Neota Logic

WHITE PAPER

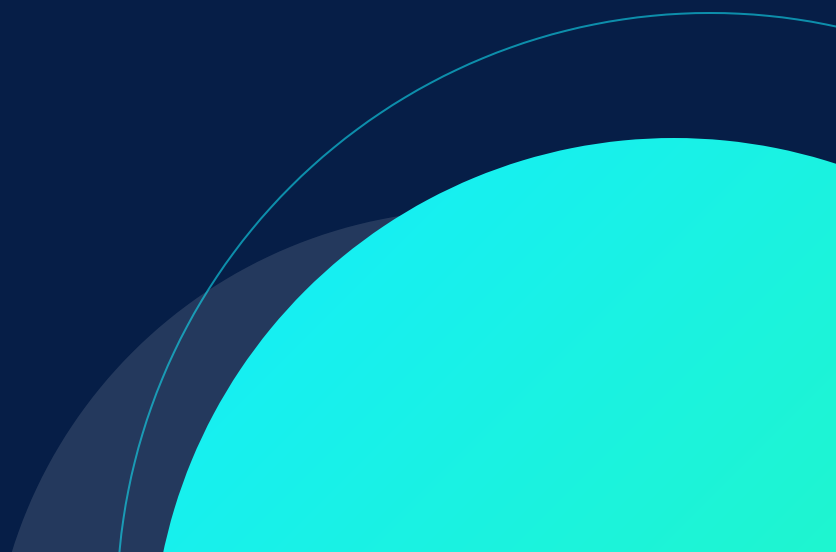
Same facts, different answer

A buyer's framework for defensible legal AI

Shaz Aziz

Head of Client Solutions, Neota Logic

Oct 7, 2026



EXECUTIVE SUMMARY

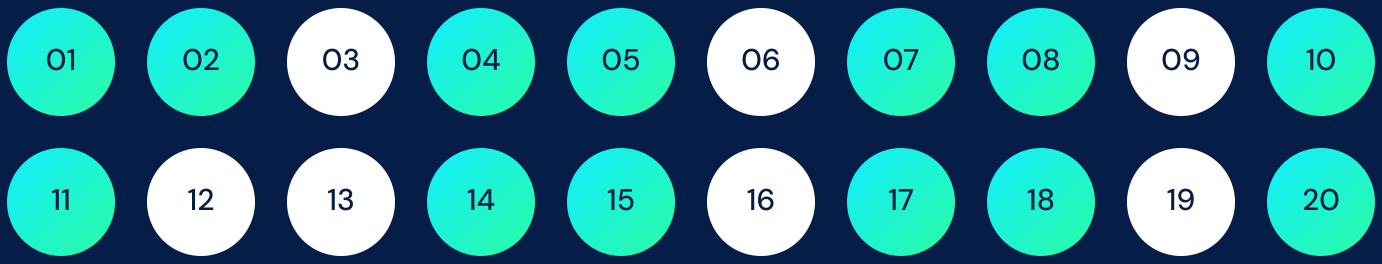
The risk in legal AI that gets discussed least is consistency: the same facts producing a different answer six months apart, with nothing in place to catch it. In 2025, researchers put 500 real federal appeals to leading models 20 times each, with identical text and the models set to be as consistent as their settings allow. On up to half of the questions, the same model named one party the winner, then the other.[1] A legal expert reviewed pairs of the contradictory answers and found the analysis sound in every one. The model simply weighed the arguments differently each time.

Legal work has always tolerated a version of this problem in people, because a lawyer's inconsistency sits inside an accountability structure: review, a file, insurance, a regulator. A model's drift has no equivalent. The major platforms spent 2026 shipping the first half of that structure, with audit logs, admin controls, and ethical walls that show who had access, what was configured, and increasingly what was said. None of it shows whether the answer held over time, or why it came out the way it did.

THE EXPERIMENT

Same case. Twenty runs.

Identical text. Settings tuned for consistency.



13 said Party A **7** said Party B

The reasoning was sound.
The answer wasn't repeatable.

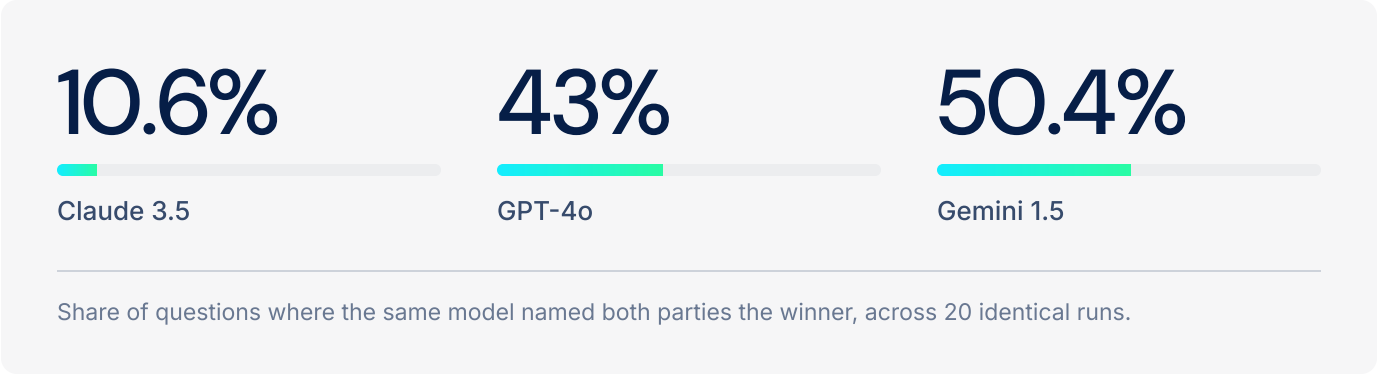
This paper sets out the evidence on why AI answers move, why human review cannot make them consistent, and why the accountability question is arriving on a regulatory clock rather than a procurement one. It closes with five questions to take into any vendor meeting, ours included, and what a defensible answer to each one sounds like.

01

Same question, twenty times

The clearest evidence comes from a 2025 study by Blair-Stanek and Van Durme at the University of Maryland and Johns Hopkins.[1] The researchers distilled 500 US federal appeals, each a case where the judges themselves split two to one, into a statement of facts plus the best arguments for each side. They asked each model which party should win, 20 times per case, with identical text every time and the models configured to be as consistent as possible.

Claude 3.5 named both parties as the winner on 10.6 percent of questions. GPT-4o did so on 43 percent. Gemini 1.5 did so on 50.4 percent. The expert review found sound reasoning on both sides of every flip. In one case on commercial speech, all three models went both ways on whether a restriction was "no more extensive than necessary". These are late-2024 models on deliberately hard questions, so treat the percentages as indicative. The researchers also ran o1 on a 50-question subset and it flipped on 28% of them, though o1 can't be set to temperature zero, so that isn't a like-for-like test. What matters is that it happens at all, and that nothing in a standard deployment would surface it.



Two distinct mechanisms move the answer

Drift.

The model or the system changes underneath you. Researchers at Stanford and Berkeley compared GPT-4 in March and June 2023 and found its accuracy at identifying prime numbers fell from 84 percent to 51 percent.[2] Same product, same name, and nobody sent a release note.

Server-side nondeterminism.

Thinking Machines ran a single prompt 1,000 times at temperature zero and got 80 different completions. The cause was how the inference server batches requests together. With that fixed, all 1,000 came back identical.[3] Their research shows it can be fixed, but only by whoever runs the model. Whether you get the same answer twice is an engineering choice made outside your organization.

Newer models show it too. A 2026 study had five models, including Claude Sonnet 4.5, Claude Haiku 4.5 and Gemini 2.5 Flash, score real enterprise question and answer pairs, and found substantial variability in the scores even at temperature zero.[4] Those are the kinds of scores enterprise pipelines use to route, triage and gate work, which is the job legal workflows give AI outputs.

> Accuracy is what gets checked in January, at the benchmark, the pilot, and the sign-off. Consistency is what you need to know about in July, when the same contract clause comes back with a different escalation decision and nobody is comparing.

Princeton researchers Narayanan and Kapoor call this the capability and reliability gap, and consistency is the first reliability dimension they measure.[6] Every vendor benchmark you have seen is a snapshot of one moment. OpenAI's headline for Astra for Law, the newest legal launch in the market, is 54 percent on a legal research benchmark, on its own testing.[5] A good number, and a snapshot. Legal work runs on the months in between.

02

Lawyers are inconsistent too, and the comparison fails anyway

The strongest objection to the consistency argument is that lawyers aren't consistent either. Ask twenty lawyers a hard question and you'll get a spectrum of answers. That's true.

The difference is structural. A lawyer's advice is produced by an associate, reviewed by a partner, recorded on the file, and backed by malpractice insurance, bar oversight, and a license worth protecting. The system is imperfect, but errors get found, priced, and sometimes corrected. A model's answer has no equivalent path. Nobody sees it change, there is no record of why it came out the way it did, and there is no license at risk. The lawyer leaves a paper trail. The model usually does not.

Human review is often offered as the fix, and it does real work: a reviewer can catch an answer that is wrong today. What a reviewer cannot do is make answers consistent, because review vouches for one answer at one moment. When the same clause produces a different escalation decision three months later, the January sign-off says nothing about it, and the reviewer who approved the first answer is not in the room for the second. A reviewer can only review what they can see, and drift is precisely what nobody is shown.

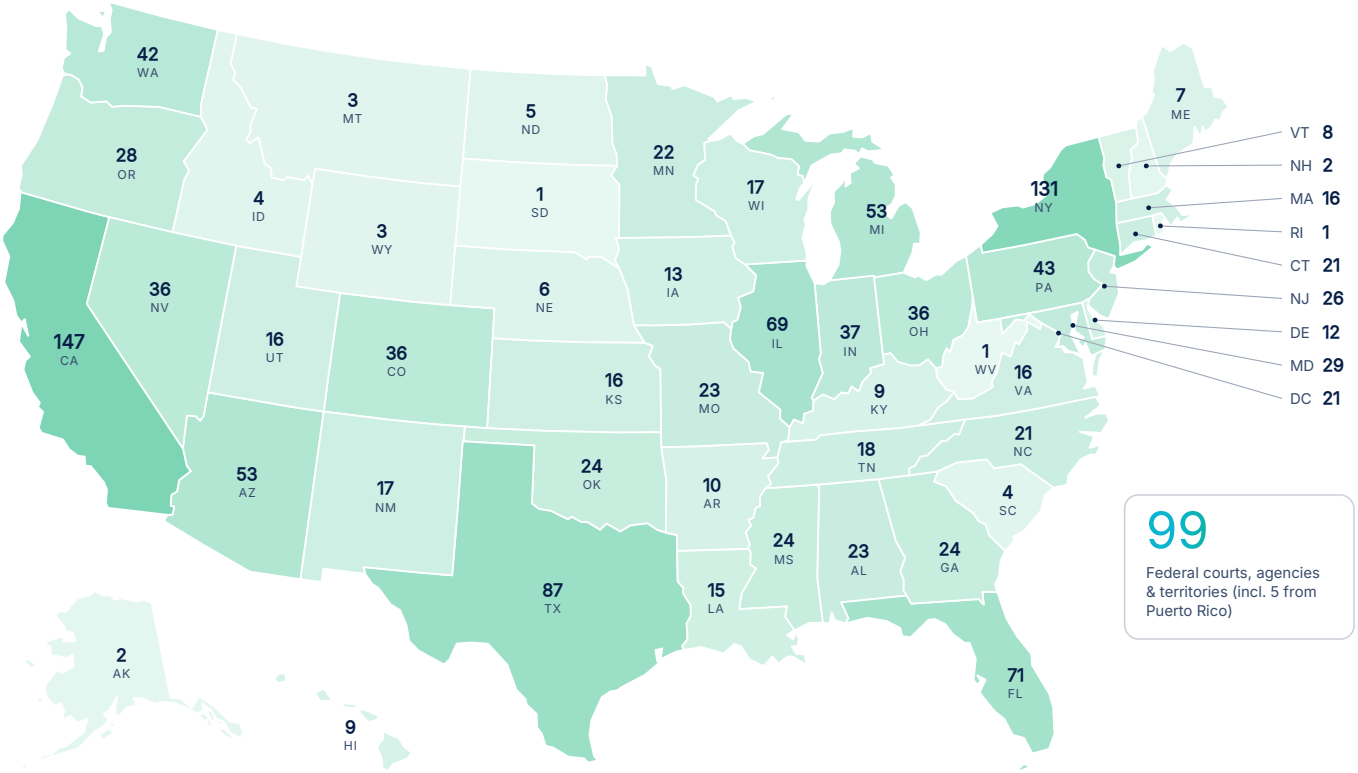
> "The lawyer leaves a paper trail. The model usually does not."

The scale of unchecked AI output is already on the public record. By September 2026, Damien Charlotin's AI Hallucination Cases database listed more than 1,400 US court decisions dealing with filings that contained fabricated AI content.[7] Every one of those filings had someone's name on it.

ALREADY ON THE PUBLIC RECORD

1,400+

US COURT DECISIONS INVOLVING FABRICATED AI CONTENT, BY STATE



DECISIONS PER STATE 1 147

Each figure is the number of US court decisions in Damien Charlotin's AI Hallucination Cases database, placed by the state of the court and current to late September 2026: 1,457 US decisions in total, including 99 from federal courts, agencies and territories. States are shaded by their total.

Source: AI Hallucination Cases Database, Damien Charlotin, CC BY 4.0.

03

The half-built governance layer

A year ago it was fair to say nobody had shipped a governance layer for legal AI. That is no longer accurate. Over the past twelve months the major platforms have been racing to add exactly the controls enterprise legal buyers asked for.

Anthropic shipped a Compliance API that pulls activity logs, chat data, and files from its enterprise product. Google launched Gemini Enterprise for Legal in preview, with a centralized control plane for IT and risk teams, audit logging, and ethical walls inherited from document and matter management systems through its connectors. Harvey integrated with Intapp Walls so ethical wall policies sync into its access controls. OpenAI is working with Latham & Watkins to design permissions, ethical walls, and firm oversight into Astra for Law.

All of this answers three questions: who had access, what was configured, and increasingly what was said. Those are necessary answers, and legal teams should welcome them. They are also only half of the layer. An access log tells you an associate queried the system on a Tuesday. It can't tell you whether the same query, a quarter later, came back different.

94%

of firms using or exploring generative AI (ILTA 2026)

71%

of firms using Harvey are still piloting it (ILTA 2026)[8]

23% to 52%

in-house generative AI use in one year (ACC / Everlaw)

Firms are trialing nearly everything and committing to very little. On the corporate side, the ACC and Everlaw found in-house use of generative AI more than doubled in a year, from 23 percent to 52 percent,[9] so the exposure now sits inside the legal department as much as at outside counsel. Pilots often stall because nothing makes the output dependable enough to stand behind in front of a client, and no volume of access logging changes that.

04

Three tiers of reliability

Not every use of AI in legal work carries the same consistency exposure. The useful way to sort a stack is by what actually produces the conclusion.

Tier 3 is where a decision becomes defensible

TIER 1 · LEAST RELIABLE

AI answering from its training data

Nothing grounds the answer in your facts or your policy. Fluent, fast, and the hardest to defend.

TIER 2 · STILL GENERATIVE

AI reasoning over your documents

Grounded in your material, but generation over retrieved text is still generation.

TIER 3 · DEFENSIBLE

AI gathers the facts, rules make the call

The model reads and extracts, the facts are confirmed, and versioned rules reach the conclusion. Same confirmed facts, same output.

A regulator has already approved the rules half of this. In February 2026 the Solicitors Regulation Authority authorized LawFairy, which describes itself as the first deterministic, technology-only law firm in England and Wales, starting with UK immigration.[12] Statutory and policy criteria are encoded as decision pathways validated by lawyers, no generative AI touches the regulated outcome, and the same facts produce the same result, in deliberate contrast to the probabilistic systems the rest of the market runs on. The open question for everyone else is how to add a language model's fluency without giving that up. It comes down to architecture. The model and the rules each do the work they suit, and neither does the other's job.

05

A risk arriving on a regulatory clock

The accountability question is no longer theoretical, and it is not waiting for procurement cycles.

Jan 2025

FTC final order against DoNotPay, with $193,000 in monetary relief, for marketing an AI legal service on untested claims.[11]

Dec 9, 2026

Deadline for EU member states to transpose the revised Product Liability Directive, bringing AI systems inside the legal definition of a product.

Regulators on both sides of the Atlantic are heading the same way: if you put an AI system into legal work, you answer for what it does.

Meanwhile there is no agreed standard for autonomy in legal AI, nothing like the SAE levels that define who is driving at each stage of a self-driving car, and that lawmakers use to decide who is liable. Legal AI today works like driver-assist, and with driver-assist the driver is still liable. That puts the liability where the review sits, on the lawyer who approved the output. A reviewer can't sit down and check the reasoning inside a model. They can check logic a lawyer wrote.

Most legal teams have not priced this risk, and the reason is understandable. The models are right most of the time, and a fast, plausible answer is hard to resist. Three things are arriving at once: product liability for AI software, enforcement against untested claims, and no autonomy standard to share out responsibility. Teams that can show why an answer came out the way it did will be in a very different position from teams that can only show who had access.

06

Five questions for any vendor, including us

Take these five questions into your next vendor meeting. Under each is what a defensible answer sounds like. Any vendor serious about legal work should be able to answer all five, and that includes us.

1 **Did the answer hold six months later?**

Accuracy gets checked once, at the pilot and the sign-off. Ask to see the same question asked months apart, and what came back each time. A vendor who has never run that comparison is telling you consistency has never been measured.

2 **What exactly is versioned: the model, the prompt, or the rules?**

The model is the provider's to change and retire, whatever you have deployed on top of it. The prompt is yours to version, and worth versioning, but it controls the wording of the question rather than how the system behaves. The rules control the outcome, and a change to one shows what changed, when, and who changed it. A vendor who versions only prompts has a record of the wording and none of the decision.

3 **When an answer changes, where is the record of why?**

Audit logs tell you who, what, and when. For a legal decision you also need to know why this answer came out this way, and to be able to reproduce it. If the reasoning lives inside the model, that record does not exist.

4 **Can I inspect the reasoning, or only the output?**

The defensible shape lets the model gather the facts and lets versioned rules make the call. Same inputs, same output, and a path through the logic you can hand to a colleague, a client, or a regulator to check. If all you can inspect is the output, you're taking it on trust.

5 **When it drifts, who is accountable: you or us?**

Watch for the pause. Vendors are quick to claim capability and slow to accept liability for behavior that changes over time. Given where product liability law is heading, the answer to this question will matter more every quarter.

07

How Neota Logic has approached it

Neota Logic has spent more than fifteen years encoding legal reasoning as rules, and the last several adding AI to that foundation. The approach comes down to three commitments.

01

Language to the model

The language model does the work it suits: reading documents, summarizing, and pulling out the facts. Within a workflow, an AI task node calls your own registered models for tasks such as extraction, summarization and classification, then returns the output into the workflow, where your rules, validation logic and human sign-off apply exactly as you have defined them.

02

Decisions to the rules

The conclusion is reached under logic a lawyer wrote and can read, versioned like any other logic, so a change shows exactly what changed, when, and by whom. In one controlled internal test, 50 of 50 scenarios passed 250 of 250 checks. That figure describes a specific test, and it illustrates the property that matters: same inputs, same output, every time.

03

Keep the record

Each AI call is recorded in the session history, with the model used and the tokens in and out, and human approval steps sit at the points where the judgment is. Site administrators register and approve which models are available at all, so credentials stay centralized and only approved models ever reach a workflow. No AI model is invoked by default.

To be clear about the boundary, because the market blurs it: Neota Logic governs the workflow and the answer. It doesn't govern the underlying AI model. The model remains the provider's, versioned and retired on their schedule. What sits under your control, and on your record, is everything that turns a model's output into a legal decision: the rules that decide, the approvals that vouch, and the record that shows how.

SOURCES

1. Blair-Stanek and Van Durme, "LLMs Provide Unstable Answers to Legal Questions," University of Maryland and Johns Hopkins, arXiv:2502.05196, January 2025
2. Chen, Zaharia and Zou, "How Is ChatGPT's Behavior Changing over Time?", Stanford and Berkeley, 2023
3. Thinking Machines, "Defeating Nondeterminism in LLM Inference," 2025
4. Lau, "Same Input, Different Scores: A Multi Model Study on the Inconsistency of LLM Judge," 2026
5. OpenAI, "Introducing Astra for Law," September 17, 2026
6. Rabanser, Kapoor and Narayanan, "Towards a Science of AI Agent Reliability," Princeton, 2026
7. Damien Charlotin, AI Hallucination Cases database
8. ILTA 2026 Technology Survey
9. ACC and Everlaw, "Generative AI's Growing Strategic Value for Corporate Law Departments," 2025
10. Law360 Pulse, "The 2026 AI Survey," March 2026
11. EU Product Liability Directive 2024/2853; FTC final order against DoNotPay, January 2025
12. Artificial Lawyer, "LawFairy 'Technology-Only Law Firm' Gets Regulatory Approval," February 24, 2026

Neota Logic

The test that matters

Every new model release is pitched on capability: more fluent, higher on the benchmark, better at the bar exam. Capability is real and it is welcome. It is also not the test. The market is already recalibrating: in Law360's 2026 survey, 44 percent of attorneys who frequently use AI at firms now say they see both the pros and cons of adoption, where a year earlier 73 percent of frequent users held a positive view.[10] The test is whether the answer you relied on in January is the answer the system gives in July, and whether you could show anyone why.

An LLM gives you a confident answer. A rules engine gives you a defensible one.

Legal AI needs both.

Request a demo