Managing AI Hallucinations

A Practical Guide for Lawyers

Colin S. Levy

2026

This document is for informational purposes only and does not constitute legal advice.

About the Author

Colin S. Levy is a legal technology advocate, writer, and advisor who works at the intersection of law, technology, and business. With experience spanning in-house legal roles, legal technology companies, and legal operations, he brings a practical perspective to how legal teams can adopt and govern emerging technologies responsibly.

Colin writes and speaks extensively on legal innovation, artificial intelligence in legal practice, and the evolving role of legal professionals in a technology-driven landscape. His work focuses on helping legal teams move beyond the hype cycle to make sound, informed decisions about the tools they use and the workflows they build.

He is the author of The Legal Tech Ecosystem and editor of the Handbook of Legal Tech, and a regular contributor to publications covering legal technology and operations. He advises organizations on responsible AI adoption, legal operations strategy, and the practical governance frameworks that make innovation sustainable.

This guide is part of a series on AI and law that includes AI for Lawyers, AI for Legal Teams, AI Agents Data Handling and Cybersecurity Guide, AI in the Courtroom, Contracting with AI Vendors, Human Judgment and AI in Legal Practice, and the AI Implementation Playbook for Legal Teams.

Table of Contents

Part One

What Hallucinations Are and Why They Matter

Defining the Problem

An AI hallucination is an output that is factually wrong, fabricated, or unsupported by any source material, yet presented as though it were accurate. The model does not flag it. It does not hedge. It delivers a false statement with the same fluency and apparent authority as a true one.

For lawyers, the most dangerous hallucinations involve legal citations. A large language model asked to find case support for a legal proposition will sometimes invent a case: a plausible party name, a realistic reporter citation, a holding that tracks the requested principle. The case does not exist. The reporter volume may not even exist. But the output reads as though it were pulled directly from Westlaw.

A Stanford HAI study tested the leading legal AI research tools and found that Lexis+ AI hallucinated on 17% of queries, Westlaw AI-Assisted Research hallucinated on 33%, and GPT-4 hallucinated on 43%.[1] An earlier companion study by the same researchers tested general-purpose models more broadly and found hallucination rates ranging from 58% (GPT-4) to 88% (Llama 2) on verifiable legal questions. These were not edge cases. The researchers used straightforward legal research queries drawn from federal case law.

The term "hallucination" itself is borrowed from psychiatry, where it describes a sensory experience without an external stimulus. Some researchers argue that "confabulation" is more precise, since what large language models do more closely resembles the neuropsychological phenomenon of filling memory gaps with plausible but incorrect information.[2] The distinction matters less than the practical consequence: AI tools generate false content that looks and reads like real content, and the burden of catching it falls entirely on the lawyer who uses it.

Not a Bug: A Design Consequence

Hallucinations are not software defects that a future update will eliminate. They are a structural consequence of how large language models work. These systems predict the next word in a sequence based on statistical patterns learned from training data. They have no mechanism for verifying claims against external reality. They do not "know" anything in the way a researcher knows something. They generate text that is statistically probable given the preceding context.

This means hallucination rates can be reduced through better training data, retrieval augmentation, and fine-tuning, but they cannot be eliminated entirely. Even the best-performing models on the Vectara Hallucination Leaderboard still hallucinate on standardized summarization tasks, with top models scoring in the low single digits.[3] On open-ended factual questions, the numbers climb sharply. Independent evaluations have found that reasoning-oriented models hallucinate at significantly higher rates on person-specific factual queries than on summarization tasks. Legal research sits squarely in this more challenging category.

The Taxonomy of Legal Hallucinations

Not all hallucinations are alike. Legal AI tools produce distinct categories of fabrication, each carrying different risks and requiring different detection strategies.

- â■¢ **Fabricated citations.** The entire case is invented. The party names, reporter volume, and court are all fiction. This is the most common type and the easiest to catch if you check. It is also the type that has generated the most sanctions.

â■¢ **Hybrid citations.** The model merges elements of two or more real cases into a single nonexistent case. The party name comes from one case, the citation from another, the holding from a third. These "legal Frankensteins" sound credible because every component has a real-world anchor.

â■¢ **Misrepresented holdings.** The case exists and the citation is correct, but the model describes a holding the court never reached, reverses the majority and the dissent, or applies the case to the wrong legal standard. These are the hardest to detect because the citation checks out in your database.

â■¢ **Fabricated quotations.** A direct quote is attributed to a real or fabricated case, but the language does not appear anywhere in the opinion. In *Mata v. Avianca*, the fabricated cases came with fabricated quotations that sounded like real judicial prose.[4]

â■¢ **Phantom statutes and regulations.** The model invents a statutory provision, assigns it a plausible section number within a real title of the U.S. Code, and describes its requirements. The provision does not exist.

â■¢ **Correct law, fabricated authority.** The legal principle the model states is accurate, but the case cited as support is invented. This category is particularly insidious because the underlying analysis is sound, which makes the false citation easy to overlook.

Key Principle:

The hallucination risk is not that AI gets the law completely wrong. The risk is that it gets the law mostly right, then invents the authority that supposedly supports it. The closer the output is to being correct, the harder the fabrication is to catch.

Part Two

Why Hallucinations Happen

Next-Token Prediction and the Confidence Trap

A large language model does not retrieve information from a database. It generates text one token at a time, selecting each word based on the statistical probability that it follows the preceding sequence. The model has no internal fact-checker, no lookup table of verified citations, and no concept of truth. It produces whatever continuation is most probable given its training.

When you ask a model for a case supporting a legal proposition, the model does not search a case database. It generates a sequence of tokens that pattern-matches to what a case citation looks like in context. If the training data contained enough examples of "courts have held" followed by a party name, reporter citation, and holding, the model will produce output that fits that pattern regardless of whether the specific case exists. The statistical engine optimizes for plausibility, not accuracy.

This architecture creates what researchers call the "confidence trap": every output carries the same surface-level confidence. The model does not distinguish between a response grounded in extensive training data and one that fills a gap with a plausible guess. Both read identically. Both use the same authoritative tone. The only way to tell the difference is to verify independently.

Training Data Gaps and Knowledge Boundaries

Large language models are trained on massive text corpora, but those corpora have boundaries. Significant categories of legal knowledge remain partially or wholly absent from training data: unpublished opinions, sealed filings, recent legislative amendments, state administrative decisions, and proprietary legal databases behind paywalls. When a model encounters a query that falls outside or at the edge of its training, it does not say "I don't know." It generates the most probable response given whatever partial patterns it has.

Knowledge cutoffs compound the problem. A model trained on data through a specific date has no access to anything published afterward. It cannot tell you about a case decided last month, but if asked, it will often generate a plausible-sounding response rather than acknowledge its limitation. For rapidly evolving areas of law (AI regulation, cryptocurrency, data privacy), the gap between the training cutoff and the current state of the law can be significant.

The Softmax Bottleneck

At a technical level, transformer-based language models use a mathematical function called softmax to convert raw prediction scores into probability distributions over the vocabulary. This function constrains the model's expressivity: when the correct output requires the model to represent a complex probability distribution with multiple plausible continuations, softmax forces a simplification. The model picks the highest-probability path, which may not be the factually correct one.[5]

For legal text, this matters because case law analysis often involves nuance: a case may support a proposition in one jurisdiction but not another, under certain facts but not others. The model cannot hold that complexity in a single probability distribution and instead collapses it into a single confident-sounding assertion.

Retrieval-Augmented Generation Is Not a Cure

Retrieval-augmented generation (RAG) connects a language model to an external knowledge base. Before generating a response, the system retrieves relevant documents and feeds them to the model as context. Legal AI tools like Lexis+ AI and Westlaw AI-Assisted Research use RAG architectures to ground their outputs in actual case law.

RAG reduces hallucination rates compared to general-purpose models, but it does not eliminate them. The Stanford HAI study tested RAG-based legal tools specifically and still found hallucination rates of 17% to 33%.[1] Retrieval can fail in several ways: the search query may not match the relevant documents, the retrieved documents may be incomplete, or the model may ignore the retrieved context and generate from its own training instead. When any of these failures occur, the RAG system hallucinates just as a standalone model would, but with the added credibility of being marketed as "grounded in real sources."

Bottom Line:

RAG makes hallucinations less frequent but no less dangerous. A tool that hallucinates 17% of the time is not a tool you can trust without independent verification. It is a tool that gives you the wrong answer roughly one out of every six queries.

Part Three

The Damage Done

Mata v. Avianca and Its Aftermath

In June 2023, Judge P. Kevin Castel of the Southern District of New York imposed sanctions in *Mata v. Avianca, Inc.*, the case that forced the legal profession to confront AI hallucinations directly.[4] Plaintiff's counsel had filed an affirmation citing six cases that did not exist. All six were generated by ChatGPT. When opposing counsel flagged the problem, plaintiff's attorneys asked ChatGPT whether the cases were real. ChatGPT confirmed they were. The attorneys then provided the court with fabricated opinion texts, still generated by ChatGPT, as supposed proof.

Judge Castel found the attorneys acted in "subjective bad faith" and imposed a $5,000 fine. He required them to send letters to every judge falsely identified as the author of a fabricated opinion, attaching the court's order and copies of the fake decisions. The underlying personal injury case was separately dismissed on statute of limitations grounds.

Mata was the warning. The profession did not fully heed it.

The Sanctions Accelerate

Since *Mata*, documented sanctioning events have multiplied. A database maintained by researcher Damien Charlotin tracked approximately two AI hallucination incidents per week through early 2025. By late 2025, the rate had accelerated to two to three per day.[6] Several cases illustrate the pattern.

In *Park v. Kim* (Second Circuit, 2024), attorney Jae S. Lee cited a case generated by ChatGPT in a reply brief. When ordered to produce a copy of the decision, she admitted the case did not exist. The court referred her to the Second Circuit Grievance Panel for disciplinary proceedings.[7]

In an ERISA case in the Southern District of Indiana (2025), a federal magistrate recommended $15,000 in sanctions against attorney Rafael Ramirez for filing three briefs containing fabricated citations generated by AI tools. The district judge ultimately imposed $6,000. Ramirez had not verified a single citation.[8]

In *Noland v. Land of the Free, L.P.* (California Court of Appeal, 2025), an attorney used ChatGPT to "enhance" appellate briefs. Nearly all case quotations in the opening brief were fabricated. The court imposed a $10,000 sanction.[9]

In *Gauthier v. Goodyear Tire & Rubber Co.* (E.D. Tex., 2024), an attorney admitted using Claude without verification. Two nonexistent cases appeared in the filing with fabricated quotations. The court imposed a $2,000 penalty and ordered mandatory continuing legal education on generative AI.[10]

These are not isolated incidents. They represent a systemic failure of verification that tracks directly to how lawyers integrate AI into their research workflows.

The Numbers Behind the Problem

The scale of the problem can be quantified from multiple angles. The Stanford HAI study tested legal AI tools on queries with known answers and found that even the best-performing specialized tool (Lexis+ AI) returned hallucinated content on 17% of queries, while GPT-4 hallucinated on 43%. Broader testing by the same team found general-purpose models hallucinating on 58% to 88% of legal research questions.[1]

In practical terms: if a legal AI tool returns ten citations, between one and three may be fabricated, misrepresented, or inaccurate. If you rely on any of those citations in a court filing without independent verification, you face sanctions under Rule 11, potential malpractice liability to your client, and disciplinary action from your state bar.

The consequences are not limited to sanctions. In July 2025, in *MyPillow/Lindell v. Dominion*, a federal judge imposed $3,000 sanctions per attorney for AI-generated briefs with fabricated citations.[11] In an Arizona Social Security appeal in August 2024, a judge found that twelve of nineteen cited cases were fabricated, misleading, or unsupported. The acceleration of these incidents tracks the adoption curve of generative AI in legal practice.

Part Four

Your Professional Obligations

The ethical framework governing AI hallucinations already exists. No new rules are required. The Model Rules of Professional Conduct, as interpreted by the ABA and state bars, impose clear obligations on any lawyer who uses generative AI.

Rule 11: The Certification You Already Make

Every time you sign a pleading, motion, or other paper filed with a federal court, you certify under Federal Rule of Civil Procedure 11 that the legal contentions are "warranted by existing law" and that factual contentions have "evidentiary support." This is not a new obligation. What is new is the risk that generative AI tools will produce contentions that appear warranted but rest on fabricated authority.

Courts apply an objective reasonableness standard when evaluating Rule 11 compliance. The question is whether a reasonable attorney would have discovered the fabrication through "reasonable inquiry." Typing a citation into Westlaw or Lexis takes seconds. Failing to do so after generating citations with an AI tool that is known to hallucinate is difficult to defend as reasonable.[12]

Rule 1.1: Competence Now Includes AI Literacy

Model Rule 1.1 requires lawyers to provide "competent representation" with the "legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation." Comment 8 to Rule 1.1 already requires lawyers to keep abreast of changes in law practice, "including the benefits and risks associated with relevant technology."

ABA Formal Opinion 512, issued July 29, 2024, made the connection explicit: lawyers need not become AI experts, but they must develop a "reasonable understanding of the capabilities and limitations" of any generative AI tool they use in client work.[13] A lawyer who uses ChatGPT for legal research without understanding that it regularly fabricates citations is not meeting this standard.

Rule 3.3: Candor Means Verification

Model Rule 3.3 prohibits a lawyer from knowingly making "a false statement of fact or law to a tribunal." It also requires correction of false statements previously made. The word "knowingly" has generated debate in the AI context: if a lawyer does not know a citation is fabricated, has the lawyer violated Rule 3.3?

The answer increasingly turns on willful blindness. A lawyer who uses a tool known to hallucinate and does not verify its output cannot credibly claim ignorance. Judge Castel's finding of "subjective bad faith" in *Mata* rested partly on the attorneys' failure to exercise any independent verification despite having ready access to legal databases. The obligation is not to guarantee accuracy; it is to conduct the kind of verification that any competent lawyer would conduct before citing a case to a court.

ABA Formal Opinion 512

Formal Opinion 512 addresses six areas of professional responsibility affected by generative AI: competence, confidentiality, communication with clients, candor toward the tribunal, supervision, and fees.[13] On hallucinations specifically, the opinion states that lawyers must "recognize potential for error and ensure AI output is accurate before relying on it in client work or court filings." Critical review and independent verification are required before court submission.

The opinion also addresses supervision. Partners with managerial authority must establish clear policies on generative AI use. When associates, paralegals, or contract attorneys use AI tools, the supervising lawyer bears responsibility for ensuring adequate verification under Model Rules 5.1 and 5.3. Delegation of research to AI without a corresponding verification protocol is a supervision failure.

State Bar Guidance

State bars have moved to supplement the ABA's framework with jurisdiction-specific guidance.

The Florida Bar issued Advisory Opinion 24-1 in January 2024, requiring lawyers to "reasonably guarantee compliance" with ethical obligations when using generative AI. The opinion addresses confidentiality (obtain informed consent before using third-party AI tools with client data), oversight (develop policies to verify AI use is consistent with ethical obligations), fees, and advertising. Chatbot-based client intake must disclose that the prospective client is communicating with AI, not a lawyer.[14]

The California State Bar released practical guidance in November 2023 requiring attorneys to verify AI-generated outputs for accuracy before court submission. In September 2025, the California Court of Appeal (Second District) reinforced this principle, stating that "no brief, pleading, motion, or any other paper filed in any court should contain citations that the attorney has not personally read and verified." California's SB 574 would codify verification requirements into statute.[15]

In the Northern District of Texas, Judge Brantley Starr's standing order requires attorneys to certify either that no generative AI was used in preparing a filing or that any AI-generated language has been "checked for accuracy by a human being." Failure to file the certification results in striking the filing from the docket.[16]

Practical Takeaway:

The ethical rules do not prohibit AI use. They prohibit unverified AI use. Every citation, every factual assertion, and every statutory reference generated by an AI tool must be independently confirmed before it appears in any filing, memo, or communication to a client or court.

Part Five

Detection and Verification

Why Hallucinations Are Hard to Spot

A hallucinated citation does not look wrong. It looks exactly like a real citation. The party names are plausible. The reporter volume and page number fall within realistic ranges. The court and year are consistent with the legal issue. The holding described aligns with the proposition you asked the model to support. Everything about the output signals credibility.

This is by design. The model was trained on millions of real citations and has learned precisely what a citation should look like. It generates text that conforms to the structural patterns of legal writing, including citation format, because doing so maximizes the statistical probability of the next token. The better the model gets at generating realistic-looking text, the harder its fabrications become to spot by reading alone.

Misrepresented holdings pose an even greater challenge. The citation is real. It pulls up in your database. But the model has described a holding the court never reached, or has subtly reframed the facts to make the case appear more favorable than it is. Catching this requires reading the actual opinion, not just confirming the citation exists.

A Verification Protocol That Works

Effective verification requires a structured process. Ad hoc spot-checking is insufficient when hallucination rates range from 17% to 33% on specialized tools. The following protocol addresses each category of hallucination risk.

Step 1: Confirm existence. Enter every case citation into a primary legal database (Westlaw, Lexis, Google Scholar). If the case does not appear, it is fabricated. Do this for every citation, not a sample. A 17% hallucination rate means roughly one in six citations may be false.

Step 2: Verify the holding. For every citation that clears Step 1, read the actual opinion. Confirm that the holding matches what the AI described. Check whether the case was reversed, vacated, or distinguished on the point at issue. Confirm the procedural posture is accurately represented.

Step 3: Check quotations. If the AI output includes a direct quote from a case, search for that exact language in the opinion. Fabricated quotations often sound plausible but cannot be found in the actual text.

Step 4: Validate statutory references. Look up every statute, regulation, and rule cited in the output. Confirm the section number exists, the provision says what the AI claims it says, and the language has not been amended since the model's training cutoff.

Step 5: Check dates and status. Verify case dates, effective dates of statutes, and procedural status (pending, settled, reversed). AI models frequently assign incorrect dates or misstate whether a case is still good law.

WARNING: THE VERIFICATION TRAP

Do not use the same AI tool to verify its own output. In *Mata v. Avianca*, the attorneys asked ChatGPT to confirm whether the cases it generated were real. ChatGPT confirmed they were. An AI tool has no ability to distinguish its own fabrications from its accurate outputs. Verification must use an independent, authoritative source.

The Red-Yellow-Green Framework

A risk-based classification system helps firms allocate verification resources where they matter most.

â■¢ **Red (prohibited without dual verification).** Any AI-generated content destined for a court filing, regulatory submission, or client-facing legal opinion. Requires independent verification of every citation and factual assertion by the attorney of record plus a second reviewer.

â■¢ **Yellow (single verification required).** Internal research memos, first drafts, and preliminary analysis. Requires the attorney to verify all citations and key factual claims before circulating to the team or using as a basis for further work.

â■¢ **Green (standard use).** Administrative tasks, document summarization from provided source texts, and initial issue-spotting. Human review for reasonableness, but citation-level verification not required because the output does not cite external authority.

Benchmarks and Evaluation Tools

The Vectara Hallucination Leaderboard provides the most widely referenced benchmark for comparing hallucination rates across models. It uses the Hughes Hallucination Evaluation Model (HHEM), a specialized model that detects when generated text is not supported by provided source material.[3] The leaderboard tests models on summarization tasks and assigns a hallucination rate score: lower is better.

For law firms evaluating AI tools, the Vectara data provides a starting point but not a complete picture. The leaderboard tests general summarization accuracy, not legal research specifically. The Stanford HAI study remains the most relevant benchmark for legal AI tools. When selecting an AI tool for legal research, ask the vendor for its hallucination rate on legal-specific tasks, not just its general benchmark score. If the vendor cannot provide this data, treat the tool's output with proportionally greater skepticism.

Part Six

Building a Firm-Level Response

Governance Structures

Individual verification is necessary but not sufficient. Firms need governance structures that make responsible AI use the default, not a matter of individual discipline.

Establish an AI governance committee. Include at least one equity partner, a representative from IT or information security, and attorneys from each major practice group. The committee's mandate should cover tool selection and vetting, usage policy development, incident review, and regular policy updates as case law and bar guidance evolve. Meet quarterly at minimum.

Maintain an approved tool list. Not all AI tools carry the same hallucination risk. Legal-specific tools with RAG architectures hallucinate less frequently than general-purpose models. Evaluate each tool's hallucination rate, data privacy protections, and terms of service before approving it for firm use. General-purpose chatbots (consumer versions of ChatGPT, Claude, Gemini) should carry the highest verification requirements or be restricted from use on client matters entirely.

Create an incident log. Track every instance where an AI-generated hallucination is caught during the verification process. This data helps the firm understand which tools and use cases carry the greatest risk and informs future policy decisions. It also demonstrates due diligence if the firm ever needs to defend its AI governance practices.

Training Requirements

ABA Formal Opinion 512 and the emerging consensus across state bars make clear that competence under Rule 1.1 now includes AI literacy.[13] Firms should implement mandatory training covering at least the following areas.

- â■¢ **What hallucinations are.** The training should explain the technical basis for hallucinations (next-token prediction, not retrieval), the taxonomy of hallucination types, and current hallucination rates for the tools the firm uses.
- â■¢ **How to verify.** Walk through the five-step verification protocol. Show real examples of hallucinated citations and demonstrate how they were detected. Use the firm's own tools and databases.
- â■¢ **What the rules require.** Cover Rule 11 certification requirements, Rule 1.1 competence obligations, Rule 3.3 candor duties, and the supervisory obligations under Rules 5.1 and 5.3. Connect each rule to specific AI use scenarios.
- â■¢ **What has gone wrong.** Review the sanctions cases: *Mata*, *Park*, *Noland*, and others. Discuss what the attorneys did wrong and what they should have done differently. These cases are more persuasive than abstract policy discussions.

Client Communication and Disclosure

ABA Formal Opinion 512 requires specific disclosure to clients about AI tool use. This is not optional. The opinion states that "generic consent in engagement letters is insufficient" and that lawyers must provide specific information about which AI tools they use, how data is processed, and what professional responsibility risks they have assessed.[13]

For hallucination risk specifically, client communication should address three points. First, which AI tools the firm uses and for what purposes (research, drafting, document review). Second, the verification protocols the firm applies to AI-generated work product before it reaches the client. Third, the known limitations of the tools, including hallucination rates and the categories of error they produce.

This disclosure protects the firm. If a hallucination survives the verification process and causes harm, documented disclosure and robust verification protocols are the firm's best defense against malpractice claims and disciplinary action.

Insurance and Coverage Gaps

Most professional liability policies do not explicitly address AI-related errors. Whether a hallucination-caused malpractice claim falls within coverage depends on how the policy defines "professional services" and whether the insurer treats AI-assisted work as falling within that definition.[17]

The risk runs in both directions. If a lawyer cannot demonstrate reasonable care in using AI tools, an insurer may argue that no covered "professional service" occurred because the lawyer delegated judgment to a machine without adequate oversight. The gap between what your policy covers and what AI-related liability you actually face may be wider than you expect.

Specialized AI insurance products are emerging. Munich Re has entered the AI liability coverage market, and Armilla provides policies that address hallucinations, model degradation, and algorithmic failures. These products are not yet widely available or standardized, and not all are tailored specifically to legal malpractice. Contact your broker and ask two questions: does your current policy cover malpractice claims arising from reliance on AI-generated output, and if not, what supplemental coverage is available?[17]

Insurance is one piece of the risk allocation puzzle. Your AI vendor contract is the other. Liability caps, indemnification scope, and performance warranties in the vendor agreement determine who bears the cost when a hallucination causes harm. For detailed guidance on negotiating these provisions, including specific clause language for accuracy thresholds, hallucination-related liability carve-outs, and vendor insurance requirements, see the companion guide in this series: *Contracting with AI Vendors: A Practical Guide for Lawyers*.

Part Seven

What Comes Next

Hallucination rates will decline as models improve. They will not reach zero. The architecture of large language models, which generate text through statistical prediction rather than factual retrieval, makes some level of fabrication inherent to the technology. Better training data, improved retrieval systems, and more sophisticated fine-tuning will push rates lower, but the fundamental trade-off between fluency and accuracy will persist.

The legal profession's response must account for this reality. The question is not whether to use AI tools. They are already embedded in legal research platforms, document review systems, and practice management software. The question is how to use them responsibly, which means treating every AI output as a draft that requires human verification before it carries legal consequences.

Courts will continue to sanction lawyers who file unverified AI-generated content. State bars will continue to tighten their guidance. Insurance carriers are already scrutinizing how firms govern AI use when evaluating coverage. The firms that build verification workflows now will be positioned to benefit from AI's genuine advantages without absorbing its most dangerous risks.

The standard is not perfection. It is diligence. Read every case you cite. Verify every statute. Check every quotation. Document your verification process. Disclose your AI use to clients. Train your team. These are not new obligations. They are existing obligations applied to a new tool. The tool is powerful. The obligations have not changed.

Glossary of Key Terms

Confabulation.

An alternative term for AI hallucination drawn from neuropsychology, where it describes the production of fabricated memories to fill gaps in recall. Some researchers prefer this term because large language models do not experience sensory perception and therefore cannot "hallucinate" in the clinical sense. In legal AI discourse, "hallucination" remains the dominant term.

Embedding.

A numerical representation of text that captures semantic meaning as a vector in high-dimensional space. Retrieval-augmented generation systems use embeddings to match queries to relevant documents. When the embedding search retrieves irrelevant or incomplete documents, the model may hallucinate despite having access to a knowledge base.

Fine-Tuning.

The process of further training a pre-existing model on domain-specific data to improve performance for particular tasks. Fine-tuning a general-purpose model on legal text can reduce hallucination rates for legal queries, but it cannot eliminate them entirely.

Foundation Model / Large Language Model (LLM).

A large AI model trained on broad text data that generates output by predicting the most probable next token in a sequence. Foundation models include GPT-4, Claude, Gemini, and Llama. Legal AI tools typically build on top of foundation models with additional retrieval and fine-tuning layers.

Hallucination.

An AI output that is factually incorrect, fabricated, or unsupported by source material, presented with apparent confidence. In legal contexts, the most dangerous hallucinations involve fabricated case citations, misrepresented holdings, invented statutes, and fabricated quotations.

Hughes Hallucination Evaluation Model (HHEM).

A specialized model developed by Vectara to detect hallucinations in AI-generated text by comparing outputs to source documents. The HHEM forms the basis of the Vectara Hallucination Leaderboard, the most widely referenced benchmark for comparing hallucination rates across models.

Inference.

The process of an AI model generating output from input, as distinct from training. When you submit a query to a legal AI tool, inference is the process that produces the response. Hallucinations occur during inference.

Model Drift.

The gradual degradation of a model's performance over time as real-world data diverges from training data. Model drift can increase hallucination rates without any visible change in the tool's interface or behavior.

Next-Token Prediction.

The core mechanism by which large language models generate text. The model evaluates the probability distribution over its vocabulary and selects the most likely next word given the preceding

context. This process has no built-in mechanism for factual verification.

Retrieval-Augmented Generation (RAG).

An architecture that retrieves relevant documents from a knowledge base before generating a response, grounding the model's output in source material. Legal AI tools use RAG to reduce hallucination rates, but the Stanford HAI study found that RAG-based legal tools still hallucinate on 17% to 33% of queries.

Softmax Function.

A mathematical function in transformer models that converts raw prediction scores into a probability distribution. The softmax function constrains the model's ability to represent complex distributions, which can contribute to hallucination when multiple plausible continuations compete.

Token.

The basic unit of text processed by a language model. A token may be a word, part of a word, or a punctuation mark. Legal citation formats involve specific token sequences that the model has learned to reproduce, which is why fabricated citations follow correct formatting conventions.

Endnotes

[1] Magesh et al., "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," 22 J. Empirical Legal Studies 216 (2025). Found hallucination rates of 17% (Lexis+ AI), 33% (Westlaw AI-Assisted Research), and 43% (GPT-4) across verifiable federal case law queries. The 58-88% range for general-purpose models comes from the same research team's earlier companion study, "Large Legal Fictions" (2024), which tested GPT-4 (58%), GPT-3.5 (69%), and Llama 2 (88%).

[2] See, e.g., Smith, Greaves & Panch, "Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models," 2 PLOS Digital Health e0000388 (November 2023); Berk, "Beware of Artificial Intelligence Hallucinations or Should We Call Confabulation?," PMC (2024). Both argue that confabulation more precisely describes the gap-filling behavior of LLMs.

[3] Vectara Hallucination Leaderboard (2025), using HHEM-2.3. Leaderboard rates are updated continuously; top models score in the low single-digit percentages on summarization benchmarks. Reasoning models showed significantly higher rates on open-ended factual queries. See also Visual Capitalist, "Ranked: AI Models with the Lowest Hallucination Rates," 2025.

[4] Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). $5,000 sanction imposed for filing six fabricated case citations generated by ChatGPT. Court found "subjective bad faith" in failure to verify.

[5] See ACM Transactions on Information Systems, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," 43 ACM Trans. Info. Sys. 1-55 (2025). The softmax bottleneck is discussed in the context of how transformer architectures constrain output distributions and contribute to hallucination risk.

[6] Charlotin, "AI Hallucination Cases Database" (ongoing). Tracked approximately two incidents per week through early 2025, accelerating to two to three per day by late 2025. Available at damiencharlotin.com/hallucinations.

[7] Park v. Kim, Second Circuit (2024). Attorney Jae S. Lee cited a nonexistent case generated by ChatGPT. Referred to Second Circuit Grievance Panel for disciplinary proceedings.

[8] ERISA case, S.D. Ind. (2025), U.S. Magistrate Judge Mark J. Dinsmore. Attorney Rafael Ramirez filed three briefs with fabricated AI-generated citations. Magistrate recommended $15,000 in sanctions; district court imposed $6,000.

[9] Noland v. Land of the Free, L.P., California Court of Appeal (2025). Nearly all case quotations in opening brief were fabricated by ChatGPT. $10,000 sanction imposed.

[10] Gauthier v. Goodyear Tire & Rubber Co., E.D. Tex. (2024). Attorney admitted using Claude without verification. $2,000 penalty plus mandatory CLE on generative AI.

[11] MyPillow/Lindell v. Dominion, Colorado Federal Court (July 2025). $3,000 sanctions per attorney for AI-generated briefs with fabricated citations.

[12] Gunder, "Rule 11 Is No Match for Generative AI," 27 Stan. Tech. L. Rev. 308 (Spring 2024), analyzing the objective reasonableness standard as applied to AI-generated court filings.

[13] ABA Standing Committee on Ethics and Professional Responsibility, Formal Opinion 512, "Generative Artificial Intelligence Tools," July 29, 2024. Addresses competence, confidentiality, communication, candor, supervision, and fees.

[14] The Florida Bar Board of Governors, Ethics Opinion 24-1, "Lawyers' Use of Generative Artificial Intelligence" (January 19, 2024). Non-binding advisory guidance on confidentiality, oversight, fees, and advertising.

[15] California State Bar, "Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law" (November 2023). The quoted verification standard is from Noland v. Land of the Free, L.P., Cal. Ct. App., 2d Dist. (September 2025). See also California SB 574, which would codify verification requirements into statute.

[16] Standing Order, Judge Brantley Starr, N.D. Tex. (May 30, 2023). Requires certification regarding generative AI use in filings. First federal standing order specifically addressing AI hallucination risk.

[17] See ABA Journal, "Does Your Professional Liability Insurance Cover AI Mistakes? Don't Be So Sure," February-March 2025. Identifies significant coverage uncertainty for AI-related malpractice claims under standard professional liability policies.