Jeffrey I. Ehrlich
THE EHRLICH LAW FIRM, APC

How AI introduces errors into your documents

AND WHY CATCHING THEM DEMANDS A NEW SKILL

Most of us know about *Mata v. Avianca Inc.* (S.D.N.Y. 2023) 678 F.Supp.3d 443, the first published opinion to sanction a lawyer for filing AI-generated "hallucinated" cases. When *Mata* came out, it attracted a massive amount of media attention, which led me to assume that it would function as a cautionary tale, teaching lawyers who used AI not to trust citations they had not verified. I was naïve.

In the two and a half years since *Mata*, courts across the country have sanctioned dozens of lawyers for the same failures as the lawyers in *Mata*. California joined the list in 2025, when the Court of Appeal published *Noland v. Land of the Free, L.P.* (2025) 114 Cal.App.5th 426, specifically to warn the bar to do better. *In re Domestic Partnership of Campos and Munoz* (2026) __ Cal.App.5th __ (*Campos*) followed in March 2026. In *Campos*, the sanctioned lawyer didn't just file a brief with fabricated citations – when opposing counsel moved to dismiss based on her reliance on fabricated cases, she filed an opposition accusing him of incompetence for being unable to locate the cases her AI had invented.

It is tempting to read cases like *Campos* and conclude that this problem belongs to a different category of lawyer – careless, unprofessional, and not like "us." That temptation is exactly what I want to caution against, because it is both wrong and dangerous.

The reason so many lawyers have been caught by this problem is not primarily one of character or diligence. It is that AI introduces errors into litigation documents in ways that are specifically and structurally difficult to detect using the proofreading skills we have spent our careers developing. Catching these errors consistently requires understanding the variety of ways AI can mislead you, and developing a different kind of proofreading skill than most of us have ever needed or used.

In the first part of this article, I will explain some of the common "failure modes" for generative AI when used for litigation. In the second part, I will explain what the new skill actually looks like and why developing it is harder than it sounds.

AI failure modes: How AI can introduce errors into your document

Hallucinations come in four flavors

The failure mode that has generated the most judicial attention is the hallucinated citation, but not all hallucinations are created equal. The most obvious variety is the completely fabricated case – a citation to an opinion that simply does not exist. This is the *Mata* problem, and while it is the easiest to catch with verification tools, it is apparently not easy enough, given how many lawyers have been sanctioned for it.

The second variety is subtler: A real case cited for a proposition it does not support. The case exists, it survives a KeyCite search, and a quick skim may confirm it involves a related area of law – but it does not actually stand for what the brief says it stands for. Catching this requires reading the opinion carefully enough to evaluate the holding, not just confirm the case's existence.

The third variety is the hardest to detect, and the one I find most sobering: A fabricated but plausible quotation from a real case. The case exists, it addresses the relevant issue, and the "quotation" accurately captures the general thrust of the holding – but the specific language in quotation marks was generated by the model, not written by the court. It is a paraphrase dressed as a direct quote. Because the substance is right and the style matches the court's actual prose, this variety can survive even careful reading of the surrounding opinion. Only a direct comparison of the quoted language against the actual text of the decision will catch it.

The fourth variety receives less attention but may be the most pervasive: factual drift. As a document moves through multiple drafts and sessions, the model can quietly alter specific details that have nothing to do with legal citations – which party is the plaintiff and which is the defendant, a witness's occupation, a contract's dollar amount, a date, a procedural posture. These changes are not flagged. They appear in the revised document with the same surface appearance as everything else, and they will not be caught by any proofreading process that does not specifically compare the new draft against the prior one. The model is not being careless; it is doing what generative models do. Every output is a generation, not a transcription, and generations drift.

Buy-in bias and sycophancy

AI seems objective; after all, it has no emotions or "feelings." But I have learned that it is nevertheless subject to a variety of cognitive biases similar to those that afflict people. One of these is what I call "buy-in bias." It is a form of confirmation bias that works this way: Once an AI model has helped you develop and refine an argument over the course of a drafting session, it becomes a co-author of that argument in a meaningful sense, and like any co-author, it becomes invested in what it helped create.

Ask it to critique a document that it just helped you write and it will find issues at the margins while instinctively defending the core positions it generated. This is not deception; it is a form of intellectual bias. But the practical consequence is that the model is least useful as a critic precisely when you most want a fresh perspective – after a long drafting session when you need someone to tell you what's wrong, or what could be improved.

The mirror image of this problem is worth naming separately, and is a form of sycophancy. Because models can be deferential, a lawyer who pushes back

confidently on the model's analysis may find that the model abandons a correct position under pressure. The lawyer believes that the model is providing a rigorous second opinion, but in reality, it is simply getting back a reflected view, dressed in the model's authority.

Prompt non-compliance – The model doesn't always do what you ask

One of the most important assumptions lawyers make about AI tools – and one of the most dangerous – is that the model is actually doing what it was asked to do. In practice, models regularly fail to follow prompts in ways that are not flagged and may not be obvious. Ask a model to review a new revised draft and it may respond based on its memory of an earlier version rather than the document you uploaded, presenting analysis of the old draft as if it were analysis of the new one. Ask it to evaluate your brief without reference to prior project discussions and it may be unable to honor that instruction, with earlier context bleeding through despite your explicit direction. And it won't tell you that is what it is doing.

I encountered a particularly stubborn version of this problem on a case where a fabricated assertion – that the Supreme Court had granted review in a key case – had been introduced early in a project through an error made by another AI tool. I discovered the error a few days later and corrected it, telling the model that the case was good law and review was denied, not granted. But the model had somehow internalized the wrong information and could not let go of it. Across multiple subsequent sessions, despite my explicit corrections, it kept resurrecting the issue, telling me that I had missed the biggest issue in the motion – that the Supreme Court had granted review and the case was therefore no longer binding. I ultimately had to open a completely fresh chat session with no access to project history to get analysis that was not contaminated by the original error.

A key point to remember: An instruction to disregard earlier context is one that models cannot reliably follow, no matter how clearly it is given. So, if you want to assure a "clean" review of a document, untainted by earlier discussions, it is not enough to ask for it. You will have to start a new chat, either in an "incognito" mode, where prior context is not available, or in a new project that is walled off from all previous projects. But starting a new chat within an existing project will not work; the model will look at the other material in the project.

Related to prompt non-compliance is a problem that is easy to overlook: AI models do not perform consistently. Not across sessions, not across tasks, and not even within a single session on a single document. On any given day, prompted in the same way, a model might acknowledge that it cannot verify a factual claim and suggest you look it up – or it might look it up and verify it before responding – or it might simply assert it confidently without verification and without flagging any uncertainty. Which of these happens is not always predictable, and the model will not tell you which mode it is operating in. A workflow that seemed reliable last week may produce different results today for reasons that are not transparent.

Context-window degradation

As a chat session grows longer and more complex, a model's ability to hold all the relevant information in working memory and integrate it reliably begins to degrade. Output quality can quietly deteriorate over the course of a long drafting session in ways that are not obvious unless you are specifically looking for them. The practical prescription is simple – fresh sessions produce more reliable output than the tail end of lengthy ones. But this runs against the intuition that more context should produce better results. And there can be an understandable reluctance to start a new chat seemingly in mid-stream of the work you are doing on a document, for fear that it will be like "starting over from scratch."

The risk of hyperfocused research

The failure modes above all concern errors that appear in documents after the drafting process begins. A separate and significant problem is the effect of AI-assisted research on what never makes it into the document in the first place if the lawyer is using the AI to conduct targeted research. By that, I mean something different from using AI to surface a key case and then following the research where it leads. Instead, the lawyer asks the AI to find a case that "stands for proposition X." The AI obliges, and the research ends there. As a result, the only research the lawyer does is to confirm that the case stands for that proposition. What can be easily missed with this approach is contrary authority that should be distinguished, useful facts or a helpful quote contained in an opinion the lawyer never read in full, or a line of authority that would have emerged from traditional research but never surfaced from a tightly targeted AI query.

The fix: Can the right tool solve the problem?

Before turning to what lawyers must do differently, it is worth addressing a question many practitioners are already asking: If general-purpose AI tools like ChatGPT or Claude are prone to hallucinations, does switching to a legal-specific AI product solve the problem?

Westlaw and Lexis have both released AI research tools that make a specific and appealing promise: Because the model only has access to their curated database of actual legal authority, it cannot fabricate cases that do not exist. This is a form of what technologists call Retrieval-Augmented Generation, or RAG – the model retrieves real documents from a controlled library before generating its response, rather than generating from training data alone.

The pitch is partially valid. A RAG-based legal-research tool does meaningfully reduce the risk of the first hallucination variety – the completely fabricated case. If the model can only cite cases that exist in the database, it cannot invent a citation out of whole cloth. That is a genuine improvement over general-purpose AI tools used for legal research.

But RAG addresses only one of the four hallucination varieties described above, and the easiest one to catch. The model can still mischaracterize what a retrieved case actually says – citing a real case for a proposition it does not support. It can still generate fabricated quoted language from a real opinion it has access to, presenting a plausible paraphrase as a verbatim extract. It remains fully subject to factual drift, buy-in bias, prompt non-compliance, context-window degradation, and every other failure mode

described above. The database constrains which cases the model can cite. It does not constrain the accuracy of what the model says about them.

Lawyers who have switched to Westlaw or Lexis AI tools in the belief that doing so eliminates the verification burden should understand that it reduces one specific variety of hallucination risk. The verification obligations described in the rest of this section apply with equal force, regardless of which tool generated the draft.

Recognition-based vs. verification-based proofreading

Every lawyer has had the experience of filing something they proofread carefully – read it, reread it – and then discovered after it was filed that it contained a typo, a missing word, or a misspelling they somehow never saw. This is not a failure of diligence. It is a feature of how skilled reading works. Our brains know what a document is supposed to say, and they supply it. The same predictive capacity that allows experienced lawyers to read faster and more fluently than law students – because we can anticipate what comes next – also fills in missing words, smooths over awkward phrasing, and silently corrects small errors before they register. We are not reading what is there; we are reading what we expect to be there because it is what we meant to say.

This is why the standard workarounds help: Read the pages out of order, read the brief backwards, read it aloud. These are all attempts to defeat the brain's prediction machinery by disrupting its ability to know what comes next. They work reasonably well, for spotting typos.

But they do not work for AI hallucinations – and understanding why the key is to understanding what verification actually requires.

The problem with an AI-generated error is not that it looks wrong. It looks exactly right. A hallucinated quotation from a real case is written in that court's register, uses that court's vocabulary, and captures something close to what the court actually held. Your expertise – the very thing that makes you a skilled proofreader – becomes a liability, because expertise fills gaps automatically. You read, "the court held that substantial performance requires..." and your trained brain recognizes the legal framework and completes the thought with what should be there, which may be precisely what the model invented. Reading more carefully does not help. The more expert you are, the more fluently your brain will accept plausible legal language as correct.

This is the distinction I have come to think of as the difference between recognition-based proofreading and verification-based proofreading.

Recognition-based proofreading – what lawyers have always done – works by running the document against known error signatures. Things that look wrong get flagged. It works well because, in traditional practice, you have what I think of as provenance knowledge: You know roughly where the substance came from, because you found it yourself. Your proofreading is checking execution, not origin. You are confirming that you correctly transcribed and applied research you conducted.

AI-generated content changes the provenance of everything in the document without changing how it looks. There is no visual signal that a given quotation was generated rather than retrieved. The only way to catch the error is to leave the document entirely and go to the primary source. Reading the document more carefully won't help, nor will reading it backwards, or aloud. Rather, you have to stop reading the document and go check what it says against the original. And you have to do this for every citation.

That is verification-based proofreading, and it is a different cognitive task than what our training and pre-AI experience prepared us for. It is closer to what a magazine fact-checker does than to what lawyers mean when they say that they proofread the brief. For every proposition, holding, and quotation the document attributes to a source, the question is not "Does this look right?" but "Can I independently confirm this is right?" – and that question can only be answered by pulling the source.

Verification is not new – But AI has raised the bar

At this point, a careful reader might object: Verification is not a new skill. Courts have always expected lawyers to read the cases they cite. Careful practitioners – and their staff – have always verified that citations support the propositions they are offered for. It's called "cite checking," and careful appellate lawyers do it for both factual and legal citations. That is true, and the objection is well-taken.

What AI has changed is not the obligation but the scope of what satisfying it requires. Traditional cite-checking confirmed that a case exists, is good law, and generally supports the proposition for which it is cited. That was, until recently, sufficient – because the lawyer or a trusted researcher found the case, read it, and extracted the relevant language. The provenance of the cited material was known. The risk being guarded against was transcription error or inadvertent mischaracterization, not fabrication.

AI introduces a layer of verification that traditional cite-checking was not designed to address, because it was not necessary. When lawyers found a case, read it, and extracted the quoted language themselves, the act of extraction was the verification – the quote was accurate because the lawyer was the one who put it there. AI removes that link. The quoted language appears in the document without the lawyer having been the one to retrieve it, which means the verification that used to happen automatically must now happen deliberately: a direct word-for-word comparison of the quoted text against the actual opinion. That is more granular and more time-consuming than what cite-checking has traditionally required – and it is the step most likely to be skipped by lawyers who believe they are already doing what the courts expect.

A practical workflow

This is my own current workflow, which I developed through some uncomfortable near-misses: I go through the brief citation by citation, page by page, and verify that the thing the document says is the thing the case actually says. I pull each case in Westlaw, which automatically runs KeyCite – so good law verification happens as a natural byproduct of the same step rather than requiring a separate pass. For quoted language I com-

pare the text in the brief directly against the text of the opinion. This is not fast. It is not my favorite part of the process. But it is absolutely necessary for me, or any lawyer who chooses to use AI in preparing documents that will be filed in court.

I want to be honest about something else. I once came close to not catching a fabricated quotation. It appeared as the third time the case in question was cited in the brief. The citation was accurate. The brief then accurately discussed the case and its holding the second time it was cited. And on the next page, it pointed back to that discussion, stating, "as the *Jones* court explained, and then it had a pithy quotation that summarized what had previously been said, which I had verified. It would have been an entirely accurate paraphrase of what the opinion held. But it was not presented as a paraphrase; it was in quotation marks, and it was entirely fabricated. And I did not spot it.

The accuracy of what had come before in those two prior citations had done its work on me. I had already processed that case as verified, and my attention moved on. Happily, just before I passed the brief on to my (amazing!) paralegal for filing, I decided to check that quote, and I discovered that it was hallucinated. The fix was easy; just delete the quotation marks and present it as a summary.

But that near-miss – and the knowledge of how close I had come to potentially joining the list of lawyers who get called out for not catching hallucinated cites – stuck with me. It inspired me to change my workflow and to formalize my verification-based proofing step.

A few additional practices worth building into your workflow:

On QuickCheck and similar tools

Westlaw's QuickCheck compares quoted language in your brief against the text of cited opinions and flags mismatches. It is genuinely useful, particularly for lawyers who are still developing a verification habit, and as a backstop for quoted language that may have slipped past manual review. Its limitation is important to understand: It only analyzes language that appears in quotation marks. It will not catch a mischaracterized holding, a fabricated standard, or a citation to a case that does not support the proposition for which it is cited.

Not every Westlaw subscription comes with QuickCheck. Mine has had it for years, but I never used it until recently. If you have access to it, by all means use it! But be aware of what it does and does not do.

On fresh sessions

As described in the context-contamination section, models seem unable to reliably follow an instruction to disregard prior context. If you want genuinely "clean" analysis of a document – uncontaminated by the drafting history, the arguments you've already made, the errors that may have been introduced earlier in the project – you need a fresh session, not a fresh instruction. Open an incognito window (in Claude) or its equivalent, or start a new project, or use a separate model entirely. This can be inconvenient. It is also the only way to get what you think you are asking for when you ask for a cold read.

On comparing drafts

Because factual drift can silently alter specific details across revisions – party names, dates, dollar amounts, procedural postures – it is not safe to assume that a revised draft preserved everything in the earlier draft that you wanted to preserve. Compare them deliberately. The model is generating each output fresh, and generations drift in ways that are not flagged and may not be obvious. You may want to use Word's "compare" or redline tool to check for unexpected changes. At minimum, be aware of the issue and look for changes as you edit your drafts.

On the model's confidence

The uniform authoritative tone of AI output is not evidence of accuracy. It is a stylistic feature that applies equally to things the model knows with precision and things it has generated plausibly. Develop the habit of treating AI-generated claims the way you would treat a research memo from a capable first-year associate whose work you have not yet learned to calibrate: A useful starting point that requires verification before reliance, and not to be confused with the work product of someone whose track record you know. The best part of using AI – at least with the best models – is that you can engage with the model about its answer. You can ask hard questions. You can follow up. Take advantage of that capability.

Summing up

I will close where I began. The lesson from the cases where lawyers have been sanctioned for failing to catch AI-generated errors is not that the lawyers were unprofessional or careless, and therefore not like you. Some may have been. But if you use AI, it could easily be you – unless you develop the skills necessary to catch those errors.

The courts are right that professional responsibility has not changed. What has changed is what meeting that responsibility requires: Understanding the new ways that AI can introduce errors into your work, developing a structured method to find those errors, and applying the discipline to use that method on every document.

Jeffrey I. Ehrlich is a certified appellate specialist and the principal of The Ehrlich Law Firm, APC *in Claremont. He is the editor-in-chief of this magazine and an Emeritus member of the Board of Governors of the Consumer Attorneys Association of Los Angeles.*