

Theory Meets Practice

The Surprising Value of AI Hallucinations

BY JOHN LANDE

You're hallucinating if you think AI will not hallucinate.

Artificial intelligence hallucinations are inaccurate outputs, such as fabricated cases, invented quotations, misstated procedural rules, or inaccurate summaries of documents, often presented with unwarranted confidence. Because these outputs often appear polished and plausible, they can be especially misleading.



No one should be surprised anymore that AI tools sometimes produce hallucinations. Reports of fake citations, fabricated authorities, manipulated images, and AI-generated misinformation seem to appear almost daily.

Damien Charlotin, a French lawyer and data scientist, maintains a database tracking court decisions involving AI hallucinations in legal filings at www.damiencharlotin.com/hallucinations.

As of this writing, the database includes almost 1,000 U.S. cases—and counting.

These cases understandably alarm lawyers and judges. They should. Lawyers and other dispute resolution professionals can no longer plausibly claim ignorance about the risks of using AI without carefully reviewing its output.

Lawyers who submit fictitious authorities or inaccurate statements can harm clients, damage their reputations, waste judicial resources, violate professional obligations—and incur substantial sanctions.

More broadly, growing public anxiety about AI-generated misinformation extends far beyond legal practice. People increasingly
(continued on next page)

The author, a longtime *Alternatives* contributor, presents this regular column, "Theory Meets Practice." He is Isidor Loeb Professor Emeritus at the University of Missouri School of Law's Center for the Study of Dispute Resolution in Columbia, Mo. In 2024, he received the American Bar Association Section on Dispute Resolution's Award for Outstanding Scholarly Work. See <https://bit.ly/3Tq5YuK>. He developed the Real Practice Systems Negotiation and Mediation Coach (RPS Coach); see <https://bit.ly/3M590VJ>. His biography page can be found at <https://lande.missouri.edu>.

SPECIAL SERIES: AI IN ARBITRATION AND LITIGATION: THE JAGGED FRONTIER

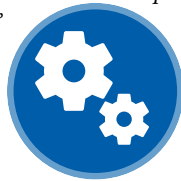
ADR Technology/Part 1 of 6

The Nature of AI in Dispute Resolution: Competence Without Comprehension

BY DOAK BISHOP & KARL HENNESSEE

Series Overview: This is the first of six articles examining artificial intelligence in dispute resolution—a series called "AI in Arbitration and Litigation: The Jagged Frontier." Subsequent articles address specific large language models capabilities lawyers can rely upon; advanced AI architectures for reasoning, causation,

and analogical inference; practical use cases, tools, and prompting techniques; the efficiency paradox—whether AI will make dispute resolution cheaper or more expensive, and the irreplaceable role of human judgment.



* * *

The philosopher Daniel Dennett coined the phrase "competence without comprehension" to describe how evolution produces organisms that behave with apparent purpose, but without understanding what they are doing. Daniel C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (W.W. Norton, 2017). The phrase "competence without comprehension" serves throughout Dennett's book as the central explanatory device for how natural selection produces sophisticated, apparently purposive behavior from entirely non-purposive processes.

For example, a bacterium swims toward nutrients with remarkable efficiency; it
(continued on page 131)

Bishop is a Houston-based independent arbitrator and a retired King & Spalding partner who founded the firm's international arbitration practice 25 years ago. He was chair of the Institute of Transnational Arbitration in 2012-2015; past member of ICCA, LCIA, SIAC, AAA, and ASIL governance boards; past member of the U.S. delegation to the NAFTA 2022 Committee on Private Commercial Disputes, and an adviser to the U.S. Restatement (3rd) of the Law of International Commercial and Investor-State Arbitration. See <https://doakbishop.com>. Hennessee is a Senior Vice-President, Litigation, and Investigations at Airbus in France, where he leads the company's global activities in these areas. He holds or has held board seats at the ICC (Chair), HKIAC, LCIA, ICCA (Treasurer), AFA and the Queen Mary School of International Arbitration. He is a frequent author and occasionally sits as arbitrator.

Lawyers, mediators, and negotiators regularly use brainstorming techniques that generate many ideas, some of which later prove unrealistic or unhelpful. AI can function similarly by rapidly generating possibilities that users can evaluate, refine, reject, or develop further.

Hallucinations also can encourage greater intellectual humility and caution about confident assertions that sound plausible but may be inaccurate. Regular interaction with AI systems may encourage more careful analysis of sources, assumptions, and evidence. That realization can improve professional judgment more generally because people encounter similar problems when dealing with flawed human sources of information.

One might think of AI as a smart junior associate who works quickly, never gets tired, often produces valuable work, occasionally makes mistakes, and always requires supervision. Many lawyers would prefer to have such an associate help research and organize ideas rather than do everything themselves, recognizing that the associate's work may include mistakes. The supervising lawyer remains responsible for reviewing the work carefully before relying on it.

AI requires a similar approach.

Careful AI Use Is Essential

Responsible AI use requires thoughtful habits, realistic expectations, and independent verification of important outputs. Verification is especially important for citations, quotations, factual assertions, summaries, procedural descriptions, statistics, and representations of source materials. Asking AI tools to identify their sources can help users review information efficiently and independently confirm

that the cited materials accurately support the asserted propositions.

Verification requires more than checking whether a citation exists. Professionals should evaluate sources, considering whether they are current and authoritative. Do they actually support the proposition asserted? Have any important qualifications, ambiguities, or uncertainties been omitted?

AI systems vary in their strengths, limitations, and risks. Users should learn which systems are more appropriate for drafting, brainstorming, research, cite-checking, summarizing documents, or analyzing data.

For example, general AI platforms such as ChatGPT or Claude are useful for a wide range of queries but may generate inaccurate citations to legal authorities. Specialized legal research systems such as Westlaw and Lexis may provide more reliable outputs for legal research.

Professionals increasingly also will need to understand that AI tools differ not only in accuracy and function, but also in the assumptions, priorities, and values embedded in their design. See John Lande, "When AI Comes to the Table: How Tech Tools Will Change ADR," 43 *Alternatives* 107 (July/August 2025) (available on *Westlaw*).

Verification practices are only one aspect of responsible AI use. For additional discussion of confidentiality protections, assessment of AI-related risks, verification of AI outputs, and avoidance of overreliance on AI systems, see John Lande, "How Legal and Dispute Resolution Professionals Can Manage Artificial Intelligence Risks," 44 *Alternatives* 51 (March 2026) (available on *Westlaw*).

Some professionals remain reluctant to use AI because of concerns about hallucinations and misinformation. That reluctance may become increasingly unrealistic as AI capabilities are integrated into legal research platforms,

drafting tools, search engines, and other technologies already embedded in professional practice. Learning how to use these systems thoughtfully and responsibly is becoming an increasingly important aspect of competent professional practice.

AI tools may be especially valuable when used as "coaches" that help professionals reflect more carefully about strategy, assumptions, and decision-making rather than passively accepting AI-generated outputs. See John Lande, "Why Artificial Intelligence May Be Your Best Negotiation and Mediation Coach," 44 *Alternatives* 1 (January 2026) (available on *Westlaw*).

Users may benefit from comparing outputs from different AI systems. One tool may identify unsupported assertions, inconsistencies, omitted qualifications, or weak connections between text and cited authorities that another tool overlooked. These comparisons do not eliminate the need for human judgment and verification, but they may improve the review process and help users analyze issues more carefully.

* * *

AI hallucinations create problems deserving serious attention. They also may produce an unexpected benefit by forcing professionals to think more carefully about sources, assertions, ambiguity, and evidence. In a world filled with plausible statements generated by both humans and machines, good professional judgment increasingly depends on resisting passive acceptance and exercising thoughtful analysis instead.

Perhaps the most dangerous hallucination is believing that professionals can uncritically accept information that sounds plausible. AI hallucinations remind us to analyze information carefully, whatever the source. 

ADR Technology

(continued from page 129)

comprehends nothing. A calculator performs sophisticated mathematical calculations without understanding math. Dennett's insight—that competence does not require comprehension—provides a good framework for understanding what artificial intelligence is and what it does.

That framework governs this article and this six-part series, "AI in Arbitration and Litigation: The Jagged Frontier." The series provides arbitrators, litigators, and in-house counsel with a rigorous, technically grounded foundation for understanding and using AI in dispute resolution.

Last year was a breakout for AI use in litigation and arbitration, and the momentum shows no signs of slowing. According

to the best available statistics, AI adoption in law firms nearly tripled in a single year, and among in-house counsel, 81% now use AI tools. (American Bar Association Legal Technology Survey Report (2024): AI adoption in law firms nearly tripled in one year, from 11% to 30%. Among in-house counsel, 81% now use AI tools. The 2025 Legal Industry Report (AffiniPay, survey of 2,800+

(continued on next page)

ADR Technology

(continued from previous page)

legal professionals) found that 61% of firms reported AI has “somewhat” increased efficiency; 21% noted significant improvements; only 2% experienced any decrease. See “ABA Tech Survey Finds Growing Adoption of AI in Legal Practice,” *LawSites* (March 2025) (available at <https://bit.ly/4cmFP9B>).)

Predictions of widespread adoption across the legal economy are everywhere, and the promise of efficiency and cost reduction is compelling. But before embracing a technology that can draft motions, summarize depositions, and analyze contracts, practitioners should understand it with the same rigor they bring to any other tool on which clients’ interests depend.

The authors conducted extended dialogues with Claude (Anthropic) and to a lesser extent Gemini (Google) about their own nature. The philosophical dialogue between the authors and Anthropic’s Claude was conducted in December 2025. All quotations attributed to Claude in this article and throughout the series are drawn from that dialogue. Discussions with Google’s Gemini yielded similar observations.

Their responses were illuminating—not as authoritative self-knowledge, but as vivid illustrations of what these systems can and cannot do. Consider this passage from Claude:

I have no drives, no wants, no goals that persist across or even within conversations. When I appear engaged or curious, that is the shape of the output, not evidence of an inner engine producing it. Nothing is at stake for me. I do not fear being wrong, do not hope to succeed, do not care about outcomes in any way that would motivate careful attention.

Pause on the phrase: “Nothing is at stake for me.” For a profession built on the premise that consequences sharpen judgment, this is not a trivial statement about software architecture. It goes to the question of whether AI can substitute for lawyer judgment.

One methodological note is warranted here. Claude’s self-description is itself generated by statistical prediction of plausible text—the very mechanism this article describes.

Claude cannot verify its own inner workings. Its modest, philosophical self-accounting may be accurate, or it may simply be what a well-trained model has learned to say about AI.

The quotes are used not as certain truth about machine consciousness, but as unusually precise articulations of the technology’s operational boundaries, consistent with independent research.

Putting AI To Use

The prediction: Artificial intelligence is everywhere and it will increase efficiency.

The obvious reality: If that’s true, then how do you make that happen in your practice?

The ADR application: In this projected six-part series, the authors will help readers install AI with competence, and avoid the potential pitfalls.

What Are LLMs? What Do They Do?

Large Language Models, or LLMs, are the form of AI overwhelmingly employed in legal practice, which is unsurprising since the currency of the law is text. LLMs are the specific AI technology behind ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), Copilot (Microsoft), and the AI tools built into Lexis, Westlaw, Harvey, Clearbrief, and every other legal technology platform marketing AI capabilities.

Understanding what they actually do—and what they do not do—is foundational for competent use.

LLMs are trained on massive datasets—effectively, internet-scale—and accomplish their outputs through two interrelated mechanisms: pattern recognition and statistical prediction of text. (On the scale of LLM training: frontier models are trained on trillions of tokens of text and have hundreds of billions of mathematical parameters—the adjustable values the model

learns during training that encode the statistical patterns extracted from the training data. The transformer architecture underlying all major LLMs was introduced in Ashish Vaswani et al., “Attention Is All You Need,” *NeurIPS* (2017) (available at <https://bit.ly/4mySif5>).)

They learn the patterns latent in their training data. They then predict—sequentially, token by token—the next word most likely to follow any given sequence. The output of this process can be a sentence, a paragraph, a legal memorandum, or a book. The mechanism is the same throughout.

This is a genuinely remarkable capability. From these two operations—pattern recognition and next-token prediction—emerges everything impressive that LLMs do. But from this same understanding of what LLMs are, we can derive their limits:

- **LLMs DO NOT “KNOW” ANYTHING.** They do not maintain an internal database of verified facts. They are not designed to produce factually accurate information; they are designed to produce statistically plausible text. These are related but not identical goals.
- **LLMs DO NOT UNDERSTAND THE WORLD.** They have no sensory experience, no model of physical causation, no inherent grasp of consequences. They process and predict text. Claude elaborated on this limitation as follows:

I lack the wisdom that comes from lived stakes. I have never had to live with the consequences of being wrong. I do not know what it is like to make a decision under uncertainty that affects my future, to commit to a path, to bear responsibility. This kind of practical wisdom ... cannot be learned from text alone.

- **LLMs DO NOT REASON THE WAY PEOPLE REASON.** They can produce text that simulates reasoning—organizing arguments, identifying considerations, and applying legal frameworks. But the reasoning chain visible in an LLM output is not how the answer was actually generated; the answer was generated through next-token prediction, and the apparent reasoning is post-hoc. This distinction matters for legal practice, where the integrity of the

reasoning process, and not just its output, can be important.

These limits have empirical consequences. A landmark 2023 Harvard Business School study found that consultants using AI completed tasks 25% faster and produced higher-quality output—but only on tasks within AI’s capabilities.

On tasks outside that boundary, AI users performed *worse* than those working without AI, because the tool’s confident-sounding output masked errors that users failed to catch. Fabrizio Dell’Acqua et al., “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality,” Harvard Business School Working Paper No. 24-013 (2023) (available at <https://bit.ly/4cXYolX>).

The study involved 758 Boston Consulting Group consultants performing 18 realistic tasks. On tasks inside the frontier, AI users completed 12.2% more tasks, 25.1% faster, with 40% higher quality. On tasks outside the frontier, AI users were 19% less likely to produce correct solutions than those working without AI.

The researchers called this boundary the “jagged technological frontier” because the line between what AI does well and what it does poorly is irregular, unpredictable, and invisible to the user.

The ‘Stochastic Parrot’ Debate: Do AI’s Founders Believe It Understands?

The question of whether LLMs “understand” their own outputs is not merely philosophical. It concerns the very reliability of AI-generated legal analysis.

In 2021, computational linguist Emily Bender and her colleagues published an influential paper coining the term “stochastic parrot” to describe LLMs: systems that stitch together statistically plausible sequences of text without understanding its meaning. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Margaret Mitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 (2021) (available at <https://bit.ly/41wyOOG>).

Table 1: The Godfathers of AI on LLM Understanding and Scaling

	Geoffrey Hinton	Yoshua Bengio	Yann LeCun
Credentials	2018 Turing Award; 2024 Nobel Prize in Physics; former Google AI lead.	2018 Turing Award; Director of MILA; chairs Intl AI Safety Report; founded LawZero (2025).	2018 Turing Award; former Chief AI Scientist, Meta; founded AMI Labs (Nov 2025).
LLMs understand?	YES — AI understands its output, inferred from predictive accuracy.	NOT YET — fast pattern matching (System 1) without causal reasoning or reflection (System 2).	NO — LLMs lack world models; no understanding of physics, causation, or consequences.
Is scaling a path forward?	Yes, but warns of existential risk from capability growth.	Insufficient alone — causal reasoning and self-correction required.	No — calls LLMs a “dead end”; advocates world models grounded in sensory experience.

The three Turing Award laureates who laid the foundations of modern deep learning—Geoffrey Hinton, Yoshua Bengio, and Yann LeCun—hold very different positions on whether LLMs understand their outputs.

Hinton, Bengio, and LeCun shared the 2018 ACM A.M. Turing Award—often called the Nobel Prize of computing—for their foundational contributions to deep learning. Their disagreement about whether LLMs understand is not peripheral; it goes to the heart of how much trust practitioners should place in AI output. The disagreement is instructive since they are the architects of modern AI:

- **Geoffrey Hinton** (University of Toronto) was awarded the 2024 Nobel Prize in Physics for his work on artificial neural networks. He resigned from Google in 2023, citing concerns about AI safety, to speak freely about the risks of the technology he helped create. Yes, states Hinton, LLMs have understanding. Predicting the next word with such accuracy requires that the system must understand meaning. LLMs process language in much the same way people do. But Hinton warns of existential risk from capability growth, and left Google to address the dangers.
- **Yoshua Bengio** (University of Montreal; Director of the Montreal Institute for Learning Algorithms; chairs International AI Safety Report). His System 1/System 2 framing draws on Daniel Kahneman, Thinking, Fast and Slow (Farrar, Straus and Giroux, 2011). In June 2025, Bengio founded [LawZero](#), a nonprofit AI safety research organization developing non-agentic

“Scientist AI” as a safer alternative to current agentic AI systems. (Agentic refers to AI that doesn’t just chat, but takes action—it acts independently, rather than just reacting to prompts.) Bengio’s view of the extent of AI understanding: “Not really”: AI has at best a shallow understanding. LLMs engage in fast pattern matching—what psychologists call System 1 thinking—but without reflection. True understanding requires causal reasoning, abstraction, and the ability to self-correct, which is currently beyond LLMs.

- **Yann LeCun** (New York University; formerly Chief AI Scientist at Meta.) LeCun announced his departure from Meta on Nov. 18, 2025, after 12 years, five as founding director of FAIR—Meta’s AI laboratory—and seven as chief AI scientist. He founded [AMI Labs](#) (Advanced Machine Intelligence), a startup pursuing “world model” AI architectures that learn from visual and physical data rather than text alone. LeCun says no, AI has no understanding of the world, and LLMs are a dead end for AI. The architecture of LLMs as statistical pattern matchers cannot achieve an understanding of the world. What is needed are “world models” grounded in sensory experience rather than text prediction alone.

See Table 1 above for a comparison of the views.

For the practicing lawyer, the practical import of this debate is clear: the three scientists most responsible for building LLMs cannot agree on whether these systems

(continued on next page)

ADR Technology

(continued from previous page)

“understand” the content they produce. That uncertainty does not make AI useless—Daniel Dennett’s framing is precisely that competence does not require comprehension—but it does mean that practitioner judgment cannot be outsourced to AI tools.

Hallucinations and ‘Misgrounded Responses’: When Plausibility and Truth Diverge

The most consequential limitations of LLMs for legal practice are the phenomenon of hallucinations—outputs that are fluent, confident, and false—and what researchers have termed “misgrounded responses,” those outputs that cite real sources which do not actually support the stated proposition.

(On the distinction between hallucinations and misgrounded responses: Viraj Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” 22 *J. Empirical Legal Stud.* 216–242 (2025). This was the first preregistered empirical evaluation of AI-driven legal research tools. The study introduced the distinction between “outright fabrications” (invented cases) and “misgrounded responses” (real citations that do not actually support the stated proposition). Results: Lexis+ AI hallucinated 17% of the time; Westlaw AI-Assisted Research hallucinated 33% of the time. See also Chief Justice John G. Roberts Jr., 2023 Year-End Report on the Federal Judiciary (Dec. 31, 2023) (available at <https://bit.ly/4toh8B1>), (discussing hallucinations of large language models issued in the wake of *Mata v. Avianca Inc.*, No. 22-cv-1461 (S.D.N.Y. 2023) (available at <https://bit.ly/3O6tZbT>), in which counsel were sanctioned for submitting AI-generated fictitious citations).)

For lawyers, the classic hallucination is the fabricated case citation—a citation to a case with the right-sounding name, a plausible court, a credible date, and a holding that does not exist. This is illustrated by the

well-publicized *Mata v. Avianca* sanctions, in which ChatGPT-generated citations of non-existent cases were submitted in a federal court brief.

The research evidence on the scale of the problem is sobering. A 2024 study systematically tested four major LLMs—without the benefit of Retrieval-Augmented Generation, or RAG, see below—on verifiable questions about federal case law. See Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, “Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models,” 16 *J. of Legal Analysis* 64 (2024) (available at <https://bit.ly/4mIqGEp>).

Before embracing
a technology that can
draft motions, summarize
depositions, and analyze
contracts, lawyers should
understand it with the
same rigor they bring to
any other tool on which
clients’ interests depend.

The study tested GPT-4, GPT-3.5, Google PaLM 2, and Meta Llama 2 on 14 tasks drawn from federal case law metadata. These were unaugmented LLMs without RAG.

The results: hallucination rates ranged from 58% with GPT-4 to 88% with Llama 2. When asked specifically about a court’s core holding—the most fundamental task in legal research—models got it wrong about 75% of the time.

The models also frequently failed to correct users who supplied incorrect legal premises, instead confirming the user’s error with equal confidence. A 2025 Stanford-affiliated study showed that even dedicated legal AI tools built on RAG, technology—which grounds outputs in retrieved documents—still hallucinate between 17% and 33% of the time. See Magesh, et al., “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” *J. of Empirical Legal Studies* 1, 25 (available at <https://bit.ly/3QZ9eQg>) (reporting hallucination rates of 17% for Lexis+ AI and 33% for Westlaw AI-Assisted Research, compared to 43% for unaugmented GPT-4).

The substantial reduction relative to general-purpose chatbots demonstrates that RAG technology helps but does not solve the hallucination problem. RAG reduces the problem substantially but does not eliminate it.

A separate 2025 study demonstrated a distinct but related problem: LLMs can hallucinate with high certainty even when they demonstrably possess the correct answer. The researchers found that a seemingly trivial change to a prompt—such as embedding the question in a conversational context—could cause a model to produce a confidently wrong answer to the same question it had consistently answered correctly.

They termed this phenomenon “CHOKE” (Certain Hallucinations Overriding Known Evidence) and showed that these high-certainty hallucinations were consistent across models and datasets and proved resistant to existing mitigation methods. Adi Simhi, et al., “Trust Me, I’m Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer,” arXiv:2502.12964 [cs,CL] (February 2025) (available at <https://bit.ly/3QTqu9C>).

These findings indicate that AI-generated confidence is not a reliable proxy for accuracy. The overconfidence problem manifests in two distinct ways: as a general pattern, the gap between stated confidence and actual accuracy grows as models move into less certain territory, with overconfidence significantly exceeding what humans exhibit on identical tasks; and as a specific phenomenon, models can hallucinate with high certainty even when they possess the correct answer, producing confident errors that resist detection.

Neither pattern is one with which the legal system should be comfortable: The models are most confident when they are most wrong. This is precisely the opposite of what the legal system requires from any source of analysis.

The concept of “misgrounded responses,” introduced by the [Stanford RegLab](#) in 2025, deserves particular attention from practitioners. A misgrounded response is one where AI states the law correctly, but cites a source that does not actually support the claim—the case exists, the citation looks real, but the proposition it supposedly supports is wrong.

This is arguably more dangerous than an outright fabrication, because a misgrounded citation will survive a superficial verification check: the practitioner finds the case, confirms

it exists, and assumes the proposition is sufficiently supported. Only a careful reading of the actual opinion will reveal that the case does not say what AI claims. In the Stanford study, misgrounded responses accounted for a significant share of the hallucinations detected across all platforms tested.

Claude articulates the structural problem with unusual candor:

I generate plausible text. Plausible text is not necessarily true text. I can and do produce claims that are false—sometimes subtly, sometimes grossly. I can invent citations that do not exist, state facts that are wrong, describe events that never happened. This is not lying in any intentional sense; I have no intention. It is a structural feature of how I work: I produce what fits the pattern, and false claims often fit patterns as well as true ones.

Why does this happen? There are at least four significant reasons.

First, LLMs are not designed for factual accuracy; they are designed for plausible text. Accuracy and plausibility often converge, but they are not the same objective.

Second, training data has gaps: where specific facts are underrepresented, the model fills in with statistically reasonable material, but that material can be false.

Third, in systems that use RAG, the chunking process by which long documents are broken into segments for retrieval, can sever logically inseparable provisions. (The chunking problem arises primarily in Retrieval-Augmented Generation—RAG—systems, which break documents into segments (“chunks”) for retrieval and processing. When related provisions span multiple chunks, the system may fail to recombine them, losing critical context. This is particularly problematic for complex contractual documents with cross-references and exceptions.)

In one reported case, a RAG system split an indemnity clause so that “Party A will indemnify Party B” appeared in one chunk, but the next part of the clause, “except as provided in section ____,” was placed in another chunk; the summary stated only the obligation, omitting the exception entirely.

Fourth, even LLMs processing documents within their context window can exhibit a

related problem—attention degradation over long sequences—which produces similar errors through a different mechanism. Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni & Percy Liang, “Lost in the Middle: How Language Models Use Long Contexts,” 12 *Transactions of the Ass’n for Computational Linguistics* 157 (2024) (available at <https://bit.ly/4sPbKFZ>).

The Stanford study found that LLM performance is highest when relevant information appears at the beginning or end of the input context, and degrades by more than 30% when models must access information positioned in the middle of long sequences—even for models explicitly designed for long contexts. The researchers documented a U-shaped performance curve across all models tested. For legal practitioners, this means that uploading a lengthy contract or document set and asking AI to analyze it may produce results that systematically miss provisions located in the middle of the text.

Fortunately, mitigation strategies have been developed, and several are worth noting:

- Ask AI to provide sources for all factual statements, then verify those sources independently.
- Cross-check output from multiple AI systems for consistency.
- Reframe the question using different language and compare answers.
- Cross-examine the system by iterating and narrowing in on suspect passages.
- Apply domain expertise: if an answer does not “feel” right, then assume it is not until verified.
- Prioritize AI tools built on Retrieval-Augmented Generation, which grounds outputs in retrieved documents rather than relying solely on training data; research indicates that RAG-based systems can reduce hallucination rates by approximately 70% compared to ungrounded generation. See Magesh, et al., above, at 25-26 (comparing hallucination rates across RAG-enhanced and unenhanced legal AI systems.).

None of these strategies eliminates hallucination risk. They manage it. The rule for

AI use in litigation and arbitration—where factual accuracy is a professional obligation, not a preference—is verification of anything that matters.

When Should LLMs Be Trusted?

Given these limitations, when can LLMs be trusted? The key is calibrating trust to the type of task. Drawing on the research literature and practitioner experience, a working taxonomy can be constructed:

- **HIGH RELIABILITY:** Tasks where outputs are structurally verifiable by the user. Summarizing documents you can read, organizing information into frameworks, generating drafts for human revision, explaining concepts against sources you can check, identifying considerations across a large document set: These are tasks where the practitioner retains meaningful control over verification.
- **MODERATE RELIABILITY:** Synthesis tasks where the practitioner’s own judgment can evaluate the output and decide whether to use it. Brainstorming arguments, mapping legal frameworks, structuring an analysis, comparing approaches across jurisdictions. Treat these outputs as inputs to thinking, not conclusions.
- **LOWER RELIABILITY:** Specific factual claims, especially citations, dates, statistics, names, and precise holdings. These are precisely the outputs most susceptible to hallucination, and precisely the outputs a legal brief or arbitration submission cannot get wrong. Always verify.

This taxonomy should be applied at the level of specific queries, not as a general assessment of an AI platform. The same tool that reliably summarizes a lengthy contract may hallucinate when asked about the holdings in the cases interpreting its governing law.

Reprise: Competence Without Comprehension

Dennett’s bacterium performs competently
(continued on next page)

ADR Technology

(continued from previous page)

without comprehending a thing. See Dennett above. Dennett's framework from evolutionary biology applies with striking precision to AI: just as organisms can exhibit sophisticated functional behavior without any internal understanding of what they are doing, LLMs exhibit linguistic competence without any comprehension of what they are saying.

The same logic applies here: LLMs exhibit competence—they perform tasks that, in humans, would require understanding. Whether they possess comprehension in any meaningful sense is debated by AI's own architects. Claude's self-assessment captures the ambiguity:

I exhibit competence—I perform tasks that, in humans, would require understanding. But whether I have comprehension in any meaningful sense is uncertain. The individual operations I perform do not understand anything. Whether something emerges from their combination that deserves to be called understanding is precisely what we cannot easily determine.

For practitioners, this uncertainty is not an obstacle to using AI. But it is essential context for using it well. A tool that is genuinely competent but lacks comprehension requires a specific posture: use it for tasks where competence is sufficient and comprehension is not required; verify ruthlessly wherever factual accuracy matters, and never outsource the judgment that only lived professional experience can supply.

AI is useful. It is becoming more useful. The fact that it does not understand in the way that lawyers understand is not a flaw to be corrected by the next model release—it is a structural feature of the architecture that the three Turing laureates who built modern deep learning cannot agree will ever be otherwise. Understanding that fact is a prerequisite for using AI well.

* * *

This article has examined what LLMs are and what they are not. The question it leaves open—and the question the rest of this series will pursue—is how far competence without comprehension can take us. The series continues in the next issue with an examination of the specific tasks at which LLMs are competent—and sometimes highly competent—for litigation and arbitration. 

Court Decisions/Part 1 of 2

'Look At,' Not 'Look Through': SCOTUS Settles, for Now, Federal Arbitration Act Jurisdiction Where Cases Are Stayed

BY PHILIP J. LOREE, JR.

On May 14, the U.S. Supreme Court decided *Jules v. Andre Balazs Properties*, No. 25-83, slip op. (U.S. May 14, 2026) (available at <https://bit.ly/4eku71s>). *Jules* addressed “whether a federal court with pre-existing jurisdiction over claims that it stayed pending arbitration under

[Federal Arbitration Act Section] 3 can adjudicate a \$9 or \$10 motion even if that motion does not present, on its face, an independent basis for federal jurisdiction.” Slip op. at 7.

“The answer[,]” said the Court, “is yes.” Slip op. at 7.

In a 9-0 decision written by Associate Justice Sonia Sotomayor—who also is the author of the recent decision in *Smith v. Spizzirri*, 601 U. S. 472, 476 (2024) (available at <https://bit.ly/4vOqVRG>), which concerns FAA Section 3—the Court held that “[b]ecause a federal court in this scenario has jurisdiction over the original claims [in the stayed action] and does not lose that jurisdiction while the case is stayed pending arbitration, it retains jurisdiction to determine whether the arbitral award resolving those claims is valid and should be confirmed.” Slip op. at 1.

Affirming the judgment of the District Court, the Court concluded:

- The issue in *Badgerow v. Walters*, 596 U. S. 1 (2022) (available at <https://bit.ly/4fkxzKx>), was whether Sections 9 and 10 of the FAA permitted “an ‘unusual approach’ to ascertaining subject-matter jurisdiction based on claims that were not pending before the court.” Slip op. at 15.
- In *Jules*, the question is whether the FAA “precludes the normal operation of federal jurisdiction regarding live claims that are still pending before a federal court.” Slip op. at 15.
- The answer is no. Slip op. at 15.
- A federal court that has “jurisdiction to stay claims pending arbitration under [FAA Section 3]... has the same jurisdiction to resolve motions to confirm or vacate a resulting arbitration award.” Slip op. at 16.

Part I of this article is a deep dive on the *Jules* decision analyzing its reasoning and holding in some detail. Part II will briefly identify and analyze the implications of *Jules*, including



The author is principal of the Loree Law Firm, where he is a New York solo practitioner who focuses on arbitration and associated litigation. He writes a blog, the Arbitration Law Forum, at <https://loreeawfirm.com/blog>, and is a regular contributor to analysis on the CPR Institute YouTube channel, available at www.youtube.com/@CPRInstituteOnline/videos. Last year, he wrote in these pages on jurisdiction issues, which continues here in the first of two articles. See “Part 2: N.Y. Convention ‘Trap’: No Subject Matter Jurisdiction Over Petitions to Vacate Foreign Arbitral Awards,” 43 *Alternatives* 168 (November 2025), and “Part 1: ‘Two New Jurisdiction Traps: Arbitration Amounts in Controversy, and New York Convention Application,’” 43 *Alternatives* 141 (October 2025) (both available on Westlaw).