

4 Ways AI Gets It Wrong and How to Catch It



Generative AI can accelerate research, but it can also produce convincing errors. At Data-Driven, we expect errors and build controls to catch them before they can do damage. AI will sometimes blend sources, invent citations, or fill gaps with guesses. How can you catch these issues before they are embedded in your work? We recommend clear verification rules, such as requiring traceable sources, along with systematic independent AI and human quality checks at each stage of the work. Below are four common ways AI can introduce inaccuracies, along with practical safeguards to prevent, detect, and correct them.



Source Blending and Provenance Loss

When working across any body of evidence, AI can combine facts, findings, or effect estimates from different papers and (confidently!) present them as one claim.

Here's how we prevent this:

- Ask the AI for verbatim excerpts before synthesis
- Label every claim with its exact source
- Have a second, independent AI check each claim against its cited source
- Restrict work to approved online sources/uploaded documents rather than the open web where appropriate



Unsupported Inference and Hidden Assumptions

AI can fill in missing information with plausible-sounding assumptions, turning uncertainty and possibility into unsupported conclusions or overstated confidence.

Here's how we prevent this:

- Instruct the model to state when the evidence is insufficient or conflicting, rather than filling in the gaps
- Ask the AI to include an "assumptions" section in all output, listing each assumption and the reasoning step that depends on it
- Have the AI classify each assumption by basis (stated in source, inferred from source, or not in the provided materials)
- Have an independent model review the output for any unstated assumptions and unsupported inferences the first model failed to flag
- Use the above classifications to target human review, verifying unsupported or high-stakes assumptions before work continues
- Require conclusions to be traceable to source evidence such as evidence tables or explicit source-to-conclusion mapping



Fabricated, Inaccurate, and Misused Citations

Models can invent references; cite real articles that do not support the claim; merge details from multiple papers into a single citation; misstate study methods or results; or pair the correct article with an incorrect title, authors, journal, DOI, PMID, or link.

Until AI shows better integrity with citations, our prevention approach:

- Treat AI-generated citations as leads, not evidence
- Confirm that every AI-cited article exists and that its title, authors, journal, DOI, PMID, and link correspond to the same source
- Confirm that each citation supports the specific claim for which it is cited by consulting the original source, including verifying the full text when the information is not checkable using the abstract alone



Analytic, Coding, and Data Cleaning/Processing Errors

AI-generated code may run without errors and still produce the wrong answer. It can run smoothly and still inappropriately recode values, use the wrong denominator, drop or duplicate records, mishandle missing data, or create erroneous matches during merges.

Here's how we prevent this:

- Maintain versioned logs of prompts, models, inputs, code, and outputs
- Review AI-generated code logic, variable definitions, and data handling decisions before running
- Independently recalculate a pre-defined subset of key statistics, percentages, and effect estimates from the analytic dataset, including at least one row and one column when checking tables; compare these to the AI output
- Check record counts, unique identifiers, missing values, and merge results before and after key data-processing steps

Good quality assurance, in AI as elsewhere, is preventive rather than corrective: anticipate where errors are likely to arise, catch them early, and keep them from propagating downstream. Underlying all of these safeguards is one consistent requirement: human judgment. **No check works unless the person running it understands the question being asked, the evidence available to answer it, and what a sound answer should look like.** Whether you are just beginning to use AI in research or already incorporate it routinely, these practices can help make it more transparent, reliable, and reproducible.

Use AI faster and more responsibly with these safeguards. The Data-Driven team is available to build an approach that fits your data and goals. Our objective is to build processes that make it easier to identify, evaluate, and address what AI gets wrong.