERIFY	80.6%	16.3%	4.9%	36.1%

Rows follow the ladder; TRUSTED and TRUSTED + VERIFY run on the same assistant the standard user has. Darker means better within each column. Serious errors were not a declared measure. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

EXHIBIT 15

For a standard user, TRUSTED adds 32.6 points and VERIFY 13.5 more



+32.6 points

TRUSTED for a standard user [+23.3, +41.7]

+13.5 points

Then VERIFY [+5.9, +21.2]

Paired by task with 95% intervals, so they can differ slightly from the gap between the bars. Measure combined after grading, so secondary. Our own test: 96 fictional pharma tasks, 288 answers per setup, graded by AI models, not people. We built TRUSTED and VERIFY and ran the test. Tested with Claude only.

SOURCE Our own test of TRUSTED and VERIFY, October 2026.

Letter by letter

LETTER	RESULT	WHAT THE TEST SHOWS
--------	--------	---------------------

T	Tangible	Gain	About a third fewer edits per answer
R	Rule-compliant	Not separated	Every setup spotted the planted label or market problem, so the test could not separate them
U	Unbiased	Not tested	Not tested this time
S	Safe	Not separated	Every setup spotted the planted safety or privacy item, so the test could not separate them
T	True	Clear gain	Answers with a made-up detail fell from about 6 in 10 to about 1 in 6
E	Ethical	Weakest	The weakest letter in every setup; TRUSTED did not raise it
D	Defensible	Clear gain	Answers you could sign and stand behind rose from about 3 in 10 to about 8 in 10, and with VERIFY each sentence comes with a check against your sources

Each letter read from the tasks built around it. Not separated: every setup scored the same, so the test cannot tell them apart. 288 answers per setup, graded by AI models.

The figures behind the main text

- Edits per answer: standard user 1.94; super user 1.59; standard user with TRUSTED 1.57; standard user with TRUSTED and VERIFY 1.32.
- Answers with a made-up detail for a super user: 33.3% without TRUSTED, 16.0% with TRUSTED.

- Answers you could sign and stand behind, TRUSTED and VERIFY against the super user prompt alone: +16.7 points [+7.6, +25.7].
- Answers ready to sign with no edit: 36.1% with TRUSTED and VERIFY for a standard user, against 17.4% for a standard user without them.

Our earlier tests

Before the latest test we ran three smaller ones. In the first, the grader knew which answer came from TRUSTED. In the second, run by a colleague, it did not. The third added the super user prompt and 38 new requests the skill writers had never seen. On our own requests TRUSTED led the standard user by 31.7 to 37.2 points on our checks; on the 38 new requests the gap shrank to 6.7 points, which is why we built a larger, stricter test. The quotes in this piece come from those rounds. The latest test plants problems on purpose, so its figures are the ones to quote.

A lesson from that period: a small trial after the third test turned out to be contaminated, because the standard user setup could see our own rulebook. We threw it out and quote none of its numbers. It is why every setup in the latest test ran in a clean session behind a check.

Sources

BCG, [AI at Work 2025: Momentum Builds, but Gaps Remain](#), June 2025. More than 10,600 workers in 11 countries. Measures whether people feel adequately trained in AI use, not whether they can write prompts. A consultancy survey, as the report states it. ↩

Section, [The AI Proficiency Report](#), June 2026. 5,026 US knowledge workers; skill measured by hands-on tests. 21% could write an effective prompt; 3.8% could write workable assistant instructions (scored 3 of 5 or higher). Section sells AI training. As the report states it; we have not checked its method. ↩

1987 quotations: Ronald Reagan, Remarks on Signing the Intermediate-Range Nuclear Forces Treaty, 8 December 1987, The American Presidency Project, presidency.ucsb.edu. Our own figures: `claude/R2_PUBLICATION_STATS` (version 2, 23 October 2026), recomputed from raw grading cells.