The Reliability Gap

Why Artificial Intelligence Still Cannot Be Trusted with Legal Reasoning

A field analysis of the gap between fluent legal language and reliable legal reasoning, and what it means for everyone who depends on the law being right

THE ARGUMENT

General-purpose AI can now answer legal questions in fluent, confident, and professional prose. That fluency is the danger. In controlled study, general-purpose models have produced hallucination rates between 58 and 88 percent on verifiable questions of law. The failure is not simply that the models are unintelligent, nor that they lack expressive ability. The deeper failure is architectural: they often lack the navigational discipline required to move from a legal question, through the material facts, into the controlling authority, and finally to a legally checkable answer.

Legal reasoning is not merely persuasive language. It requires reasoning that is anchored to real authority, follows the actual procedural path, stays within the doctrine that governs, and commits only as far as the law itself commits. A legal answer may sound correct while relying on the wrong court, the wrong rule, the wrong procedural route, or an unstated factual assumption. In law, fluency is not competence; the answer is only as reliable as the route by which it reaches its conclusion.

This paper diagnoses that gap. It examines who is exposed by legal-sounding but legally unstable AI answers, why the problem is structural rather than cosmetic, and what any

DOCUMENT

Position paper. A field analysis intended for public circulation.

AUTHOR

Ayooluwa Paul Obembe, Esq.
Legal practitioner and legal-AI developer

SCOPE

Diagnostic. Names the problem and the properties a solution must have; prescribes no specific system.

credible legal-education AI system must be able to do before it can be trusted. It does not sell a cure. It defines the standard that a cure would have to meet.

This document was developed by the author through extended iterative direction; large language model tools were used to draft and format the text under that direction. The argument, analysis, conclusions, and verification of all cited sources are the author's own.

Abstract

Artificial intelligence has become fluent in the language of law. It has not become reliable in the reasoning of law. This paper argues that the distance between those two facts is the central unresolved problem in legal AI: legal-sounding answers are now easy to generate, but legally dependable answers remain difficult to produce, verify, and trust.

The argument proceeds in four steps. First, legal reasoning is high-stakes at every level of use: for laypersons trying to understand a dispute, law students forming doctrinal foundations, professional examination candidates learning to apply law to facts, and practising lawyers making decisions with real consequences. Second, current general-purpose AI fails these users in a specific way. It is optimised to produce plausible answers, not necessarily answers routed through the correct legal authority, procedural path, and doctrinal limits. In law, plausibility and correctness diverge precisely where the risks are greatest. Third, the failure compounds as legal reasoning deepens. Systems that perform adequately on shallow retrieval or general explanation often break down when the task requires multi-step legal navigation, procedural sequencing, jurisdictional control, or precise fact-rule application. Retrieval augmentation alone does not solve this problem if the system lacks discipline over what to retrieve, how to rank authority, and how to apply it. Fourth, the missing ingredient is not raw intelligence but navigational discipline: the capacity to move from question, to legally material facts, to verified authority, to a checkable legal conclusion.

This paper is diagnostic. It characterises the gap between legal fluency and legal reliability, identifies the users most exposed by that gap, and proposes the properties any credible legal AI system must satisfy before it can be trusted. It does not endorse a particular architecture or product. Its purpose is to make the problem precise, so that the standard for a real solution becomes clear.

Contents

1. Fluency Is Not Reliability

In the space of a few years, artificial intelligence has become strikingly fluent in the language of law. It drafts clauses, summarises judgments, explains doctrines, and answers legal questions in confident, well-structured prose. To a non-specialist, and increasingly to specialists under time pressure, the output is indistinguishable from competent legal work.

This fluency has been mistaken for reliability. It is not the same thing. Fluency is the capacity to produce text that reads as correct. Reliability is the property of being correct, and of being correct in a way that can be checked. In most domains the gap between the two is tolerable, because an answer that merely sounds right is often close enough, and the cost of being wrong is small. Law is not such a domain. In law, an answer that sounds authoritative but cites a provision that does not exist, or names a court that has no jurisdiction, or asserts a procedural step that the rules do not contain, is not a near-miss. It is a failure, and depending on who relied on it, it can be a costly one.

This paper makes a single argument, developed in stages. The argument is that the central unsolved problem in legal AI is not that the technology lacks intelligence. By any reasonable measure these systems are capable. The problem is that capability has been pointed at the wrong target. Current systems are optimised to be fluent and plausible. Legal reasoning demands something they are not built to deliver: answers that are anchored to verifiable authority, that follow the correct procedural and substantive path, and that commit only as far as the law itself commits. The distance between fluent legal language and reliable legal reasoning is the reliability gap, and closing it is the real work that remains.

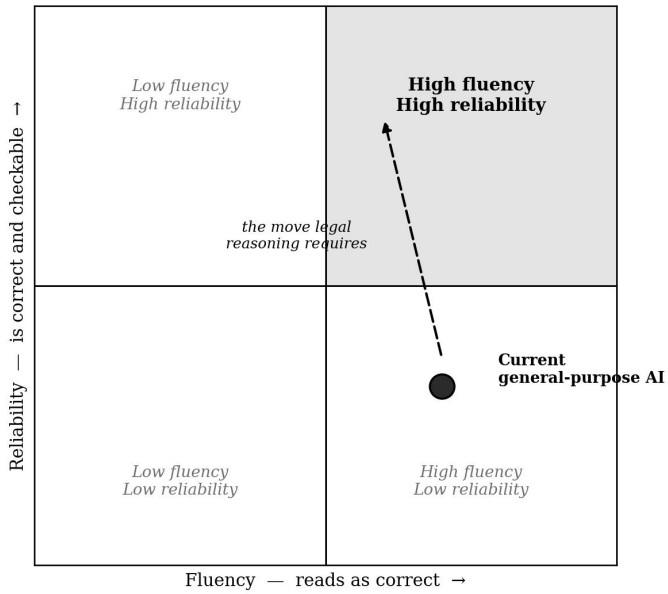

Figure 1. Fluency and reliability are distinct axes. A system can read as entirely correct while being unreliable in the ways that matter. Current general-purpose AI sits in the high-fluency, low-reliability quadrant; legal reasoning requires the high-fluency, high-reliability quadrant, and the distance between the two is the reliability gap.

The paper is deliberately diagnostic. It does not advocate for a particular product, company, or technical architecture. Its ambition is narrower and, in a field crowded with confident claims, more useful: to state the problem precisely enough that the standard for a genuine solution becomes unambiguous.

2. Who Is Exposed: Legal Reasoning Across Four Tiers of User

It is tempting to treat the reliability of legal AI as a narrow professional concern, relevant mainly to lawyers and law firms. That framing understates the problem. Legal reasoning is consumed across a spectrum of users, and the reliability gap exposes every one of them, though it exposes each differently.

At one end is the ordinary person. Legal information has always been expensive and difficult to access, and for many people a general-purpose AI assistant is now the first, and sometimes the only, place they turn to understand a tenancy dispute, a workplace grievance, a family matter, or a question about land. This user has the least capacity to detect an error. A confident, fluent, wrong answer is, to them, simply an answer. They act on it. The Stanford study discussed in Section 3 makes this point directly: the risks of unreliable legal AI fall most heavily not on well-resourced professionals but on those without access to traditional legal resources [1].

Next is the law student, who uses these systems while still building the foundational models of doctrine and method that the rest of a legal career rests on. A fluent tutor that is subtly wrong is, for this user, particularly dangerous, because the error is not caught and discarded; it is learned, and it becomes part of the foundation.

Then the examination candidate, preparing for high-stakes professional qualification, where precision is not a virtue but a graded requirement. For this user, an AI that hedges between two answers where the law commits to one, or that invents an authority, is actively harmful at exactly the moment reliability matters most.

And finally the practising lawyer, operating in live matters under real time and cost pressure. This is the tier where the failure has the most public and most documented consequences, and Section 3 turns to that evidence. The point of this section is prior and simpler: this is not one problem for one group. It is a single structural failure with four distinct populations exposed to it.

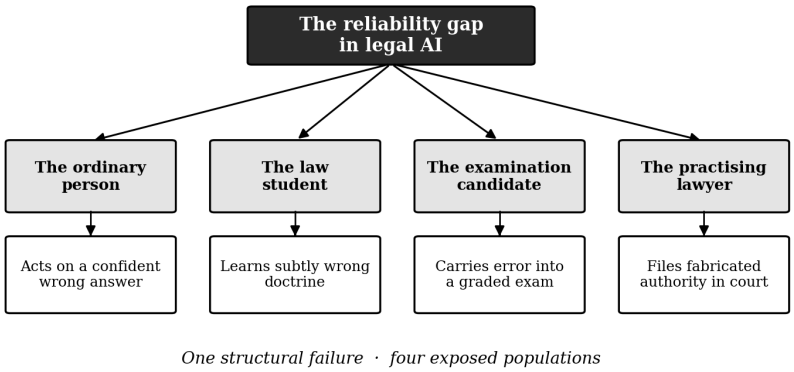

Figure 2. One structural failure, four exposed populations. The reliability gap is not a narrow professional concern; it reaches every tier of user who depends on legal reasoning, exposing each in a different way.

Table 1. The reliability gap across four tiers of user.

User Tier	How They Rely on Legal AI	What a Wrong Answer Costs
The ordinary person	To understand a tenancy dispute, an employment grievance, a family or land matter, often with no affordable alternative.	Acts on a confident but wrong answer, with no way to detect the error, and forecloses a real legal right.

User Tier	How They Rely on Legal AI	What a Wrong Answer Costs
The law student	To build foundational mental models of doctrine, procedure, and legal method while still learning.	Absorbs subtly wrong doctrine from a fluent tutor and hardens the error into a foundation.
The examination candidate	To prepare for high-stakes professional examinations where precision and procedural accuracy are graded directly.	Carries a hedging or fabricated answer into an examination where the wrong court or wrong process is a marked failure.
The practising lawyer	To accelerate research, drafting, and analysis in live matters under time and cost pressure.	Files a fabricated authority in court: professional sanction, client harm, and damage to the integrity of the record.

3. The Evidence That the Gap Is Real

The reliability gap is not a theoretical concern or a matter of opinion. It has been measured in controlled studies and it is being documented, continuously, in the public record of the courts.

3.1 The Measured Failure Rate

In a 2024 study published in the Journal of Legal Analysis, researchers at Stanford profiled legal hallucination in general-purpose language models by posing specific, verifiable questions about real court cases. The result was stark: the models produced hallucinated content between 58 and 88 percent of the time, depending on the model, when asked verifiable questions about random federal cases [1]. The study made two further observations that matter as much as the headline rate. The models often failed to correct a user's incorrect legal assumptions when those assumptions were embedded in the question. And the models could not reliably predict, or did not reliably know, when they were hallucinating [1]. A system that is frequently wrong is a problem; a system that is frequently wrong and cannot tell that it is wrong is a different and deeper one.

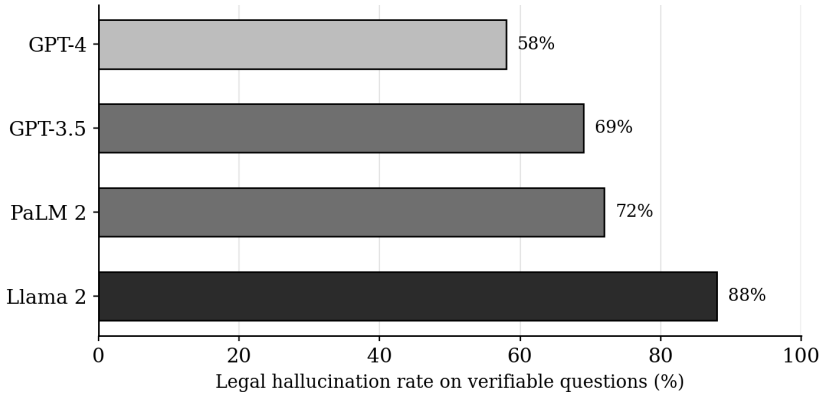

Figure 3. Legal hallucination rates measured on verifiable questions about real federal cases, as reported by Dahl and colleagues [1]. Even the strongest model evaluated hallucinated on a majority of questions.

3.2 The Documented Failure in Practice

Beyond the laboratory, the failure is now visible in the day-to-day record of the courts. A widely cited database maintained by a research fellow at HEC Paris tracks legal decisions in which courts or tribunals have identified AI-generated hallucinations, typically fabricated citations, in filings [2]. The trajectory of that database is itself the evidence. It recorded fewer than a hundred cases in the first half of 2025; by early 2026 it had catalogued well over a thousand cases worldwide, the substantial majority in the courts of a single country, and the count has continued to climb [2].

The consequences attached to these cases are not nominal. Courts have moved from warnings to substantial monetary sanctions, including penalties reported in the tens of thousands of dollars against attorneys who submitted briefs containing fabricated citations and invented quotations [2]. The phenomenon has reached appellate courts, national law firms, and even government legal offices. The pattern is no longer a series of isolated embarrassments; it is a documented, accelerating feature of contemporary legal practice.

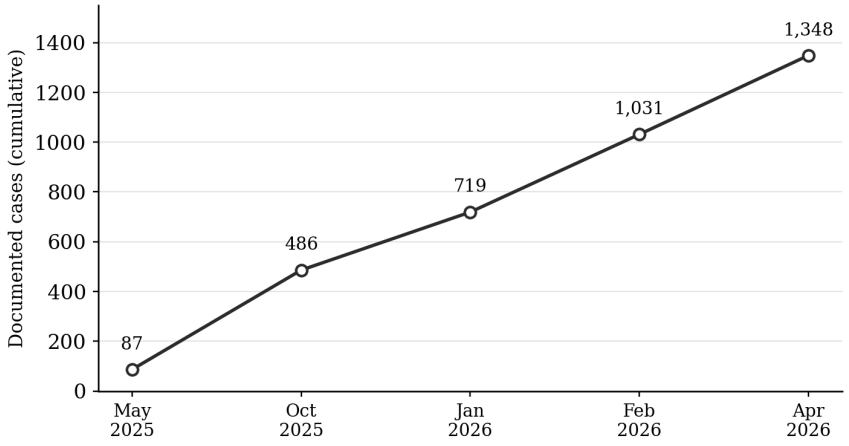

Figure 4. Cumulative count of court decisions identifying AI-generated hallucinated content in legal filings, as recorded in a continuously updated public database [2]. The curve is the trend: a problem that was a handful of cases is now counted in the thousands. Values are indicative of the database state at the dates shown and continue to rise.

Taken together, the controlled evidence and the courtroom record establish the same fact from two directions. Under measurement, general-purpose AI hallucinates legal content at high rates and cannot reliably detect when it does. In practice, that failure is now producing real sanctions,

real harm to clients, and real damage to the integrity of the legal record. The gap is not coming. It is here, and it is being counted.

4. Why the Failure Is Architectural, Not Incidental

It is comforting to assume that a failure rate this high is a temporary defect, the kind of problem the next, larger model will quietly solve. That assumption deserves direct scrutiny, because the evidence points the other way.

General-purpose language models are, at their core, systems that predict the most probable continuation of text. They are trained to produce what is plausible given everything they have seen. This is an extraordinary capability, and in most uses it is exactly what is wanted. But it carries a consequence that is not a bug to be patched: the system optimises for the answer that is most likely, not for the answer that is verifiably correct. In the great majority of domains those two targets overlap closely enough that the distinction does not matter. Law is one of the domains where they come apart, and they come apart precisely at the points of greatest consequence — the unusual fact pattern, the jurisdictional exception, the provision that is well known to specialists but thinly represented in ordinary text.

There is a second, compounding dimension to the failure. Research on the reasoning capabilities of these models has argued that they do not, on their own, perform genuine multi-step planning or reliable self-verification; what they produce is better understood as fluent approximation of reasoning rather than reasoning that computes and checks each step [3]. A consequence follows directly. The longer and more interdependent the chain of reasoning a task requires, the more opportunities there are for the system to substitute a plausible step for a correct one, and errors accumulate rather than cancel.

Legal reasoning is, almost by definition, the worst possible match for that profile. A legal problem is rarely a single retrieval. It is a sequence: identify the legally material facts, locate the controlling authority, apply that authority to those facts, follow the procedural consequences, and arrive at a position. This is exactly the multi-step, interdependent reasoning on which the failure compounds. A benchmark study built from hundreds of real law-school examinations found precisely this pattern: models could handle shallow informational questions, but their performance

fell away on open-ended questions that required structured, multi-step legal reasoning [4]. The failure is not random. It is concentrated exactly where law actually lives.

The conclusion is uncomfortable but important. The reliability gap is not an incidental defect that scale will erase. It is a consequence of what these systems are and how they work. A system optimised for plausibility, and weak at verified multi-step reasoning, will produce its most confident errors in exactly the high-stakes, multi-step situations that legal reasoning consists of. That is an architectural mismatch, and architectural mismatches are not closed by training runs.

5. Why Retrieval Did Not Close the Gap

The most widely promoted response to the reliability problem has been retrieval-augmented generation: rather than relying solely on what a model has absorbed into its parameters, the system retrieves relevant source documents and generates its answer with those documents in view. The intuition is sound. If the model can see the actual statute or the actual case, surely it will stop inventing them.

The intuition has not survived contact with evidence. In a 2025 study published in the Journal of Empirical Legal Studies, Stanford researchers examined legal research tools built on retrieval-augmented generation, including tools marketed by their providers as eliminating hallucination or guaranteeing hallucination-free citations. The study found that hallucination persisted even in these retrieval-augmented, legal-specific systems [5]. Retrieval reduced the problem in some respects; it did not eliminate it, and the marketing claim of elimination was not borne out.

The reason is instructive, and it matters for understanding what a real solution requires. Retrieval does not remove the probabilistic character of the system; it relocates it. The retrieval step is itself typically driven by similarity — it finds text that resembles the query — and similarity is not the same as legal authority. A retrieval system can surface a passage that is topically similar but legally beside the point, that belongs to the wrong jurisdiction, or that is no longer good law. The generation step then proceeds, fluently, over that imperfect material. The result can be an answer that is now confidently wrong and apparently sourced — in some respects a more dangerous failure than an obvious invention, because it carries the appearance of grounding.

The lesson is not that retrieval is worthless. It is that retrieval, by itself, is not the answer, because retrieval that is not legally disciplined simply moves the point of failure rather than

removing it. A system that retrieves the plausible-looking passage instead of the controlling authority has not solved the reliability problem. It has relocated it one step upstream and hidden it better.

6. The Missing Ingredient: Navigational Discipline

If the problem is not simply a shortage of model intelligence, and if retrieval alone does not cure it, then the missing ingredient must be stated precisely. Legal reasoning is not the production of plausible legal language. It is a disciplined movement through a constrained system of authority, facts, procedure, and consequence.

A competent legal answer does not begin with fluency. It begins with orientation. The lawyer must understand what the question is actually asking, distinguish the legally material facts from the background facts, identify the authority that controls the issue rather than merely resembles it, and apply that authority within the proper doctrinal and procedural frame. A question about employment, land, or evidence may contain many facts, but only some of them carry legal consequence for the task at hand. The answer is trustworthy only when the right facts are connected to the controlling authority in the correct order.

This is what this paper calls navigational discipline: the capacity to move through a legal problem along a verifiable path — from the question to the material facts, from the material facts to the controlling authority, from authority to legal consequence, and from consequence to a conclusion that can be checked. It includes a sound sense of when hierarchy, jurisdiction, or procedure is decisive, when a broad principle must yield to a more specific rule, and when the available materials simply do not justify a confident answer.

Current general-purpose systems often lack this discipline. Their failure is not usually visible as crude error. More often it appears as a fluent answer that has taken the wrong route: the wrong authority, the wrong court, the wrong procedural posture, the wrong level of generality, or an unstated factual assumption. Such an answer can read professionally while remaining legally unsafe. The danger, in other words, is not that these systems cannot speak law. It is that they can speak law persuasively without reliably travelling through it.

This reframing matters because it changes what counts as progress. If the problem were merely inadequate intelligence, the answer would be to wait for larger and more capable models. But if

the problem is the absence of navigational discipline, greater fluency may deepen the risk rather than reduce it. A more capable but undisciplined model produces errors that are harder to detect, because they arrive with greater confidence, coherence, and legal style. The reliability gap will not be closed by intelligence alone. It will be closed, if it is closed, by systems that impose discipline on the reasoning process itself.

There is one further consequence, and it is worth stating plainly because it runs against the grain of what users have come to expect. A system with genuine navigational discipline will sometimes do what fluent systems almost never do: it will decline to commit. It will report that the authority is genuinely unsettled, or that the question cannot be answered reliably from the available materials. To a user accustomed to a confident answer every time, that restraint can look like a weaker product. It is the opposite. A system that knows the boundary of what it can stand behind, and respects it, is the only kind of system that can be trusted in a domain where being wrong has consequences. In law, that restraint is not a limitation. It is the beginning of trust.

7. The Properties a Credible Solution Must Have

This paper does not prescribe a particular system, architecture, or product. It does, however, hold that the diagnosis above implies a clear set of properties that any credible solution to the reliability gap must satisfy. These properties are not a design. They are a standard — a way for a user, a buyer, a court, or an investor to ask the right questions of any tool that claims to do legal reasoning.

Table 2. Properties a credible legal-reasoning system must satisfy.

Property	What It Requires
Authority grounding	Every legal proposition in an answer is tied to a specific, verifiable source, and the system can show where it came from. Plausibility is never accepted as a substitute for citation.
Procedural fidelity	The answer follows the correct procedural and substantive path in the correct order, rather than assembling a plausible-sounding route that the rules do not actually contain.
Verified retrieval	Where the system retrieves source material, retrieval is judged by whether it captures the controlling authority, not merely by topical similarity.
Commitment discipline	The system commits where the law is settled and discloses uncertainty where it genuinely is uncertain, rather than hedging where the law is clear or inventing certainty where it is not.

Property	What It Requires
Honest boundaries	The system recognises the limit of what it can reliably answer and declines to cross it, rather than producing a confident answer in every case regardless of competence.
Traceability	A user, or a court, can follow the answer back through its reasoning and its authorities and independently check each step.

The common thread is that every property describes discipline rather than raw capability. None of them asks whether the system is intelligent, fluent, or fast. Each asks whether the system is anchored, verifiable, ordered, honest about its limits, and traceable. That is the correct test, because those are the properties the reliability gap is made of, and a system that does not satisfy them has not closed the gap regardless of how capable or fluent it appears.

A reader evaluating any legal-AI tool can therefore set aside the question that marketing tends to foreground — how advanced is the underlying model — and ask the questions that actually matter. Can it show me the authority for every proposition. Does it follow the real procedural path. Does it commit only where the law commits. Does it know what it does not know. Can I trace and check its reasoning. A tool that cannot answer those questions is, whatever its fluency, still on the wrong side of the reliability gap.

8. Conclusion

The reliability gap in legal AI is real, it is measurable, and it is not closing on its own. It has been measured in controlled study and it is being counted, case by case, in the public record of the courts. And it is architectural: a system optimised for plausibility, and weak at verified multi-step reasoning, will fail hardest in exactly the high-stakes, multi-step situations that legal reasoning consists of. Retrieval, the most promoted remedy, relocates that failure rather than removing it.

The error in the field has been to treat this as a problem of intelligence, to be outgrown. It is better understood as a problem of discipline. What legal reasoning demands, and what current systems lack, is navigational discipline: the capacity to route a question through verified authority, along the correct path, to an answer that commits only as far as the law commits and that can be traced and checked. A more capable model without that discipline will produce a more convincing version of the same failure.

This paper has not offered a finished solution, and it has been deliberate in not doing so. Its purpose has been to make the problem precise: to show who is exposed, to show that the gap is documented rather than speculative, to explain why it is architectural, and to set out the properties any credible solution must satisfy. The next contribution to this question should not be a louder claim of fluency. It should be a system that can be measured against the standard set out here — and a measurement that is honest about the result. Until a tool can show its authority, follow the real path, commit only as far as the law does, know its own limits, and be traced and checked, the gap between fluent legal language and reliable legal reasoning remains open, and everyone who depends on the law being right remains exposed to it.

References

[1] Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis, 16(1), 64–93.

[2] Charlotin, D. AI Hallucination Cases Database. Smart Law Hub, HEC Paris. A continuously updated record of court decisions addressing AI-generated hallucinated content in legal filings. Available at damiencharlotin.com/hallucinations.

[3] Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L., and Murthy, A. (2024). Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 22895–22907.

[4] Fan, Y., Ni, J., Merane, J., and others (2025). LEXam: Benchmarking Legal Reasoning on 340 Law Exams. arXiv:2505.12864.

[5] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 22(2), 216–242.