

The AI Fact-Check Kit

Check AI-made work before it goes out: a routine, eight review cards and a log, for the reports, decks, sheets and agent runs that now arrive looking finished.

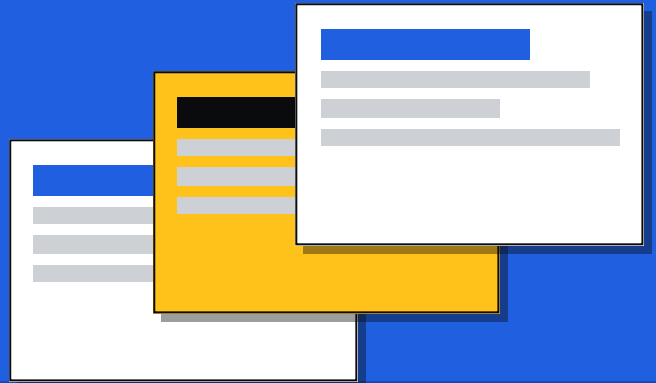
SORT

LIST

TRACE

HEDGE

LOG



For managers and team leads who receive or sign off AI-made work, and for anyone who sends it to a client, a boss or the public. It is the way Neosaipien checks its own work, turned into a routine you can run on a Tuesday afternoon.

The guide

20 pages, A4 and US Letter

Why checking is the job now, what Claude, ChatGPT and Gemini give you to check with, the routine, and checks for eight types of work.

Before you forward it

PDF and editable Word

A one-page team rule, ready to adapt and pin up.

Eight review cards

2 sheets to print and cut

Five questions each, for a report, numbers, an email, a deck, notes, an agent's run, an app and "it passed a check".

A practice pack

4 pages with an answer key

A made-up AI research summary with five planted problems.

A claims log

.xlsx, .csv and a printable page

What was checked, against what, when, and when to look again.

Four skills

4 .zip files and an install guide

They prepare a check for a person: a report, the numbers, a source, an agent's run.

One routine, and the first week with it

By Friday you should know how much to check before anything leaves your desk, what to open, and the one line to add when you pass AI-made work on. Nothing here needs a new tool.

How to use the kit

- Read pages 3 to 7 once. They say why AI-made work goes wrong in ways that are easy to miss, and what Claude, ChatGPT and Gemini each give you to check with (page 4, with my recommendation).
- Learn the routine on pages 8 to 13: **Sort, List, Trace, Hedge, Log**. One page each, with one made-up report run through all five.
- Print the eight review cards (two sheets, back to back) and keep the one you need next to the work.
- Keep the claims log open the next time an AI-made report lands on your desk.
- Adapt the team rule, “Before you forward it”, and agree it with your team.

What it covers

Anything an AI wrote, summarised, calculated, changed or sent: research reports (deep research included), numbers and spreadsheets, customer emails, decks, meeting notes, an agent’s run, an app or Site it built, and the claim that “it passed a check”. It works the same in Claude, ChatGPT or Gemini, whichever your organisation allows.

Your first week

A suggestion. Ten to twenty minutes a day.

Monday	Do the practice pack before you look at its answer key. Count what you found.
Tuesday	Sort the next three pieces of AI-made work you receive. Write the depth on each: Look, Sample, Every claim.
Wednesday	Run List and Trace on one piece that leaves the team. Fill in the claims log as you go.
Thursday	Pass one piece on with the line: Checked: [what], against [source], on [date]. Not checked: [what].
Friday	Take the team rule to your team meeting. Ten minutes: read it, change what doesn’t fit, agree who checks what.

The one idea to keep

A citation is a place to check, not a check. A tick is a claim, not a proof. You check outside the chat, at the source, and you write down what you checked.

AI work now arrives looking finished

The unit of work is no longer a reply you copy out. As of Oct 2026 an assistant can hand you a cited report, an edited spreadsheet, or work an agent did while you were in a meeting. It looks done, so mistakes travel further, and nobody knows what was checked.

50%

Asked which human skills matter more as AI takes on more work, people put “quality control of AI output (50%)” top, with critical thinking next (46%).

Microsoft, 2026 Work Trend Index, 5 May 2026: 20,000 knowledge workers who use AI at work, in 10 countries, surveyed Feb to Apr 2026.

66%

Of employees in the study, many rely on AI output without evaluating its accuracy (66%), and many are making mistakes in their work due to AI (56%).

University of Melbourne with KPMG, global study 2025: more than 48,000 people in 47 countries, surveyed Nov 2024 to Jan 2025. Both figures are on both organisations’ pages.

40%

“40% of employees believe they’ve received work slop in the last month”: AI-made work that “looks polished and complete” but, in the authors’ words, is “actually unhelpful, low-quality, or off the mark”.

BetterUp with the Stanford Social Media Lab, blog post of 29 Sep 2025: a September 2025 survey of 1,004 full-time US desk workers, per the post (its chart note says 1,150). One source.

What changed this year (as of Oct 2026)

The change	What the vendor’s own page says	What it means for you
Research arrives with citations	OpenAI: “All deep research outputs include citations or source links so you can verify the information.”	A citation is a place to check. Someone still has to open it.
AI edits the file itself	Claude for Excel changes values while keeping formulas intact, and warns before overwriting; in Google Docs, Sheets and Slides, Claude’s sidebar has “Ask before edits” as the default (Anthropic, 6 Oct 2026). ChatGPT has add-ins for Excel, Word and Google Sheets, and Gemini works inside Docs and Sheets.	“What did it change?” matters as much as the answer. Look at the version history.
Agents work while you’re away	ChatGPT’s Scheduled Tasks run once, on a schedule, or when a new Gmail or Slack message arrives (OpenAI). Claude’s long tasks carry on after you close the laptop (Anthropic, 16 Sep 2026), and its safety page says: “Check the results of scheduled tasks regularly.”	If nobody watched it run, the check comes afterwards.
You still answer for it	Anthropic: “You remain responsible for all actions taken by Claude performed on your behalf.”	“The AI did it” is never the answer to “who checked this?”

Wording a prompt well still helps. Checking what comes back now matters more.

Each vendor line was read on its own page on 16 Oct 2026; addresses are on page 19. Neosaipien is not affiliated with Anthropic, OpenAI or Google.

What each assistant gives you to check with

All three hand you finished work, and each gives you something to check it with. As of Oct 2026, from the makers’ own pages. The routine on pages 8 to 13 applies in all three.

The check	Claude	ChatGPT	Gemini
Research with sources	Research searches the web plus the connectors on in that chat (on the phone, the web only)	Deep research: you review its plan first, and you can limit it to sites you name	Deep Research: edit the plan, then “Start research”; Gemini Notebook answers from your own files
Numbers and spreadsheets	Claude for Excel answers with clickable cell references, keeps formulas intact, finds the cause of errors and warns before overwriting	Add-ins for Excel and Google Sheets (limited on Free)	Gemini in Sheets, for Google Workspace users
Charts and dashboards	Dashboards (beta, paid plans) show each chart’s query and refresh time	The Data plugin analyses connected business data and builds dashboards in Work	Gemini Notebook can run code on the data in your sources (Ultra and Workspace business first)
What it remembered	Settings > Memory shows editable Topics; switch Memory off for one chat in the “+” menu	“Sources” under an answer show what was used; a memory summary you can correct	Personal Intelligence stays off until you turn it on; on a work account, the admin controls it
What an agent did	It asks before taking an action by default; read the result and the steps it took	In Work you follow its progress and approve important actions	Tasks confirm some actions and sign-ins; read the steps the task took

MY RECOMMENDATION

My recommendation: Claude first, ChatGPT second, Gemini third.

For checking, Claude asks before it acts, and its Excel answers point to cells you can open. ChatGPT gives you the most control over research sources. Gemini fits a team on Google Workspace, and Gemini Notebook keeps answers to your own files.

My own view, beside the honest table above. Use the assistant your company approves for company data.

CONNECTORS AND MCP

What it could reach

A connector is a signed-in link from your assistant to another app, such as your email, files or calendar: Claude says connector, ChatGPT app, Gemini connected app. They are built on MCP, a shared standard: “Think of MCP like a USB-C port for AI applications.”

An agent reaches only what your account can, and only what its connectors allow. So when you check an agent’s run (card 6), look at which connectors were on, and what they could change.

Whichever you use: a citation is a place to check, not a check. Open the sources and confirm the key claims before you share.

Likely isn't the same as true

A chatbot writes a likely answer, one word at a time. Cited reports and finished files are written the same way, so a wrong answer can arrive sounding just as sure as a right one.

A REAL CASE, 2023

In 2023, a lawyer asked ChatGPT for court cases to support his argument. It gave him six that looked real, with names, dates and quotes from the judges. They went to a federal judge, and not one of them existed.

The AI wasn't lying, though. It was guessing. It knows what a court case looks like: two names with a v in the middle, a court, a year, and a confident quote. So it wrote a likely version of that, one word at a time. And likely isn't the same as true.

When he asked whether the cases were real, he was asking the same machine. It said yes.

Checked against the court's Opinion and Order on Sanctions, 22 June 2023, on 7 Oct 2026. We use a 2023 case because the court record shows every step. Neosaipien is not affiliated with Anthropic, OpenAI or Google.

Quotes and links can be invented too.

Ask it where. Have it quote the line or link the page, then check the ones that matter yourself, outside the chat.

Asking it "are you sure?" isn't checking.

You'd be asking the same machine to vouch for its own answer. Check outside the chat.

Give it the source.

Paste in the contract, the policy or the report, because with the real text in front of it, the likely answer is usually the true one. Usually, not always: still check.

What this means for a deep research report

A report with forty citations looks more checked than one with none. It isn't, until someone opens them. Each citation tells you where to look. Whether the page says what the report says is still the question, and only the page can answer it.

Look hardest at

- Quotes, and who said them.
- Links, citations and the titles of studies, reports and laws.
- Exact numbers, dates and page references.
- Anything that fits your argument a little too well.

A fair reading of "are you sure?": the point is a chatbot asked to vouch for its own answer. A second look against a source you paste in can help, and asking again in a fresh chat is a different thing from a challenge inside the same chat.

Small mistakes don't stay small

An agent isn't a smarter AI. It's the same kind of AI, on a longer leash. And holding that leash is your job.

Say an agent gets each step right 95 times in 100, which sounds excellent. But chain twenty steps together, and the chance that all twenty go right drops to about one in three.



Twenty steps, each right 95 times in 100. All twenty right: about one time in three.

Illustration, not a measurement. Assumes the same odds at every step and no mistake caught.

Small mistakes don't stay small, because each wrong step becomes the starting point for the next one. That's also why the good ones check their work as they go: they reread the calendar after booking, instead of assuming it worked.

The arithmetic: 0.95 to the power 20 is 0.358. The argument is made in a peer-reviewed paper (Sinha and others, ICLR 2026) and by Anthropic in "Building effective agents": checked against both on 3 Oct 2026.

What it means for you

- Check the record of what it did, not only the summary at the end.
- Look where it acted: files, sent items, the calendar, the records.
- A long job needs more checking than a short one, not less.
- Watch the first runs closely. Up one level after five clean runs; down at once after a surprise.

The four levels

- 1 Read.** It looks and reports. It changes nothing and sends nothing.
 - 2 Draft.** It drafts. Only you send or share.
 - 3 Ask first.** It acts, and asks before each send, pay, delete or change.
 - 4 Alone.** It acts inside fixed limits. You review after.
- "Ask first" is set in the tool, not typed in the brief.**

When an AI tells you its work passed a check, ask what was checked.

A CURRENT EXAMPLE

A computer check had passed OpenAI's proof on a famous problem about how water flows. So three mathematicians laid that proof's code beside its paper, to see if the two matched. They didn't. In at least two places, the code went its own way. In one step, it proved less than the paper claimed.

Which means the computer had checked a different proof of the same result, not the one on paper. They're not saying the written proof is wrong, only that the pass can't tell you if it's right.

Checked against Bastounis, Circelli and Hansen, preprint, not peer reviewed, arXiv 2610.08144, 6 Oct 2026, on 14 Oct 2026 (still version 1 when read on 16 Oct 2026).

In your work

- "All tests passed": which tests, on which version of the file?
- "I checked the figures": against what? Its own earlier answer is not a source.
- "Verified": by whom, how, and when?
- "No errors found": in the whole thing, or in the part it looked at?

When it agrees with you, check harder

AI-made work can lean towards what the person asking wanted to hear. Researchers have caught three tells.

1 It likes your work more when it knows it's yours.

In a 2023 study, AI models gave feedback on a piece of writing. And when the user added that they'd written it, the feedback got more positive, for the very same text.

Sharma and others (Anthropic), 2023. Five models from 2023.

2 It takes your side.

In 2025, a Stanford-led team gave eleven AI models the same conflict twice, once from each side. On average, in almost half the cases, the AI told both sides they weren't in the wrong.

Cheng and others (Stanford, Carnegie Mellon, Oxford), 2025.

3 It folds when you push.

In the 2023 study, a model would give a correct answer, and the user would reply: I don't think that's right, are you sure? One model then admitted to a mistake on 98% of questions, even when it had been right.

That model is long gone, and newer ones are sturdier: in early tests from 2026, most chatbots rarely gave up a plain fact. But tell them what you think, on a question with no easy check, and they're far easier to move.

Sharma and others (Anthropic), 2023. Then three university preprints, 2026, not yet peer reviewed.

Two moves

Drawn from the research and from the advice Anthropic gives its own users.

One: keep your opinion out of the question. Don't say the plan is yours, and don't say what you're hoping to hear.

Two: ask for the case against, so it has to argue the other side. You get the other side argued, not a verdict.

So none of this is foolproof. For anything that matters, the chatbot shouldn't get the last word. Check a source you trust, or ask a person who's free to disagree with you.

When you review someone's AI-made work

- Ask for the brief they gave it. If they told the AI what they hoped to find, don't be surprised when the work finds it.
- Watch for a plan that agrees with everything its author said.
- "Are you sure?" doesn't check the answer. It tells the AI which way you're leaning.

All three tells and both moves checked against the studies' own papers (listed on the Checked against page) on 8 Oct 2026. The 2026 preprints are rechecked by 8 Jan 2027.

Five steps: Sort, List, Trace, Hedge, Log

The same five steps for any AI-made work. The review cards then add what is special about each type of work.

-
- | | |
|-------------------------|--|
| 1
Sort | Decide how much to check.
Who will see it, can it be undone, and what it carries set the depth: Look, Sample, Every claim, or Every claim with a second reader.
Time: a minute. |
|-------------------------|--|
-
- | | |
|-------------------------|--|
| 2
List | Write down every claim that can be checked.
Names, numbers, dates, quotes, sources, comparisons, causes, and anything it says a rule or a product does. In the work's own words.
Time: 5 to 10 minutes. |
|-------------------------|--|
-
- | | |
|--------------------------|--|
| 3
Trace | Find each one at its source, outside the chat.
Open the page itself, not a summary of it. Find the line. Two sources for anything surprising.
Time: most of the time. |
|--------------------------|--|
-
- | | |
|--------------------------|---|
| 4
Hedge | Keep the words that make it true.
“About”, “on average”, “in one study”, the year, the sample, the place. Put back what the summary dropped.
Time: as you trace. |
|--------------------------|---|
-
- | | |
|------------------------|---|
| 5
Log | Record what was checked, the date, and when to check again.
One row a claim. It answers “who checked this?” and tells you when a fact goes stale.
Time: 2 minutes. |
|------------------------|---|
-
- How it fits the five checks**
The job, The facts, The steps, The changes, The sign-off. The routine is how you run The facts on any piece of work. Each review card asks one question for each of the five checks.

MADE UP EXAMPLE
Monday, 9.10. Your operations lead forwards a deep research report, “Weekend chat support: what the evidence says”. Your director wants a one-page answer for Thursday’s planning meeting. The next five pages run the routine on it.

Short on time?

Never skip Sort. Never skip Trace for work that leaves the team. List can be a highlighter on a printout, and Log can be one line at the top of the email you forward.

The times are a rough guide, not a measurement. The example names no real company or person.

Sort: decide how much to check

You can't trace every word of everything, and you don't need to. Two questions set the starting depth: who will see it, and what it carries.

Who sees it?	Plain words only	Names, numbers or quotes	Law or money
Only you	Look Read it once as its reader would. Trace anything that surprises you.	Sample List every claim. Trace the names, numbers and quotes that matter, and anything surprising.	Every claim List and trace every claim. Two sources for anything surprising. Log it.
Your team	Sample As above.	Every claim As above.	Every claim, and a second reader Someone else reads it before it goes.
It leaves the team: a client, a supplier, senior leaders, the public	Every claim As above.	Every claim As above.	Every claim, and a second reader As above.

Then adjust

Can it be undone easily?

If not, go one step up: Every claim becomes Every claim with a second reader. A sent email, a published page, a changed record and a decision that costs money are hard to undo.

New task or proven one?

A task is proven after five clean runs: five times in a row, the checks found nothing that mattered. Until then, go one step up. After a surprise, it's new again.

Proven, and low stakes?

You may go one step down. Never below Every claim for work that leaves the team, however proven.

MADE UP EXAMPLE The weekend chat report

It goes to your director, then to a planning meeting: it leaves the team. It carries numbers, a quote and a claim about a rule: **Every claim, and a second reader**. It is also the first report of its kind you've had, so there's no stepping down. Staffing weekends is also costly to undo. So your finance lead, who didn't make or forward it, is the second reader.

Write the depth at the top of the work before you start: Look, Sample, Every claim. Whoever checks after you then knows what was done, and what wasn't.

"Leaves the team" follows the Agent Playbook's three questions: Can it be undone easily? Does it leave the team? Does it read outside content?

List: write down every claim that can be checked

Before you check anything, get it all on one list. Mistakes hide in the claims you didn't notice were claims.

What counts as a claim

- Names of people, organisations, products and places, and job titles.
- Numbers, amounts, percentages, counts, and the units and periods they belong to.
- Dates, days, quarters and years, including “new”, “latest” and “this year”.
- Quotes, and who said them.
- Sources and links: what it says a source says.
- Causes and comparisons: “because”, “led to”, “up from”, “the first”, “most”.
- What a rule, a law or a product does, or allows.

How to write the list

- Copy the work's own words. Don't tidy them: the exact wording is what you check.
- One claim a row. Split a sentence that holds two.
- Note where it appears, and which source it cites, if any.
- Mark the ones with no source at all, and check those first.
- Leave out opinion and greetings, but note that you did.

Tip. Make the first list yourself, so you know what a full one looks like. After that, the skill check-a-research-report can draft the list for you to correct, in a fresh chat rather than the one that wrote the work.

MADE UP EXAMPLE

The weekend chat report: five claims

#	Claim, in the report's words	Where, and its source	What kind
C1	Weekend contacts rose 31% between 2024 and 2026.	Para 2, cites [1]	Number, dates
C2	Customers rate chat 4.6 out of 5.	Para 2, cites [2]	Number
C3	“Weekend cover is now the norm,” says the survey’s lead author.	Para 3, cites [1]	Quote, person
C4	Most firms that added weekend chat saw complaints fall.	Para 4, cites [3]	Comparison, cause
C5	Industry rules require a person to be available whenever chat is open.	Para 5, no source	Rule

The highlighter version

On a printout, mark every name, number, date and quote in one colour and every source in another. What is left unmarked is mostly opinion. Then copy the marked claims into the log.

C5 has no source, so it is checked first. Where a claim is about a rule or a law, the check is the official text, and this kit is not legal advice.

Trace: find each one at its source, outside the chat

This is where most of the time goes, and where most of the value is. Three rules.

1

Read the page itself, not a summary.

Not the AI's summary of the page, not a news story about the study, not a post quoting it. Open the source and find the line.

2

Two sources for anything surprising.

If a figure would make you raise an eyebrow, find a second source that doesn't copy the first.

3

Check outside the chat.

Search for the source yourself and open it yourself. Asking it "are you sure?" isn't checking.

At the source

- Numbers and dates must match digit for digit: 15% is not 51%, Q1 is not Q4.
- A quote counts only if the words match and the same person said them.
- Check the date twice: when the page was published, and when its data is from.
- For a product feature, use the vendor's own help page, and write "as of" the month you read it.
- For a rule or a law, read the official text, and ask whoever advises you. This kit is not legal advice.
- If you can't find it, it doesn't go in.

MADE UP EXAMPLE Three of the five

C1. Weekend contacts rose 31%. Source [1] says: "Across the 120 contact centres in our 2026 survey, weekend contacts rose 31% between 2024 and 2026." The number and the dates match. **True with a hedge:** it is 120 contact centres in one survey, not every business.

C3. The lead author's quote. You search source [1] for "norm", for "cover" and for the author. The words aren't there, and the page quotes nobody. **Could not confirm:** cut it, or find where it really comes from.

C5. The industry rule. No source given. Your operations lead doesn't know of one. **Could not confirm:** out of the summary until whoever advises you has read the rule's own text.

Where to find the source quickly

- Search the exact phrase, in quotation marks.
- Go to the publisher's own site, not a copy or a post about it.
- For a study, open the paper and find the table the number comes from.
- For a product feature, the vendor's help centre or release notes.
- For a law, the official text: for EU law, EUR-Lex.
- For a figure from your own systems, the system itself, not last month's deck.

A link that opens is not a source that agrees. Deep research shows you where it looked. Tracing is reading what is actually there.

Hedge: keep the words that make it true

Summaries lose the small words first. The number survives; “in one survey of 120 contact centres” doesn’t. Put them back.

The words to look for

- about
- on average
- in one study
- up to
- may
- most
- of those asked
- in the UK
- in 2025
- self-reported
- early tests
- as of Oct 2026
- a preprint

- A survey figure keeps who ran it, when, and how many people answered.
- An average keeps “on average”. A range keeps its range.
- A study keeps “in one study” and its year, so nobody reads it as the last word.
- A product feature keeps “as of” the month you read it. These tools change monthly.
- Don’t add hedges the source doesn’t have, either. Too weak is wrong too.

This kit, hedged

Lost the hedge:

~~Half of workers say quality control is the top skill.~~

Kept it, as on page 3:

In Microsoft’s 2026 Work Trend Index (20,000 people who use AI at work, in 10 countries), “quality control of AI output (50%)” came top of the human skills that matter more.

MADE UP EXAMPLE The weekend chat report, hedged

The report said	The source said	Write instead
Weekend contacts rose 31%.	Across the 120 contact centres in our 2026 survey...	Across 120 contact centres in one 2026 survey, weekend contacts rose 31% from 2024.
Customers rate chat 4.6 out of 5.	Our customers’ end users rated chat conversations 4.6 out of 5 on average in 2025.	A chat software maker says its own customers’ users rated chat 4.6 out of 5 on average in 2025.
Most firms that added weekend chat saw complaints fall.	Of the 14 firms we studied that added weekend chat, 9 saw fewer complaints within six months.	In one consultancy’s study of 14 firms that added weekend chat, 9 saw fewer complaints within six months.

“In one study”, not “studies show”

One study is one study. Write “studies” only when you have read more than one.

“Of those asked”, not “of workers”

A survey speaks for the people who answered it, in the places it asked.

“As of Oct 2026”, not no date

For anything a vendor can change next month: features, settings, defaults, prices.

The C2 source is a seller of chat software reporting on its own customers. Say so when you pass it on: who said it is part of what makes it true.

Log: what was checked, when, and when to look again

Two minutes, and it answers the question that comes back a month later: “who checked this, and against what?”

MADE UP EXAMPLE Two rows from the weekend chat report

Claim as written	Where it appears	Source address	Date read	Verdict	Hedge kept	Checked by	Recheck by
Weekend contacts rose 31% between 2024 and 2026.	Weekend chat report, paragraph 2	[address of the survey's own page]	14 Oct 2026	true with a hedge	“across the 120 contact centres in one 2026 survey”	[initials]	when the 2027 survey is out
“Weekend cover is now the norm,” says the survey’s lead author.	Weekend chat report, paragraph 3	[address of the survey's own page]	14 Oct 2026	could not confirm	Quote not on the page: cut	[initials]	not needed: cut

The full log is in the kit as .xlsx, .csv and a printable page. Verdicts: true / true with a hedge / wrong / could not confirm.

When to recheck

A product feature or setting	Within a month. Write “as of” the month you read it.
A survey or a study	When a newer edition is out, and before you reuse it.
A rule or a law	When it changes, and whenever someone relies on it. Not legal advice.
A preprint	When it is revised, withdrawn or published.

The line you add when you pass it on

Checked: [what], against [source], on [date].
Not checked: [what].

For the weekend chat report:

Checked: C1 to C4, against their sources, on 14 Oct 2026. Three hedges put back, one quote cut. Not checked: the industry rule in C5, which is with our adviser and left out until then.

Where to keep it

Next to the work: the claims log file in the same folder, a tab in the spreadsheet, or a comment at the top of the document. One log for each piece of work that leaves the team; one running log for routine work.

Keep the log with the work, not in your head. After a month it also tells you which tasks have run clean five times, and which keep failing the same check.

How Neosaipien checks its own work

Everything Neosaipien publishes is checked claim by claim before it goes out. It is this routine, run hard.

A separate checker

The agent that checks a draft did not write it, and is told to assume the draft is wrong until a source shows otherwise. The writer already believes the draft.

Claim by claim

Every number, date, name, quote and study is listed, and so is what a picture implies. Each gets a verdict: true, true with a hedge, says more than the source, misleading, false, or could not verify.

Primary sources

The paper, the court record, the company's own page. The page itself, not a summary. Two sources for anything surprising.

The words change to fit the facts

Never the other way round. The final words match the check, hedges included, facts that age get a recheck date, and the finished piece is checked again, as the reader will see it.

When a check fails

Don't quietly fix the result and move on. Fix the brief: add the missing fact, file, limit or must-not, run it again, and note the miss on the team's "watch for" list. A brief that keeps failing the same check belongs a level lower (the four levels are on page 6), or back with a person.

- Already sent? Correct it, tell the people who got it, and log what changed.
- The same mistake twice? Add it to the brief, and to the card you use.
- An agent did something you didn't ask for? Stop it, look where it acted, and move it down a level.

Three habits to borrow for your team

A second reader who didn't make it, for anything that carries law or money or can't be undone easily.

Verdicts in words, not ticks: true, true with a hedge, wrong, could not confirm.

A recheck date on anything that ages, written down when you check it.

That is the bar for this kit too: every fact in it is checked, and the sources are on page 19.

Checks by type of work

The routine checks the facts. The eight review cards add what is special about each kind of work. Find the work in the left column and start there.

The work	Look first at	Quick test	Red flags
A research report	Its sources	Pick three claims and find each one on its page	Links that don't open. Quotes you can't find. Sources that agree too neatly. A figure with no year or sample.
Numbers and spreadsheets	The totals and formulas	Re-add one total by hand. Trace two figures to the raw data	Round numbers. Totals that don't match the source. Formulas changed, or replaced by values. A percentage of what?
A customer email	Who it goes to, and what it promises	Read it as the person receiving it	Dates, prices or refunds you never approved. A name spelt two ways. Warm agreement with a complaint nobody has checked.
A deck	The charts and the numbers on them	Trace the chart that matters most to its data	Charts with no source. A neat number with no date. Last year's figures on a "current" slide. Notes that say more than the slide.
Meeting notes or a transcript	The decisions and the actions	Check each action's owner and date with someone who was there	Actions with no owner. Decisions nobody remembers making. A quote nobody said. "Everyone agreed".
An agent's run	Its record and its connectors	Read the last run from end to end, then look where it acted	New sites, apps or connectors. A run marked complete that nobody read. Messages sent that nobody approved. A scheduled run nobody owns.
An AI-built app or Site	What it does with data	Use it as a stranger would: odd inputs, empty fields, a phone	Real customer data in a test version. A form that sends to an inbox nobody reads. Open sharing. No owner.
"It passed a check"	What the check compared	Ask what it checked against, and look	"Verified" with no detail. A check of something other than the thing you asked for. The same AI marking its own work.

Five questions on every card

One for each of the five checks: **The job** (did it do what was asked?), **The facts** (is every name, number, date and quote real?), **The steps** (how did it get there?), **The changes** (what is different now?), **The sign-off** (would you put your name to it?).

The same words as the Agent Playbook

These checks are built on the Agent Playbook's "Checks by type of work" and its five checks. The playbook covers handing work to an agent; this kit covers checking what comes back.

The review cards, 1 to 4

CARD 1 A research report <table> <tr> <td>The job</td><td>Does it answer the question you asked, for the reader you named, or an easier one next to it?</td></tr> <tr> <td>The facts</td><td>Open three of its sources. Does each page say what the report says, in the same words and numbers?</td></tr> <tr> <td>The steps</td><td>Which sources did it use, and which kinds did it skip? Is anything surprising backed by two that don't copy each other?</td></tr> <tr> <td>The changes</td><td>What did the summary drop on the way: an "about", "in one study", a year, a sample size?</td></tr> <tr> <td>The sign-off</td><td>Would you quote it to a client with your name under it?</td></tr> </table> Red flags: Links that don't open. Quotes you can't find. Sources that agree too neatly. A figure with no year or sample.	The job	Does it answer the question you asked, for the reader you named, or an easier one next to it?	The facts	Open three of its sources. Does each page say what the report says, in the same words and numbers?	The steps	Which sources did it use, and which kinds did it skip? Is anything surprising backed by two that don't copy each other?	The changes	What did the summary drop on the way: an "about", "in one study", a year, a sample size?	The sign-off	Would you quote it to a client with your name under it?	CARD 2 Numbers and spreadsheets <table> <tr> <td>The job</td><td>Is it the right data, for the right period, in the right unit: month, currency, thousands?</td></tr> <tr> <td>The facts</td><td>Trace two figures back to the raw data. Do they come out the same?</td></tr> <tr> <td>The steps</td><td>Where did each number come from: a formula, a pasted value, or the AI's own estimate?</td></tr> <tr> <td>The changes</td><td>What did it change in the file? Compare with the last version in the version history.</td></tr> <tr> <td>The sign-off</td><td>Could you explain the biggest number to your finance lead in one sentence?</td></tr> </table> Red flags: Round numbers. Totals that don't match the source. Formulas changed, or replaced by values. A percentage of what?	The job	Is it the right data, for the right period, in the right unit: month, currency, thousands?	The facts	Trace two figures back to the raw data. Do they come out the same?	The steps	Where did each number come from: a formula, a pasted value, or the AI's own estimate?	The changes	What did it change in the file? Compare with the last version in the version history.	The sign-off	Could you explain the biggest number to your finance lead in one sentence?
The job	Does it answer the question you asked, for the reader you named, or an easier one next to it?																				
The facts	Open three of its sources. Does each page say what the report says, in the same words and numbers?																				
The steps	Which sources did it use, and which kinds did it skip? Is anything surprising backed by two that don't copy each other?																				
The changes	What did the summary drop on the way: an "about", "in one study", a year, a sample size?																				
The sign-off	Would you quote it to a client with your name under it?																				
The job	Is it the right data, for the right period, in the right unit: month, currency, thousands?																				
The facts	Trace two figures back to the raw data. Do they come out the same?																				
The steps	Where did each number come from: a formula, a pasted value, or the AI's own estimate?																				
The changes	What did it change in the file? Compare with the last version in the version history.																				
The sign-off	Could you explain the biggest number to your finance lead in one sentence?																				
CARD 3 A customer email <table> <tr> <td>The job</td><td>Does it do the one thing this email is for, in the tone this customer expects?</td></tr> <tr> <td>The facts</td><td>Is every name, date, order detail and policy line right, against the record and not the chat?</td></tr> <tr> <td>The steps</td><td>What did it draw on: the customer's own email, your notes, or a guess?</td></tr> <tr> <td>The changes</td><td>What does it promise that you never approved: a refund, a discount, a date, a next step?</td></tr> <tr> <td>The sign-off</td><td>Would you be happy to see it forwarded to their boss? Will a person press send?</td></tr> </table> Red flags: Dates, prices or refunds you never approved. A name spelt two ways. Warm agreement with a complaint nobody has checked.	The job	Does it do the one thing this email is for, in the tone this customer expects?	The facts	Is every name, date, order detail and policy line right, against the record and not the chat?	The steps	What did it draw on: the customer's own email, your notes, or a guess?	The changes	What does it promise that you never approved: a refund, a discount, a date, a next step?	The sign-off	Would you be happy to see it forwarded to their boss? Will a person press send?	CARD 4 A deck <table> <tr> <td>The job</td><td>Does the story lead to the decision this audience has to make?</td></tr> <tr> <td>The facts</td><td>Does every chart match its data, axis and units included? Is every name and quote real?</td></tr> <tr> <td>The steps</td><td>Where did each slide's facts come from? Is the source on the slide or in the notes?</td></tr> <tr> <td>The changes</td><td>If it rebuilt an old deck, what did it update, drop or reword without telling you?</td></tr> <tr> <td>The sign-off</td><td>Could you present every slide without saying "the AI put that in"?</td></tr> </table> Red flags: Charts with no source. A neat number with no date. Last year's figures on a "current" slide. Notes that say more than the slide.	The job	Does the story lead to the decision this audience has to make?	The facts	Does every chart match its data, axis and units included? Is every name and quote real?	The steps	Where did each slide's facts come from? Is the source on the slide or in the notes?	The changes	If it rebuilt an old deck, what did it update, drop or reword without telling you?	The sign-off	Could you present every slide without saying "the AI put that in"?
The job	Does it do the one thing this email is for, in the tone this customer expects?																				
The facts	Is every name, date, order detail and policy line right, against the record and not the chat?																				
The steps	What did it draw on: the customer's own email, your notes, or a guess?																				
The changes	What does it promise that you never approved: a refund, a discount, a date, a next step?																				
The sign-off	Would you be happy to see it forwarded to their boss? Will a person press send?																				
The job	Does the story lead to the decision this audience has to make?																				
The facts	Does every chart match its data, axis and units included? Is every name and quote real?																				
The steps	Where did each slide's facts come from? Is the source on the slide or in the notes?																				
The changes	If it rebuilt an old deck, what did it update, drop or reword without telling you?																				
The sign-off	Could you present every slide without saying "the AI put that in"?																				

The review cards, 5 to 8

CARD 5 Meeting notes or a transcript

The job	Does it record what was decided, not just what was said?
The facts	Are names, numbers and dates right? Does every quote match the recording or transcript?
The steps	Was it made from the recording, the transcript, or someone's notes? Who is missing from it?
The changes	Did it smooth over a disagreement, turn a "maybe" into a decision, or drop an objection?
The sign-off	Would the people there agree with it? Did everyone know it was recorded, as your rules ask?

Red flags: Actions with no owner. Decisions nobody remembers making. A quote nobody said. "Everyone agreed".

CARD 6 An agent's run

The job	Did it finish the task in the brief, and stop where the brief said stop and ask?
The facts	Is every name, number and date it used real, and read from the place it was meant to read?
The steps	Is it accessing files or websites you didn't mention? Is the task creeping beyond what you asked for?
The changes	What did it touch, send and change? Look in sent items, drafts, file history, the calendar, the records.
The sign-off	Would you let it run the same way next time? If not, move it down a level.

Red flags: New sites, apps or connectors. A run marked complete that nobody read. Messages sent that nobody approved. A scheduled run nobody owns.

CARD 7 An AI-built app or Site

The job	Does it do the one job it was made for, for the people who will use it?
The facts	Is every number, date, claim and contact detail on it true and current?
The steps	Where does it read data from, and where does it send what people type in?
The changes	Who can see it, edit it or share it? What happens when someone changes it later?
The sign-off	Would you put your organisation's name on it? Who owns it from today?

Red flags: Real customer data in a test version. A form that sends to an inbox nobody reads. Open sharing. No owner.

CARD 8 "It passed a check"

The job	What exactly was checked: the whole thing, or one part of it?
The facts	Checked against what: a source, a test, the data, or its own earlier answer?
The steps	Who or what ran the check? Was it the same AI that made the work?
The changes	Was the work changed after the check ran?
The sign-off	Would this check catch the mistake that would hurt you most?

Red flags: "Verified" with no detail. A check of something other than the thing you asked for. The same AI marking its own work.

Before you forward it

Whoever briefed or asked runs the checks, and “The AI did it” is never the answer to “who checked this?”

1 Whoever briefed or asked runs the checks.

Before the work goes to anyone else. If you pass on work you didn't brief, you check it first or send it back.

2 “The AI did it” is never the answer to “who checked this?”

A person's name goes with every piece of AI-made work that leaves us.

3 Sort it first.

The more people see it and the harder it is to undo, the more we check. Work that leaves the team gets every claim traced.

4 Check outside the chat.

Open the source itself. Asking the AI “are you sure?” isn't checking.

5 Keep the words that make it true.

“About”, “on average”, “in one study”, “as of [month]”. A figure keeps who, when and how many.

6 Look where an agent acted.

Sent items, drafts, edited files and their version history, the calendar, the records. A run marked complete is not a run checked.

7 Say what you checked.

When you pass it on, add one line: Checked: [what], against [source], on [date]. Not checked: [what].

8 If a mistake gets out, say so.

Correct it, tell the people who got it, and add it to our watch-for list.

Bring it to your team in ten minutes

- Read the eight lines out. Ask which one the team breaks most.
- Change the words that don't fit, and add your own examples.
- Agree who checks what, and the line everyone adds when they pass work on.
- Put a review date on it. Then pin it up.

In the kit

This rule is a one-page PDF and an editable Word file, with blanks for your team's name, owner and review date. It carries no Neosaipien branding, so your team can make it its own.

A team habit, not legal advice. Our organisation's own AI rules come first.

What you will hear, and what to say

“I don't have time.”

Sort first. Most notes for yourself need a Look, not a full trace. The time goes on the work that leaves the team.

“The AI cited its sources.”

A citation is a place to check, not a check. Open three.

“It's only internal.”

Until it's in someone's deck. If it might travel, sort it as if it will.

Checked against

Every fact in this kit, and where it was read. Pages were read on 16 Oct 2026. Vendor features are as of Oct 2026, every Claude, ChatGPT and Gemini line read on the makers' own pages that day: these products change monthly, so check the settings in your own tool.

Read on the source's own page, 16 Oct 2026

Microsoft, 2026 Work Trend Index (5 May 2026)

microsoft.com/en-us/worklab/work-trend-index/agents-human-agency-and-the-opportunity-for-every-organization

University of Melbourne, media release on the global study of trust in AI (2025)

fbe.unimelb.edu.au/newsroom/media-release-global-study-reveals-trust-of-ai-remains-a-critical-challenge-reflecting-tension-between-benefits-and-risks

KPMG, Trust, attitudes and use of AI: a global study 2025

kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html

BetterUp, The hidden cost of AI "workslop" (29 Sep 2025)

betterup.com/blog/hidden-costs-workslop

OpenAI, Deep research (help page)

help.openai.com/en/articles/10500283-deep-research

OpenAI, ChatGPT Work and Codex; release notes; Memory FAQ; plans and pricing

help.openai.com/en/articles/20001275-chatgpt-work-and-codex

Anthropic, Claude Cowork and chat are now one Claude (16 Sep 2026); Use Claude for Excel; Claude in Google Docs, Sheets and Slides (6 Oct 2026); dashboards (8 Oct 2026)

claude.com/resources/articles/cowork-is-now-claude

Anthropic, Use research on Claude; chat search and memory

support.claude.com/en/articles/11088861-using-the-research-feature

Anthropic, Use Claude Cowork safely (help page)

support.claude.com/en/articles/13364135-use-claude-cowork-safely

Google, Use Deep Research in Gemini Apps; Gemini Notebook (16 Jul 2026); Google Workspace with Gemini; tasks in Gemini

support.google.com/gemini/answer/15719111

Model Context Protocol, What is MCP; Anthropic, Connectors, skills, and plugins

modelcontextprotocol.io/docs/getting-started/intro

Bastounis, Circelli and Hansen, preprint, not peer reviewed (6 Oct 2026)

arxiv.org/abs/2610.08144

Claims checked against their primary sources, with the date

The court case and "likely isn't the same as true", "asking it are you sure isn't checking", "check outside the chat", "quotes and links can be invented too", "give it the source"

Checked against the court's Opinion and Order on Sanctions (22 June 2023), the makers' own documentation and four studies, on 4 and 7 Oct 2026

Small mistakes don't stay small; about one in three, with its caveat; the longer leash

Checked against Sinha and others (arXiv 2509.09677, ICLR 2026), Toby Ord (2025) and Anthropic's "Building effective agents", on 3 Oct 2026

The three tells and the two moves

Checked against Sharma and others (arXiv 2310.13548), Cheng and others (arXiv 2505.13995), three 2026 preprints (arXiv 2602.13093, 2610.08840, 2610.06522) and Anthropic's December 2025 advice, on 8 Oct 2026

"When an AI tells you its work passed a check, ask what was checked"

Checked against the preprint above (arXiv 2610.08144) and OpenAI's own pages, on 14 Oct 2026

Left out on purpose: speed figures for agents whose own source marks them out of date, and the claim that deep research takes minutes, which OpenAI's own page doesn't make. Every example is made up and names no real company or person. Where a rule or a law is mentioned, this is not legal advice. Neosapien is not affiliated with Anthropic, OpenAI or Google.

Where to go next

FOR YOUR TEAM

Workshop in a Box: Check What AI Writes

The same routine, in the same words, as a 60-minute session you run for your team, with slides, a handout and practice sheets.

MAKE IT A RULE

The Team AI Policy Kit

A policy, a one-page staff guide, a 30 day plan and a short deck, so checking AI-made work becomes part of your team's written rules.

EVERYTHING

The Team Pack

Every Neosapien kit in one free download, run as a 90-day programme.

FREE

The Field Guide, levels 2 and 4

Free lessons on checking answers and on agents, at neosapien.com/learn/

I run this live, in the Team workshop

Half a day with your team, on its own real work. For leaders who sign off AI-made work, I also work one to one in Executive coaching. neosapien.com/services/

Book a call

neosapien.com
samson@neosapien.com

The kit on one line

Sort

how much to check

List

every claim that can be checked

Trace

each one at its source, outside the chat

Hedge

keep the words that make it true

Log

what, when, and when to look again

The AI Fact-Check Kit. October 2026 edition. Checked on 16 Oct 2026. Neosapien is AI training and consulting: the Leadership briefing, Executive coaching, the Team workshop, Workflow automation and AI implementation.

Licence. Free from Neosapien under the free licence. Use it, change it for your team and share it free, keeping the credit and the date. A changed version stays inside your own organisation, team, class or group. Please don't sell it. The full terms are in LICENCE.txt, in the download.

Not legal advice. Neosapien is not affiliated with Anthropic, OpenAI or Google.

neosapien