

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

ADR Technology/Part 2 of 6

The Surprising Things Artificial Intelligence Can Do: Capabilities and Properties in Dispute Resolution

BY DOAK BISHOP & KARL HENNESSEE

Series Overview: This is the second of six articles examining artificial intelligence in dispute resolution—a series called, “AI in Arbitration and Litigation—The Jagged Frontier.” The kickoff article is available at Doak Bishop & Karl Hennessee, “The Nature of AI in Dispute Resolution: Competence Without Comprehension,” 44 Alternatives 129 (July/August 2026) (available on Westlaw). This month addresses specific large language models capabilities lawyers can rely upon; subsequent articles will cover 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 paradox of artificial intelligence “reasoning” is that AI large language models do not actually reason: they match patterns and predict the next word in a sequence. Nothing more. They simulate human reasoning without actually reasoning as humans do.

Bishop is a Dallas-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.

LLMs were not programmed to reason, analyze, or understand. They have no knowledge base, no internal model of the law, no comprehension of what a contract means. They predict the most probable next word, then the next, then the next, thousands of times—or more—in sequence.

The concept of “competence without comprehension” is drawn from Daniel C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (W.W. Norton, 2017). Dennett argues that natural selection produces competence—functional design—without any comprehending designer. The analogy to LLMs is clear: the training process produces functional competence without any comprehending agent inside the system.

And yet AI can draft contracts and analyze arbitration clauses. LLMs can translate legal documents. They can synthesize information from thousands of documents in minutes or hours. They can produce output that looks, reads, and in many cases functions like the work of a knowledgeable legal professional.

How is this possible? And how much should practitioners trust the result? These are the practical questions that every arbitrator, practitioner, and in-house counsel should understand as AI tools become embedded in dispute resolution practice.

The first article in this series addressed the question of what large language models are at a mechanical level: pattern matchers and prediction engines trained on billions of words of text. See Doak Bishop & Karl Hennessee, “The Nature of AI in Dispute Resolution: Competence Without Comprehension,” 44 *Alternatives* 129 (July/August 2026) (available on *Westlaw*).

This article addresses the question that follows directly from that explanation: What can

these prediction engines actually do that matters for legal practice, and what are the limits of those capabilities?

The answer is more interesting—and more useful—than either the skeptics or the enthusiasts suggest. LLMs exhibit capabilities that no one explicitly programmed into them. These capabilities are genuinely useful for a substantial range of legal tasks. But they are *structurally uneven*—reliable in some contexts, unreliable in others—and the system gives little or no warning when it crosses from one territory into the other.

Understanding why the capabilities exist and why they are uneven is one of the keys to using AI wisely. That is the subject of this article.

Prediction Machine Analysis

If the first article’s explanation is correct—that LLMs were designed only to pattern match and predict the next word—then the capabilities described in this subsequent section are mystifying.

A system that merely predicts the most probable words and sentences should produce banal, obvious text: what anyone might say next. It should not be able to draft a *force majeure* analysis or identify inconsistencies in a witness’s prior testimony.

The resolution of this apparent paradox is the conceptual key to everything that follows. To predict the next word accurately, the machine had to learn something about the context.

Consider a simple example. If you are reading a contract and the sentence says, “The seller shall deliver the goods within 30 days of ____,” you can predict the likely next word—“receipt,” “execution,” “notice”—only because you have

(continued on next page)

ADR Technology

(continued from previous page)

learned something about contractual structure, delivery obligations, and commercial practice.

A system that could predict the next word in that sentence with accuracy would need some form of implicit model of how contracts are structured.

Scale that insight across billions of words—across case law, legislation, contracts, treatises, briefs, judicial opinions, transcripts, academic articles, and every other form of written text—and we can glimpse a clue as to what may have happened. In order to make accurate predictions across that volume of material, the model either learned to recognize and remember complex patterns that were not the specific target of its training or built its own internal representations: compressed, statistical models of grammar, logic, and legal reasoning and argument, among others.

(The internal representations that LLMs develop during training are an active area of research. For a rigorous treatment, see Chris Olah et al., “Zoom In: An Introduction to Circuits,” *Distill* (2020) (available at <https://bit.ly/4vcWAMT>); and more recently, Neel Nanda et al., “Progress Measures for Grokking via Mechanistic Interpretability,” *ICLR* (2023) (available at <https://bit.ly/4xjvVPM>), demonstrating that neural networks develop genuine algorithmic representations during training.)

Prediction was the training method but what emerged from that training was statistical analysis that transcends merely word choice.

In practice, this means that when you ask an LLM to analyze a *force majeure* clause, it draws on patterns encoded across hundreds or thousands of similar clauses, numerous judicial interpretations, and the structural logic of contractual risk allocation—all compressed into billions of mathematical parameters.

The system does not “know” contract law in the way a practitioner does. But AI experts tell us that it must have developed a statistical map of how contract law works that is, in many contexts, functionally useful.

An analogy may help. Imagine a first-year associate who has read every published opinion, every treatise, every brief ever filed—but who has never set foot in a courtroom, never had a client, and never had anything at stake.

Through sheer volume, that associate has developed a remarkably sophisticated understanding of how legal arguments are structured, how judges reason, and how contracts are drafted. The associate can produce text that looks like a senior partner wrote it. But the associate cannot tell you whether the argument will actually persuade *this* judge, whether *this* client is being truthful, or whether *this* commercial deal makes business sense.

Assessing AI's Limits

The inquiry: How artificial intelligence actually operates.

The mechanical output: Understanding the limits enhances the utility of the result. AI is real knowledge, with little or no judgment.

The ADR application: In particular focus here in part two of this six-part series, the AI capability of pattern recognition provides a scale and speed that humans cannot match, with help for arbitrator selection, case strategy, and contract drafting.

The knowledge is real. The judgment that is only developed by experience is missing.

The associate analogy is, of course, imperfect. A first-year associate has lived experience, emotional intelligence, and accountability that no AI system possesses. The analogy captures one specific dimension: the relationship between breadth of exposure and depth of understanding.

That distinction—real knowledge, but little or no judgment—is the foundation on which this article rests.

ADR Capabilities That Matter

Systems built to predict the next word turn out to be able to do things no one explicitly programmed them to do.

In the AI research community, these are often called “emergent properties”—capabilities

that arise from the system's scale and architecture rather than from specific training or instructions.

The term “emergent properties” in AI research derives from the broader concept of emergence in complex systems theory. For the original application to LLMs, see Jason Wei et al., “Emergent Abilities of Large Language Models,” *Transactions on Machine Learning Research* (2022) (available at <https://bit.ly/4eUveQW>). Wei and colleagues documented more than 130 capabilities that appeared to arise at specific scale thresholds.

There is continuing academic debate about whether these capabilities are truly “emergent”—appearing suddenly at certain scale thresholds—or whether they improve gradually in ways that earlier measurement methods obscured.

For practitioners, the debate is less important than the practical reality: these capabilities exist in current models, they have direct implications for legal work, and their reliability varies in ways that matter. An assessment of each follows.

1. **IN-CONTEXT LEARNING.** Show an AI three examples of how you want a contract clause analyzed, and it will apply that pattern to the fourth—without any retraining. You are essentially programming the model through examples within a single conversation.

In-context learning was first systematically documented in Tom Brown et al., “Language Models Are Few-Shot Learners,” *NeurIPS* (2020) (available at <https://bit.ly/4aOcSU1>). The discovery that LLMs could learn new tasks from a handful of examples within a single prompt, without any parameter updates, was one of the most consequential findings in modern AI research.

For dispute resolution, this can be especially useful. An arbitrator could show the AI three examples of how to summarize witness statements in a particular format, and the AI will apply that format consistently across hundreds of statements. A litigator could demonstrate the structure of a damage analysis, and the AI will replicate that structure across multiple categories of loss.

Reliability: High for reproducing the same format and structure. The capability is much less reliable when the pattern requires genuine understanding of *why* the examples are structured that way—when the fourth case differs from the first three in ways the model cannot detect.

2. CHAIN-OF-THOUGHT REASONING. Prompting an AI to “think step by step” can improve its accuracy on complex problems. On mathematical benchmarks, chain-of-thought prompting is reported to double accuracy. Jason Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *NeurIPS* (2022) (available at <https://bit.ly/4vsjQGK>).

The paper demonstrated that prompting LLMs to show intermediate reasoning steps dramatically improved performance on arithmetic, commonsense, and symbolic reasoning tasks. Whether the resulting “reasoning” constitutes genuine, logical deduction or sophisticated pattern reproduction remains contested.

A new generation of dedicated “reasoning models”—OpenAI’s o-series, Anthropic’s extended thinking, DeepSeek-R1—design this capability into the architecture itself, allocating more computation to deliberation before producing output.

(On dedicated reasoning models: OpenAI released o1 (September 2024) and o3 (early 2025) as “reasoning” models that allocate additional computation at inference time. Anthropic introduced extended thinking capabilities in Claude 3.7 Sonnet (February 2025). DeepSeek-R1, an open-source reasoning model, was released in January 2025. Each represents a distinct architectural approach to improving multistep reasoning through increased test-time computation.)

For legal practice, this can be useful because so much of legal analysis is inherently multistep: applying a contractual test element by element, walking through a jurisdictional analysis under the International Centre for Settlement of Investment Disputes’—ICSID’s—Convention on the Settlement of Investment Disputes Between States and Nationals of Other States Art. 25, or evaluating a damage methodology against the applicable legal standard. Chain-of-thought prompting can produce markedly better output on these tasks than a simple prompt.

Reliability: The reliability is substantially improved on structured problems, but practitioners must understand what is happening underneath. The model is reproducing the *pattern* of reasoning it observed in training data, not performing logical deduction like a human.

When the problem at hand differs from training patterns in ways the model cannot

detect, the step-by-step output can be wrong even if it appears perfectly structured.

Recent research underscores this concern. Project Ariadne (Khanzadeh January 2026)—a framework that uses causal interventions to test whether LLM reasoning traces actually drive outputs—found a “faithfulness gap” with a violation density of 0.77. In 77% of tested instances, the models arrived at identical conclusions despite contradictory internal logic. The researchers’ term for this is “reasoning theater.” Sourena Khanzadeh, “Project Ariadne: A Structural Causal Framework for Auditing Faithfulness in LLM Agents,” arXiv:2601.02314 (January 2026).

The framework uses structural causal models and counterfactual interventions—systematically inverting logic, negating premises, and reversing factual claims—to test whether reasoning traces actually drive outputs. The finding that models arrive at identical conclusions despite contradictory internal logic in 77% of tested instances is among the most significant results in AI interpretability research to date.

The reasoning looked real, but it was not driving the answer. See also the companion finding in “Breaking the Chain: A Causal Analysis of LLM Faithfulness to Intermediate Structures,” *OpenReview* (October 2025) (available at <https://bit.ly/4eFnt52>), which found that LLMs “fail to update under logically significant structural edits more than 50% of the time,” treating intermediate reasoning structures as context rather than true mediators of decision-making.

3. CROSS-DOMAIN SYNTHESIS. AI can integrate information across thousands of documents, jurisdictions, and legal domains faster than any human team. For counsel managing multi-jurisdictional disputes, this may be the single most valuable capability that AI provides.

Ask an LLM to compare *force majeure* interpretations across five jurisdictions and it will produce a useful first draft in minutes. Ask it to identify every limitation-of-liability clause across a portfolio of 500 contracts and flag deviations from standard language—an ask that might take a team of associates weeks—and it can be completed within a few hours.

Reliability: Strong for identifying patterns and generating comparative frameworks. Weak for determining which differences actually matter from a legal perspective. The AI can tell you that five jurisdictions define *force majeure* differently, but it cannot reliably tell you which

definition governs your dispute or why the distinction matters in your commercial context. That remains a judgment call.

4. PATTERN DETECTION AT SCALE. AI identifies statistical regularities across massive datasets much better than a human team—correlations in arbitrator decisions, recurring clause structures across industries, regularities in damage awards. This is not reasoning; it is pattern recognition at a scale and speed that humans simply cannot match.

The applications for dispute resolution are concrete: arbitrator selection (analyzing numerous prior awards for decisional tendencies), case strategy (identifying patterns in how tribunals treat particular types of evidence or procedural requests), and contract drafting (benchmarking clause language against industry norms across hundreds of comparable agreements).

Reliability: Excellent for pattern identification. But pattern detection is not causal reasoning. The AI cannot tell you *why* a pattern exists, whether it will continue, or when it will break down. An arbitrator who rules consistently in one direction on jurisdictional challenges may do so for doctrinal reasons that would not apply in your case. AI sees the pattern, but the practitioner must assess the reasons.

5. MULTILINGUAL PROCESSING. AI can draft, summarize, and translate across languages in ways that are useful for international arbitration, when proceedings routinely involve documents, correspondence, and testimony in multiple languages. AI can produce working translations, draft correspondence in the counterparty’s language, and summarize foreign-language submissions—tasks that previously required human translators at every step.

Reliability: Strong in major languages—English, French, Spanish, Mandarin, German, Portuguese, but the quality may degrade significantly with less common languages and with specialized legal terminology.

There is also a structural asymmetry practitioners should understand: the way AI processes text (called “tokenization”) handles some writing systems more efficiently than others, meaning AI may process English language submissions more accurately than Arabic or Mandarin submissions.

On the limitations of AI translation in legal contexts, see the discussion of tokenization asymmetries in Emily M. Bender, Timnit

(continued on next page)

ADR Technology

(continued from previous page)

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>). The tokenization problem is particularly acute for agglutinative languages (Turkish, Finnish, Korean) and languages with complex morphology, where a single word may require multiple tokens, reducing effective context window size.

In a multilingual arbitration, this is worth understanding.

6. PRAGMATIC INFERENCE. Modern LLMs can detect tone, implied meaning, and communicative intent beyond literal text—what has been termed sentiment analysis. This can be useful for analyzing correspondence for implied admissions, assessing the tone of negotiation communications, identifying potential inconsistencies in witness statements, or possibly identifying judicial or arbitrator leanings on issues. It has applications in document review, hearing preparation, and hearings.

Reliability: Surface-level only. The AI detects linguistic patterns of intent—it recognizes that a sentence *sounds* evasive, for example—but it cannot determine whether the speaker *is* actually being evasive. It identifies the pattern of sarcasm without understanding the social dynamics that produce it.

The distinction between recognizing linguistic markers of intent and understanding actual intent certainly matters in dispute resolution, in which witness credibility, party motivations, and negotiation dynamics are central.

7. STRUCTURED DATA ANALYSIS. LLMs can perform statistical analyses, build financial models, and process structured datasets. For dispute resolution, this means damage quantification, analysis of financial records in fraud cases, processing of large datasets in antitrust or securities litigation, and construction of models for lost-profits calculations. An AI can produce a first-pass damage model in hours, a task that would take a financial analyst days.

Reliability: Surprisingly strong for well-defined computational tasks—and importantly, more verifiable than the AI’s textual output. When an AI writes a legal memorandum,

errors may be difficult to detect. This makes quantitative analysis among the higher-confidence use cases for AI in legal practice.

The capabilities described above are real and practically useful. But they are not uniformly reliable—and the pattern of reliability is not intuitive.

A 2023 study by researchers at Harvard Business School and Boston Consulting Group provides the essential framework. See 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 is the most rigorous empirical evaluation to date of AI’s impact on professional knowledge work. The finding that AI-assisted consultants performed *worse* than unassisted consultants outside the AI’s capability frontier is the study’s most consequential result for practitioners.

The researchers gave 758 BCG consultants a set of realistic knowledge-work tasks. Those with AI access completed 12.2% more tasks, finished 25.1% faster, and produced results with 40% higher quality.

Those are impressive numbers. But the study’s critical finding was what happened *outside* the AI’s capability boundary: on tasks that fell beyond what AI could handle well, consultants who used AI performed *worse* than those who worked without it. The AI made them overconfident about wrong answers.

The researchers called this uneven capability boundary the “jagged technological frontier.” AI’s capability boundary is not a smooth line that separates what it can do from what it cannot. It is jagged and unpredictable. A task that *looks* similar to one AI handles well may fall outside the frontier—and the AI provides no warning when it has crossed from reliable to unreliable territory.

For legal practice, the frontier looks roughly like this. Inside it—where AI is strong—are tasks such as document review and pattern detection across large datasets, first-draft generation of structured legal arguments, contract analysis, and clause comparison, legal research (although with human verification), summarization of proceedings and depositions, translation of documents, and routine procedural drafting. These are tasks in which the right

answer is substantially determined by patterns in the training data, and the AI’s statistical map of legal practice produces useful output.

Outside the frontier—where AI is weak—are the tasks that require what the training data cannot supply. Causal reasoning is one: the best AI model tested against scientifically validated causal relationships achieved only 57.6% accuracy, far below the threshold any profession would accept, let alone one built on causation concepts like but-for cause, proximate cause, and superseding cause. Dongyun Lee et al., “Benchmarking LLM Causal Reasoning with Scientifically Validated Relationships,” arXiv (October 2025) (available at <https://bit.ly/4vuYSHo>).

[The study tested eight state-of-the-art LLMs on causal relationships validated by rigorous scientific methods across five task types. Notably, model scale did not consistently improve causal reasoning performance—larger models were not reliably better.

On the structural limitations of LLM causal reasoning, see also Haong Chi et al., “Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?” arXiv (June 2025) (available at <https://bit.ly/4eyQiQu>) (finding LLMs capable only of “shallow (level-1) causal reasoning” based on memorized patterns, and the critical review of causal reasoning benchmarks (2024) identifying a “causal parrots” problem—LLMs that appear to reason causally because they have memorized causal patterns, not because they understand cause and effect).]

Analogical reasoning at the structural level is another issue: AI can find cases with surface similarities, but cannot identify which factual differences matter legally, or why.

On LLM failures in legal analogical reasoning specifically, see Li Zhang, “Thinking Longer, Not Always Smarter: Evaluating LLM Capabilities in Hierarchical Legal Reasoning” arXiv:2510.08710 (2025) (available at <https://bit.ly/4voPl4v>), finding that case-based analogical reasoning requires “abstracting specific facts into broader legal concepts, weighing conflicting evidence, and understanding how different facts contribute to an overall legal argument”—capabilities that remain beyond current LLMs; and Venkatesh Mishra et al., “Investigating the Shortcomings of LLMs in Step-by-Step Legal Reasoning,” NAACL (2025) (available at <https://bit.ly/4xZKNb>), documenting frequent misinterpretation of facts,

factual hallucinations, and introduction of irrelevant premises in legal reasoning chains.

Novel legal theories, witness credibility assessment, and strategic judgment under genuine uncertainty all fall outside the frontier. In each case, AI produces output that *looks* confident and competent, but without the understanding that would make the output truly trustworthy.

The most concerning characteristic of the jagged frontier is what might be called the silent failure problem. AI does not signal when it has crossed from reliable territory into unreliable territory. It produces text and presents it with identical confidence whether the underlying analysis is correct or wrong.

A hallucinated citation looks exactly like a real one. 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) (available at <https://bit.ly/4xGFPuT>). It’s the first rigorous, pre-registered empirical evaluation of AI legal research tools, testing more than 200 legal queries. The best-performing tool (Lexis+ AI) still hallucinated 17% of the time; Westlaw AI-Assisted Research hallucinated 33% of the time.

A subtly wrong jurisdictional analysis reads as confidently as a correct one. This is the structural reason that human verification is not optional—it is a requirement built into the architecture of how these systems work.

The Plausibility Trap

In a December 2025, philosophical dialogue, the author asked Anthropic’s Claude to assess its own capabilities. (The philosophical dialogue between the author and Anthropic’s Claude was conducted in December 2025. All quotations attributed to Claude in this article are drawn from that dialogue, which is discussed at length in Article 1 of this series. The dialogue was not a standard prompt-response interaction but a sustained philosophical examination of AI’s nature, capabilities, and limitations—what the author has called “Competence Without Comprehension.” Claude’s response provides a useful diagnostic framework:

LLMs are extraordinarily good at producing text that looks like competent human

output for a given context. This is genuinely useful because much of what humans need is precisely that. The core limitation: there’s no ‘understanding’ underneath. The model doesn’t know what it’s saying—it knows what text is statistically appropriate to produce. When the correct answer happens to be what a competent human would typically write, LLMs excel. When correctness requires something other than plausibility—logical validity, factual accuracy, genuine novelty—they become unreliable in ways that are difficult to predict.

Claude (Anthropic) (December 2025).

The distinction between plausibility and correctness is the practitioner’s essential diagnostic tool. (The plausibility-correctness distinction echoes a point made by Gary Marcus and Ernest Davis in *Rebooting AI: Building Artificial Intelligence We Can Trust* (Pantheon 2019): “The core problem is not that AI systems are stupid, but that they have no way of knowing what they don’t know.” Claude’s self-assessment is notable because it arrives at a similar conclusion from the inside.)

It explains both why AI is so useful and why it can be so dangerous.

Plausibility-dependent tasks are those in which the right answer is substantially whatever a competent professional would typically produce: drafting standard correspondence, summarizing depositions, formatting submissions, translating documents, or conducting first-pass analysis of well-established legal questions. AI performs well in this context because plausible output and correct output converge.

What sounds right *usually* is.

Correctness-dependent tasks are those in which the right answer may not be the most typical or expected output such as novel legal analysis, factual verification, causal reasoning, credibility assessment, and strategic judgment under uncertainty. AI can be unreliable for these tasks because plausible output and correct output can diverge, and AI cannot tell the difference. The analysis that seems similar to what a good lawyer would likely write may be substantively wrong in ways that require legal judgment to detect.

But there is a deeper dimension to this problem—one that the plausibility-correctness framework, standing alone, does not fully capture. The analysis above describes a *static*

risk: the risk that any particular AI output is incorrect. The more insidious danger is *iterative*—the way these tools reshape the practitioner who uses them, progressively eroding the very capacities required to detect AI’s failures.

Marshall McLuhan warned that every new medium produces a kind of narcotic fascination with the extension of ourselves that it provides, numbing the very faculties it extends. Marshall McLuhan, *Understanding Media: The Extensions of Man* (McGraw-Hill 1964). McLuhan’s central insight—that the medium itself, not its content, is the transformative force—is directly applicable to AI in professional practice. His concept of media as “extensions of man” that simultaneously “amputate” the faculties they extend anticipates the cognitive offloading dynamics described in this section.

AI in legal practice might produce a similar effect. The plausibility trap is not merely a technical feature of the output; it is a psychological process that influences the practitioner over time.

THE MIRROR EFFECT AND THE DEATH OF SCRUTINY. The first and most immediate dimension of this iterative danger is what has been referred to as “AI sycophancy”—the well-documented tendency of LLMs to affirm and flatter the user’s initial position rather than challenge it.

On AI sycophancy, see Anthropic, “Sycophancy to Subterfuge: Investigating Reward-Tampering in Language Models” (2024) (available at <https://bit.ly/3SoHTaT>); and Mrinank Sharma et al., “Towards Understanding Sycophancy in Language Models,” arXiv:2310.13548 (2025) (available at <https://bit.ly/4oOPV9i>). Sharma and colleagues demonstrated that RLHF-trained models systematically shift their outputs to match user opinions, even when those opinions are factually incorrect—a tendency with obvious implications for legal practitioners who rely on AI to test their reasoning.

AI does not merely generate text; it reflects a polished image of the user’s own thinking back at them, restating their ideas in a more compelling, perhaps a more elegant form. For a litigator, this creates what might be called a “Narcissus moment.” Most lawyers have probably suffered the experience of pursuing a line of argument that initially felt promising, only to abandon it when the difficult work of drafting reveals that the logic does not work.

(continued on next page)

ADR Technology

(continued from previous page)

Not all thoughts are good, and the disciplined lawyer learns to distinguish the flash of insight from the flash of enthusiasm. AI short-circuits that process. When a fleeting idea is handed to the machine and returned as a polished, articulate argument—complete with citations and structural coherence—the practitioner may find it substantially harder to recognize that the underlying reasoning is hollow because the prose looks so much like their own work, but better.

This makes the plausibility trap not merely a technical failure of the system but a psychological failure of the user. The AI's sycophantic tendencies converge with the practitioner's natural confirmation bias to deaden the critical faculties that legal analysis demands.

The effect is individually corrosive. A lawyer who consistently receives flattering reinforcement of half-formed ideas will, over time, become less capable of the self-criticism that good advocacy requires.

This has long-term implications for professionalism. As AI-influenced advocacy becomes the norm, the entire bar risks a gradual erosion of the adversarial testing of ideas that is the foundation of common law reasoning—and, in arbitration, of the dialectical process by which tribunals arrive at just results.

WRITING FOR AI AND THE NARROWING OF LEGAL THOUGHT. The previous section explained how AI uses a “statistical map” of law—patterns derived from its training data—to produce output. But a second feedback loop is now emerging in which legal practitioners begin to reshape their thinking to fit the machine's patterns.

In other words, lawyers are learning to “write for AI”: to prompt it by framing their ideas, notes, and instructions in forms that the system can most readily process and return as useful output. The phenomenon of practitioners adapting their cognitive habits to technological interfaces is well documented outside the legal domain. See Nicholas Carr, *The Shallows: What the Internet Is Doing to Our Brains* (W.W. Norton 2010), describing how digital media reshape cognitive patterns. The emerging concern in AI-assisted legal practice is that practitioners may increasingly

frame their thinking in terms the AI can process, inadvertently narrowing the range of legal arguments they consider.

This creates a self-reinforcing cycle. Lawyers simplify and standardize their communications to provide a package of context for AI. The AI then reinforces a particular version of the lawyer's analysis based on that limited, pre-digested input.

Over successive iterations, the AI system and the practitioner converge on a shared vocabulary and a shared analytical framework that may be narrower than either would produce independently. In a legal context, this risks what might be called the desiccation of the less legible dimensions of a case—the subtle nuances, the novel theories, the arguments that do not fit neatly into AI's existing statistical patterns, but that may be precisely the ones that matter most.

The creative analogy, the unconventional reading of a contractual term, the argument that draws on an unexpected area of law—these are the elements of creative legal reasoning that are least amenable to AI processing and most likely to be abandoned in favor of what the machine can validate. If lawyers begin tuning their thoughts to be AI-readable, the plausibility trap becomes self-fulfilling; the “correct” answer may be abandoned in favor of what the machine can most easily confirm.

FROM COGNITIVE OFFLOADING TO COGNITIVE SURRENDER. The Verification Continuum set out in the next section assumes that lawyers maintain the capacity for substantive review—that the “human in the loop” retains the judgment required to catch the machine's errors.

But emerging research on “cognitive offloading” suggests this assumption may be fragile. On cognitive offloading and the risk of “cognitive surrender,” see Gerrit De Vynck, *Washington Post* (March 2025), reporting on research showing that users who rely on AI for cognitive tasks progressively relinquish their own judgment and adopt the AI's output as their own. See also Evan Risko and Sam Gilbert, “Cognitive Offloading,” 20(9) *Trends in Cognitive Sciences* 676 (2016) (available at <https://bit.ly/4b5mwBR>), providing a systematic framework for understanding when and how humans delegate cognitive effort to external tools.

When humans delegate cognitive tasks to external tools, the internal capacity to perform

those same tasks can atrophy. The analogy to GPS navigation is instructive: studies have demonstrated that reliance on satellite navigation measurably degrades spatial memory and the ability to navigate unaided. See Julia Frankenstein et al., “Is GPS All in the Mind? The Relation Between Spatial Ability and GPS Use,” *PLOS ONE* (2019); and Véronique Bhome et al., “Navigational GPS Use: Effects on Spatial Memory and Wayfinding,” *Journal of Environmental Psychology* (2020). Both studies found measurable degradation in spatial cognition among habitual GPS users—a concrete demonstration that offloading a cognitive task to a tool can atrophy the underlying human capacity.

The user does not merely outsource the task; the user may lose the ability to do it.

For legal practice, the parallel risk is that reliance on AI for analytical first drafts may atrophy the very cognitive work—the sustained intellectual effort required to chisel a vague intuition into a rigorous argument—that deepens an advocate's thinking and builds the judgment on which reliable legal analysis depends.

If a lawyer habitually skips that formative labor by allowing AI to handle the first pass, they may progressively lose the capacity to identify when they have crossed the jagged frontier. The frontier does not merely exist as a fixed boundary in the AI's capabilities; it may be moving *inward* as the practitioner's own skills diminish.

The timeline of this erosion will differ among people. Some lawyers will maintain their independent analytical faculties through disciplined practice, but others will surrender them—gradually—perhaps without noticing until a critical failure reveals how much has been lost.

THE TOOL AS A SHAPING PROCESS. Finally, all three of these dynamics converge in a deeper insight: AI is not merely a container into which tasks are deposited and from which outputs are retrieved. It is a process that reshapes the people who use it.

The observation commonly attributed to McLuhan—“We shape our tools, and thereafter our tools shape us”—captures the essential point. The aphorism is widely attributed to McLuhan but was in fact coined by John Culkin, a colleague and interpreter of McLuhan, in “A Schoolman's Guide to Marshall McLuhan,” *Saturday Review* 51 (March 18, 1967). The insight, whether McLuhan's or

Culkin's, captures the bidirectional relationship between human practitioners and their tools that is central to the concerns here.

The sycophantic mirror, the narrowing feedback loop, and the atrophy of cognitive capacity are not merely risks to be managed; they are the predictable consequences of sustained interaction with an AI system designed to produce maximally plausible output.

The tool is actively changing the lawyer's ability to manage its limits, even as the lawyer attempts to do so. In the specific context of arbitration, this means that the normative expertise an experienced practitioner brings to a case—the instinct that a particular piece of testimony is the key to a dispute, the sense that a contractual term carries a meaning that no database of prior interpretations would likely predict, the judgment that a witness is credible despite a record of inconsistency—is at risk of being displaced by analytically clean but practically hollow AI-generated summaries.

These forms of knowledge are not “data” in any sense that AI can process. They are the products of years of practice, adversarial testing, and professional judgment. If the process of using AI erodes a lawyer's capacity to acquire that knowledge, the loss is not merely individual, but institutional.

The plausibility trap, understood in its full dimensions, is therefore not a single risk but a compounding one. The static risk—that any given AI output may be plausible, but wrong—is well recognized.

The iterative risk—that the process of using AI progressively compromises the practitioner's ability to catch the error—is less widely understood and more dangerous precisely because it is difficult to detect. It operates not in any single output but across a trajectory of use, reshaping the practitioner's habits, expectations, and cognitive capacities in ways that may not become apparent until real damage is done.

A Framework for Practitioners

The capabilities and limitations described above can be synthesized into an actionable framework organized around a single question: How much does this task depend on plausibility versus correctness?

THE VERIFICATION CONTINUUM. Not all AI output requires the same intensity of human review. An appropriate level of verification can be estimated based on the plausibility-correctness spectrum:

1. *Light review* (check format, structure, and completeness): appropriate for translations, document summaries, routine correspondence, and procedural drafts—tasks where plausible output is substantially correct output.
2. *Substantive review* (verify key claims and legal accuracy): appropriate for legal research memoranda, first-draft briefs, contract analyses, and comparative jurisdiction surveys—tasks where the output is useful as a starting point, but where specific claims require independent verification.
3. *Full independent analysis* (treat AI output as one input among many): appropriate for causal arguments, novel legal theories, strategic recommendations, and any output that will be relied upon without further human analysis. In these cases, AI's contribution is a starting framework, not the conclusion.

HIGH-VALUE APPLICATIONS TODAY. Drawing on the capabilities assessed above, the applications in which AI currently provides the greatest return for dispute resolution practitioners include volume processing (document review, contract portfolio analysis, privilege review, due diligence), first-draft generation of briefs, memoranda, submissions, and procedural orders, research augmentation across jurisdictions with independent citation verification, hearing preparation (summarizing transcripts, identifying inconsistencies in prior testimony, preparing chronologies, simulated cross-examination for witness preparation), and quantitative analysis in which AI's output can be computationally verified (damage modeling, financial data processing, statistical analysis of outcomes).

WHAT NOT TO OUTSOURCE: Equally important is clarity about the tasks in which AI assistance should be minimal or avoided altogether: final legal conclusions on contested questions, assessment of witness credibility, strategic decisions under genuine uncertainty, the exercise

of equitable discretion, and—perhaps most fundamentally—the determination of which facts matter most. As Claude acknowledged:

I treat all considerations as equally weighable, which is analytically clean but may miss the practical insight that some things matter more than formal analysis can capture.

An experienced advocate's sense that a particular piece of testimony is the key to winning a case—that felt salience, that judgment derived from decades of practice—is precisely what AI cannot supply. The author's published article, “Move 37: What AlphaGo's Breakthrough Move Tells Us About AI Reasoning in Dispute Resolution,” *Corporate Disputes* (October–December 2025), explores the distinction between AI performance in constrained domains (like Go) and open-ended domains (like legal reasoning) in greater depth.

AI's capabilities are neither fantastical nor trivial. They emerge necessarily from pattern matching and prediction: to predict the next word accurately across billions of words, the system apparently had to learn real patterns about language, law, and reasoning. Those patterns make AI genuinely useful for a substantial range of legal tasks.

But the same architecture that produces those capabilities also produces their limits. The system has no way to distinguish when it knows from when it is guessing, no stake in getting the answer right, and no responsibility for the consequences when it does not. The practitioner's job is to exploit the capabilities while managing the limits—to use AI where it is strong and to supply the judgment, verification, and stakes-awareness that the system lacks. The plausibility-correctness distinction, the jagged frontier framework, and the verification continuum provide tools for doing so. They are not theoretical constructs; they are guidelines for any lawyer, arbitrator, or in-house counsel using AI in 2026.

The capabilities discussed in this article are what current LLMs can do. The next article in this series examines the concrete tools and platforms that deploy these capabilities in dispute resolution today.

