

Evaluation AI

THE M&E AI HANDBOOK • FREE EDITION

AI you can defend to a donor.

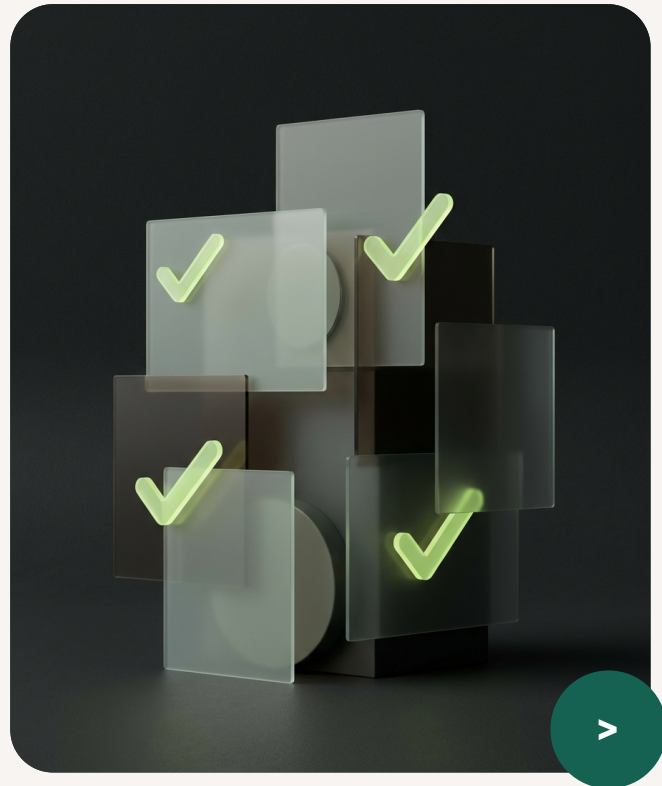
Most monitoring and evaluation professionals now have an AI tool open while they work. This short handbook shows you — in plain English, with nothing assumed — how to use it without ever putting your name on something you cannot prove.

Your free copy — no charge, no sign-up

Written for busy evaluators, not engineers.
Every term is explained.

■ WHAT YOU WILL BE ABLE TO DO BY THE END

Spot the sentence an AI invented... check a whole report with one page... and hand a donor a clear record of where every fact came from. About forty minutes of reading. No software to install.



■ BEFORE YOU BEGIN

Who this is for, and how to use it

This handbook assumes no technical knowledge at all. If you can write a document and you have used a chatbot once or twice, you have everything you need. Read it in order — each short part builds on the one before it.

Who it is for

Evaluators, M&E officers, consultants, project staff and students — anyone whose report carries their name.

How to read it

About forty minutes, start to finish. Or dip into one part and come back later.

What you need

Any AI tool you already use — ChatGPT, Claude, Gemini or Copilot — and a document you have written.

WHAT IS “AI TOOL”?

A program you type a request into — such as ChatGPT — which then writes text back for you. It is also called a chatbot or a language model.

WHAT IS “PROMPT”?

Simply the instruction or question you type into the AI tool. If you have ever asked a chatbot a question, you have already written a prompt.

■ IN SUMMARY

There are no prerequisites. Read at your own pace, and keep a real report beside you to practise on.

What is a “harness”, and why do we need one?

Strictly speaking, a harness is equipment that holds something steady while you work on it. In this handbook, a harness means the same thing for your writing: a small set of routines that holds your AI steady, so nothing unsupported slips through. It is made of four ordinary documents, in plain English.

1 • The Skill

The rules your AI must follow every time it helps you.

2 • The Golden Set

A few practice tasks with known answers, to test the AI.

3 • The Checklist

One page of questions to run over any output.

4 • The Proof Table

A simple record of where each fact came from.

■ GO FURTHER — THE PRO EDITION

This free handbook teaches all four files. The Pro edition gives you fourteen ready-to-use rules, six checklists and three complete procedures — copy, paste and go.

\$79 · downloadable PDF with every template. Or get it included with the video course. For teams, write to contact@evaluationai.com.

Why AI gets things wrong

This is the single most important idea in the handbook, so we will take it slowly. When you ask an AI to write something, it is not looking the answer up in a filing cabinet. It is guessing what words are most likely to come next. Most of the time that guess is remarkably good. Occasionally it is completely wrong — and it never sounds unsure unless you teach it to.

WHAT IS “HALLUCINATION”?

When an AI produces something that sounds true but is not — a number that does not exist, a report that was never written, or a quote nobody said. It is not lying on purpose; it is simply producing the most likely-sounding answer. The word comes from the idea that the machine “sees” something that is not there.

WHAT IS “EVIDENCE”?

A fact that you can point to in a real document — a figure on a specific page, a quotation from a named person. Evidence can be checked by someone else.

WHAT IS “INFERENCE”?

A reasoned judgement that an AI (or a person) makes. It may well be right — but it is not in the document, so it cannot be verified. Inferences are useful; they simply must be labelled honestly.

Confidence is not evidence. A sentence can be beautifully written and still be unsupported.

■ IN SUMMARY

AI predicts likely text rather than checking facts. So our whole method comes down to one habit: separate what is in the source from what was inferred, and label each one.

The seven mistakes that cost reputations

These come up again and again. You do not need to memorise them — the checklist later will do the work — but recognising them is half the battle.

The mistake	What it looks like, and why it is serious
1 · An invented indicator	A target or a baseline the AI made up. It enters your results framework and is repeated for years.
2 · A fake source	A reference to a report that does not exist. The moment a donor checks it, trust is gone.
3 · False certainty	It says the programme caused a change, when the study design cannot show that.
4 · The missing finding	A tidy summary that quietly drops the bad news — the most dangerous of all.
5 · Fake agreement	“Everyone agreed”, when in fact two people raised concerns.
6 · Number drift	You wrote 47%; by the next paragraph it has become “nearly half”, then “most”.
7 · Opinion as fact	The AI's own reasoning, presented as if it were one of your findings.

■ KEY TAKEAWAYS

- None of these are about AI being foolish — they are about AI being fluent.
- The cure is not a cleverer model; it is a habit of checking.
- Every one of the seven is caught by the same routine, taught next.

The one habit that catches most mistakes

If you remember nothing else, remember this. Every sentence an AI writes contains claims. Some are evidence — they came from the document you gave it. Others are inference — the AI's own reasoning. Make the AI label each one as it writes. You are not judging it yourself; you are asking it to show its working.

■ TRY SAYING EXACTLY THIS

“Rewrite the paragraph below. After every sentence, add a tag: [EVIDENCE] if the fact is in the source you were given, or [INFERENCE] if it is your own reasoning. Never invent a number or a source. If something is not in the source, write ‘not in source’.”

1 Paste the source first

Give the AI the document you want it to rely on — for example, the programme report or the survey table.

Why this matters: Without a source, everything it writes is inference.

2 Ask for the labels

Use the wording in the box above, so the AI marks evidence and inference as it goes.

Why this matters: It turns a hidden judgement into something you can see.

3 Scan the [INFERENCE] lines

These need your professional judgement before they can become findings.

Why this matters: This is where over-claiming hides.

■ IN SUMMARY

That single instruction catches five of the seven mistakes at once. Try it on one paragraph you wrote recently — the unsupported sentences stand out immediately.

■ PART THREE

The harness, file by file

Here is each of the four files explained in one plain paragraph. None of them is software. You could write all four in a single afternoon.

■ FILE ONE • THE SKILL

A page of rules — in English — that you paste into your AI so that it behaves the same way every time. Think of it as the instruction sheet you would give a new colleague on their first day.

■ FILE TWO • THE GOLDEN SET

A short set of practice tasks where you already know the correct answer. It is how you test whether your AI is still reliable — much as you would check a new member of staff before trusting them with a donor report.

■ FILE THREE • THE CHECKLIST

One page, six questions. You run it over any output before it leaves your desk. It takes about four minutes and it catches most errors.

■ FILE FOUR • THE PROOF TABLE

A simple list: each claim, its source, and whether it is evidence or inference. This is the page you give a donor when they ask “how do you know?”.

The Skill: telling your AI how to behave

A Skill is nothing more than a short list of instructions. Once it is saved in your AI tool, it applies to everything that tool does for you. Here are the four rules we recommend, each explained in ordinary words.

WHAT IS “LABEL EVERY CLAIM”?

After each sentence, say whether the information is evidence (from the source) or inference (your own reasoning). This is the heart of the method.

WHAT IS “NEVER INVENT”?

Do not create a number, a quotation, a date or a source that is not in the material provided. If it is missing, say “not in source”.

WHAT IS “MINIMUM CLAIM”?

State only what the evidence actually supports. If the evidence is weak, write that it is weak.

WHAT IS “PROVE IT”?

For every important claim, give the source and a confidence level — high, medium or low.

Once these rules are in place, a surprising amount of the risk simply disappears.

■ A QUESTION PEOPLE ASK

“Won't all this slow me down?” At first, yes — by a few minutes a page. It quickly becomes second nature, and it is far faster than repairing a report after a donor has questioned it.

■ IN SUMMARY

A Skill is your standing instructions. Write it once, paste it in, and every answer your AI gives will be shaped by it.

The Golden Set: testing your AI

How do you know whether your AI is still doing a good job today? You test it — with a small set of tasks where you already know the right answer. We call it a Golden Set. It is the closest thing this method has to quality control.

1

Choose five to ten past tasks

Pick tasks you have already completed and marked, with a short, checkable right answer.

Why this matters: You cannot test something you cannot grade.

2

Remove any private details

Replace names with “person A”, and locations with “the region”, so the file can be stored safely.

Why this matters: Client confidentiality still applies, even to test files.

3

Write down the correct answer

One line describing what a good answer must contain, and the mistakes to watch for.

Why this matters: This becomes your mark scheme.

4

Run the test whenever something changes

Every few months, or after the AI tool is updated, run the same tasks again and compare.

Why this matters: AI tools change silently; a good result in January can be a bad one in June.

■ IN SUMMARY

The Golden Set turns a vague feeling — “the AI seems fine” — into something you can actually measure and show.

The Checklist: one page, six questions

Before anything leaves your desk, run these six questions. Score each one — zero or one. Anything below six needs another look. It takes about four minutes.

- Does every number appear in the source, exactly?
- Does every citation exist, and say what it claims?
- Is each claim labelled evidence or inference?
- Is any inference written as though it were a finding?
- Has anything important been left out?
- Is there a proof table for the main claims?

WHAT IS “WHY WE SCORE IT”?

A score gives you something objective to point to, and a simple rule to follow when you are tired or rushed — which is exactly when errors slip through.

Below six out of six, the work does not leave your desk. That single sentence has saved more reputations than any software.

The Proof Table: answering “how do you know?”

This is the file donors value most, and it is the simplest. For every important claim in your report, you keep one row: the claim, where it came from, and whether it is evidence or inference. It turns your report into an auditable record.

Claim	Source	Type
Enrolment rose 12%	MIS report, Q3, page 3	Evidence
Training drove the increase	—	Inference
Attendance 41%	MIS report, Q3, page 5	Evidence

WHAT IS “AUDIT TRAIL”?

A record that lets someone else follow the same path you did, from a claim back to its source. It is what separates a professional report from a persuasive one.

It is the difference between “trust me” and “here is exactly how I know”.

■ **IN SUMMARY**

Keep one proof table per report. It costs a few minutes and it is the strongest evidence of care you can show a funder.

A full example, from start to finish

Let us do the whole thing on one paragraph, so you can see how quickly it works.

The AI wrote: “The programme was highly effective, increasing income by roughly thirty per cent. Stakeholders agreed to scale up. The evidence confirms the model works.”

Claim	Verdict
“Highly effective”	Inference — needs review
“Income up roughly 30%”	Wrong — the source says 12%
“Stakeholders agreed”	Misleading — 8 of 10; two objected
“Evidence confirms the model”	Wrong — the design cannot show cause

■ THE CORRECTED PARAGRAPH

Participants' reported income rose 12% over three months (40 people; MIS report, page 3). Eight of the ten stakeholders supported scale-up; two raised cost concerns. The study design cannot prove that the training caused the change.

■ IN SUMMARY

Four minutes, four errors caught. None of them required technical skill — only the habit of labelling claims and checking them.

Keeping client information safe

Evaluators work under promises of confidentiality. That does not mean you cannot use AI — it means you must be careful about what you paste in, and honest about what you did.

WHAT IS “REDACTION”?

Removing or replacing private details before text is shared — names become “participant”, places become “the region”, and personal details are left out.

WHAT IS “YOUR OWN ACCOUNT”?

Using an AI tool on your own paid plan, with a setting that prevents your text being used to train the company's models.

Before you paste

Remove names. Remove locations.
Combine anything personal so no individual can be recognised.

When you send

Add one honest line: “AI tools were used to assist with drafting; all findings were verified against source documents.”

Honesty about your method is a strength. It tells the donor you understand exactly what you are doing.

Questions people ask

Do I need to be good with technology?

No. If you can write a document and type a question into a chatbot, you can do everything in this handbook.

Does this replace my professional judgement?

No — it protects it. The harness makes the AI's guesses visible so that your judgement is applied where it matters most.

Which AI tool should I use?

Any of the well-known ones — ChatGPT, Claude, Gemini or Copilot. The method works the same way with all of them.

How long does it take to set up?

An afternoon for the four files. After that, it is a few minutes per report — the small cost that keeps your work credible.

■ KEY TAKEAWAYS

- No technical skill, and no new software, is required.
- The method protects your judgement rather than replacing it.
- Setup takes one afternoon; using it takes minutes.

■ QUICK START

Your first thirty minutes

If you would rather start doing than reading, follow these five steps with a report you have already written. It will take about half an hour, and by the end you will have a working habit.

1

Pick one short report

Choose a finished report of one or two pages that you know well.

Why this matters: You can only spot an error if you know what the truth looks like.

2

Paste in the four rules

Copy the Skill rules from the next pages into your AI tool.

Why this matters: They shape every answer from that point on.

3

Ask for the claims

Have the AI list each sentence as evidence or inference.

Why this matters: This exposes anything it invented.

4

Check with the six questions

Run the checklist and score it out of six.

Why this matters: The score tells you plainly whether it is safe to send.

5

Keep a proof table

Write down each claim and its source in three columns.

Why this matters: This becomes your answer to “how do you know?”.

■ IN SUMMARY

Half an hour, one real report, and you have the whole method working. Everything that follows simply explains the details.

■ WORKED EXAMPLE

Checking a logframe

A logframe is a one-page grid showing how a programme's activities lead to results. The AI is good at drafting them — and just as good at inventing numbers. Here is a typical repair, line by line.

An AI-drafted line	What the check found, and the fix
"80% of trainees find work within 3 months."	The 80% was not in the source. The MIS report shows 68%. Corrected.
"Baseline: 0 employed."	No baseline exists. Marked "not available — needs collection".
"Training caused higher incomes."	The design has no comparison group, so "caused" was changed to "was followed by".
"Assumption: the market remains stable."	Fine as an assumption — kept, because it is a genuine risk.

WHAT IS "LOGFRAME"?

Short for "logical framework": a grid of a programme's goal, outcomes, outputs and activities, with an indicator for each level. Donors usually require one.

A single invented baseline can follow a programme for years. The check finds it in minutes.

Five mistakes beginners make

As people start using AI in evaluation, the same five errors show up. Knowing them in advance will save you time.

1 • Trusting the tone

A confident, well-written paragraph feels reliable. Read it for sources first, fluency second.

2 • Checking only at the end

It is far easier to check one section at a time than to unpick a whole report under deadline.

3 • Letting the AI summarise the source

Ask it to work from the document itself. Summaries drift, and drift becomes an error.

4 • Skipping the omission check

The most dangerous mistake is not a wrong sentence — it is the missing one. Always ask what was left out.

5 • Pasting confidential material

Remove names and places first. Confidentiality applies to everything you share.

■ KEY TAKEAWAYS

- Check as you go, not at the end.
- Work from the source, never from a summary.
- Ask what is missing, every time.

Every term, in plain English

More on the next page. Keep this open while you work.

AI tool	A program you type a request into — such as ChatGPT — that writes text back for you.
Prompt	The instruction or question you type into the AI tool.
Model	The engine behind an AI tool. It is updated from time to time, which can change its behaviour.
Hallucination	When AI produces something that sounds true but is made up.
Evidence	A fact you can point to in a real document, with a location.
Inference	A reasoned judgement that is not written down anywhere.
Source	The document, table or report a fact came from.
Claim	Any single statement in a piece of writing.
Verify	To check that something is true by going back to its source.
Citation	A reference showing where a fact came from.
Harness	The set of files and habits that keep your AI honest.
Skill	A page of rules you paste into your AI so it behaves consistently.

Every term, in plain English

Golden set	A few practice tasks with known answers, used to test the AI.
Checklist	A one-page set of yes-or-no questions used to review an output.
Proof table	A record of each claim, its source, and whether it is evidence or inference.
Audit trail	A record that lets someone else follow your steps from a claim to its source.
Baseline	The starting value, measured before a programme begins.
Indicator	A measure used to show whether something has changed.
Means of verification	How you will prove an indicator — where the number comes from.
Logframe	A one-page grid linking activities to results, with an indicator at each level.
Theory of change	A short explanation of how and why a programme is expected to work.
Redaction	Removing private details before text is shared.
Anonymised	With all identifying details removed.
Sign-off	Your formal statement that a report has been checked and is fit to send.

■ IN SUMMARY

Two pages of plain-English definitions. Nothing in this handbook assumes you already knew any of them.

The free handbook is the beginning. The Pro edition is the system.

This free edition gives you the mindset, the seven mistakes, the four-file method and the checklist. When you are ready to run it across your team and your documents, the Pro edition gives you everything, ready to use.

Free edition — this handbook

The method, explained in plain English, with a worked example and the checklist.

Pro edition — \$79

Fourteen ready-made rules, six checklists, three step-by-step procedures, the privacy pack and downloadable templates.

Video course — \$299

Build your first harness with us, slowly and clearly, on screen. Early-bird \$199.

Teams — on request

We can set the whole system up for your organisation. Write to contact@evaluationai.com and we will send a short proposal.

Get the Pro edition · evaluationai.com · contact@evaluationai.com