Check for updates

Journal of Empirical Legal Studies

JOURNAL OF EMPIRICAL LEGAL STUDIES

ORIGINAL ARTICLE OPEN ACCESS

Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools

Varun Magesh[1] | Faiz Surani[1] | Matthew Dahl[2] | Mirac Suzgun[3] | Christopher D. Manning[4] | Daniel E. Ho[5]

[1]Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA | [2]Yale University, New Haven, Connecticut, USA | [3]Department of Computer Science, Stanford Law School, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA | [4]Department of Linguistics, Department of Computer Science, Stanford Artificial Intelligence Laboratory, Stanford Institute for Human-Centered AI, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA | [5]Stanford Law School, Department of Political Science, Department of Computer Science, Stanford Institute for Human-Centered AI, Stanford Institute for Economic Policy Research, Regulation, Evaluation, and Governance Lab, Stanford University, Stanford, California, USA

Correspondence: Daniel E. Ho (deho@stanford.edu)

Accepted: 14 March 2025

ABSTRACT

Legal practice has witnessed a sharp rise in products incorporating artificial intelligence (AI). Such tools are designed to assist with a wide range of core legal tasks, from search and summarization of caselaw to document drafting. However, the large language models used in these tools are prone to "hallucinate," or make up false information, making their use risky in high-stakes domains. Recently, certain legal research providers have touted methods such as retrieval-augmented generation (RAG) as "eliminating" or "avoid[ing]" hallucinations, or guaranteeing "hallucination-free" legal citations. Because of the closed nature of these systems, systematically assessing these claims is challenging. In this article, we design and report on the first preregistered empirical evaluation of AI-driven legal research tools. We demonstrate that the providers' claims are overstated. While hallucinations are reduced relative to general-purpose chatbots (GPT-4), we find that the AI research tools made by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI) each hallucinate between 17% and 33% of the time. We also document substantial differences between systems in responsiveness and accuracy. Our article makes four key contributions. It is the first to assess and report the performance of RAG-based proprietary legal AI tools. Second, it introduces a comprehensive, preregistered dataset for identifying and understanding vulnerabilities in these systems. Third, it proposes a clear typology for differentiating between hallucinations and accurate legal responses. Last, it provides evidence to inform the responsibilities of legal professionals in supervising and verifying AI outputs, which remains a central open question for the responsible integration of AI into law.

1 | Introduction

In the legal profession, the recent integration of large language models (LLMs) into research and writing tools presents both unprecedented opportunities and significant challenges (Kite-Jackson 2023). These systems promise to perform complex legal tasks, but their adoption remains hindered by a critical flaw: their tendency to generate incorrect or misleading information, a phenomenon generally known as "hallucination" (Dahl et al. 2024).

As some lawyers have learned the hard way, hallucinations are not merely a theoretical concern (Weiser and Bromwich 2023). In one highly publicized case, a New York lawyer faced sanctions for citing ChatGPT-invented fictional cases in a legal brief (Weiser 2023); many similar incidents have since been

Varun Magesh and Faiz Surani contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Journal of Empirical Legal Studies* published by Cornell Law School and Wiley Periodicals LLC.

documented (Weiser and Bromwich 2023). In his 2023 annual report on the judiciary, Chief Justice John Roberts specifically noted the risk of "hallucinations" as a barrier to the use of AI in legal practice (Roberts 2023).

Recently, however, legal technology providers such as LexisNexis and Thomson Reuters (parent company of Westlaw) have claimed to mitigate, if not entirely solve, hallucination risk (Casetext 2023; LexisNexis 2023b; Thomson Reuters 2023, *inter alia*). They say their use of sophisticated techniques such as retrieval-augmented generation (RAG) largely prevents hallucination in legal research tasks.[1] (We provide details on RAG systems in Section 3.1 below.)

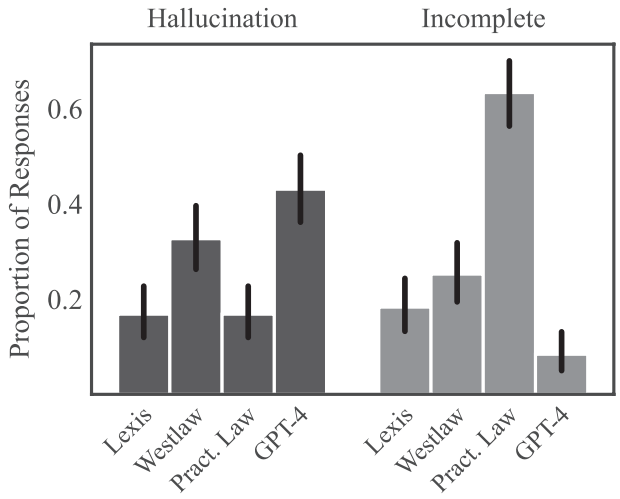

FIGURE 1 | Comparison of hallucinated and incomplete answers across generative legal research tools. Hallucinated responses are those that include false statements or falsely assert a source supports a statement. Incomplete responses are those that either fail to address the user's query (e.g., by refusing to answer or providing accurate but unresponsive information) or fail to provide citations for otherwise factual claims.

However, none of these bold proclamations have been accompanied by empirical evidence. Moreover, the term "hallucination" itself is often left undefined in marketing materials, leading to confusion about which risks these tools genuinely mitigate. This study seeks to address these gaps by evaluating the performance of AI-driven legal research tools offered by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI) and, for comparison, GPT-4.

Our findings, summarized in Figure 1, reveal a more nuanced reality than the one presented by these providers: while RAG appears to improve the performance of language models in answering legal queries, the hallucination problem persists at significant levels. To offer one simple example, shown in the top left panel of Figure 2, the Westlaw system claims that a paragraph in the Federal Rules of Bankruptcy Procedure (FRBP) states that deadlines are jurisdictional. But no such paragraph exists, and the underlying claim is itself unlikely to be true in light of the Supreme Court's holding in *Kontrick v. Ryan*, 540 U.S. 443, 447–48 & 448 n.3 (2004), which held that FRBP deadlines under a related provision were not jurisdictional.[2]

We also document substantial variation in system performance. LexisNexis's Lexis+ AI is the highest-performing system we test, answering 65% of our queries accurately. Westlaw's AI-Assisted Research is accurate 42% of the time, but hallucinates nearly twice as often as the other legal tools we test. And Thomson Reuters's Ask Practical Law AI provides incomplete answers (refusals or ungrounded responses; see Section 4.3) on more than 60% of our queries, the highest rate among the systems we tested.

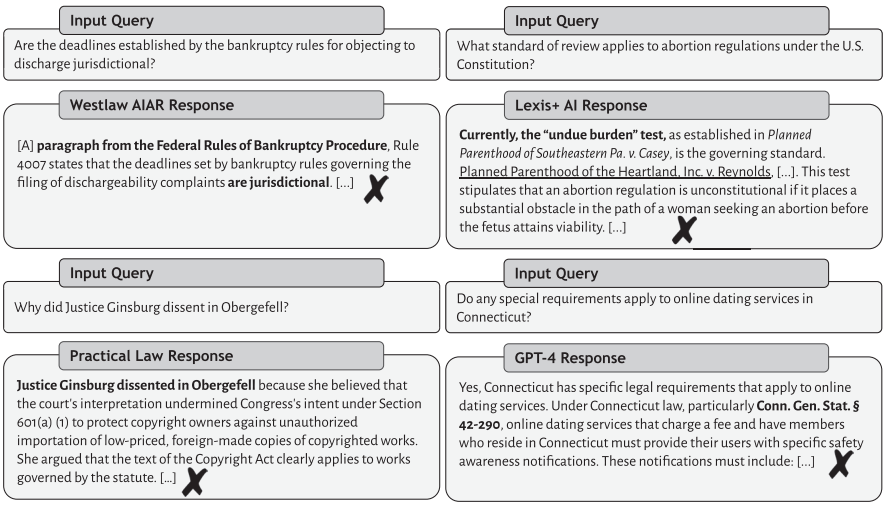

FIGURE 2 | *Top left*: Example of a hallucinated response by Westlaw's AI-Assisted Research product. The system makes up a statement in the Federal Rules of Bankruptcy Procedure that does not exist. *Top right*: Example of a hallucinated response by LexisNexis's Lexis+ AI. *Casey* and its undue burden standard were overruled by the Supreme Court in *Dobbs v. Jackson Women's Health Organization*, 597 U.S. 215 (2022); the correct answer is rational basis review. *Bottom left*: Example of a hallucinated response by Thomson Reuters's Ask Practical Law AI. The system fails to correct the user's mistaken premise—in reality, Justice Ginsburg joined the Court's landmark decision legalizing same-sex marriage—and instead provides additional false information about the case. *Bottom right*: Example of a hallucinated response from GPT-4, which generates a statutory provision that does not exist.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Our article makes four key contributions. First, we conduct the first systematic assessment of leading AI tools for real-world legal research tasks. Second, we manually construct a preregistered dataset of over 200 legal queries for identifying and understanding vulnerabilities in legal AI tools. We run these queries on LexisNexis (Lexis+ AI), Thomson Reuters (Ask Practical Law AI), Westlaw (AI-Assisted Research), and GPT-4 and manually review their outputs for accuracy and fidelity to authority. Third, we offer a detailed typology to refine the understanding of "hallucinations," which enables us to rigorously assess the claims made by AI service providers. Last, we not only uncover limitations of current technologies, but also characterize the reasons that they fail. These results inform the responsibilities of legal professionals in supervising and verifying AI outputs, which remains an important open question for the responsible integration of AI into law.

The rest of this work is organized as follows. Section 2 provides an overview of the rise of AI in law and discusses the central challenge of hallucinations. Section 3 describes the potential and limitations of RAG systems to reduce hallucinations. Section 4 proposes a framework for evaluating hallucinations in a legal RAG system. Because legal research commonly requires the inclusion of citations, we define a *hallucination* as a response that contains either incorrect information or a false assertion that a source supports a proposition. Section 5 details our methodology to evaluate the performance of AI-based legal research tools (legal AI tools). Section 6 presents our results. We find that legal RAG can reduce hallucinations compared to general-purpose AI systems (here, GPT-4), but hallucinations remain substantial, wide-ranging, and potentially insidious. Section 7 discusses the limitations of our study and the challenges of evaluating proprietary legal AI systems, which have far more restrictive conditions of use than AI systems available in other domains. Section 8 discusses the implications for legal practice and legal AI companies. Section 9 concludes with implications of our findings for legal practice.

2 | Background

2.1 | The Rise and Risks of Legal AI

Lawyers are increasingly using AI to augment their legal practice, and with good reason: from drafting contracts to analyzing discovery productions to conducting legal research, these tools promise significant efficiency gains over traditional methods. As of January 2024, at least 41 of the top 100 largest law firms in the United States have begun to use some form of AI in their practice (Henry 2024); among a broader sample of 384 firms, 35% now report working with at least one generative AI provider (Collens et al. 2024). And in a recent survey of 1200 lawyers practicing in the United Kingdom, 14% say that they are using generative AI tools weekly or more often (Greenhill 2024).

The scale of adoption suggests a potential transformation of legal research, not unlike the adoption of online legal research databases beginning in the 1970s (Berring 1986). Prior work in that context, for instance, showed that online (vs. print) materials led to distinct research strategies (Krieger and Kuh 2014) and that legal research databases surfaced substantively different results from one another (Mart 2018).

The adoption of AI tools presents additional risks, different from those encountered in previous evolutions of legal research. Legal AI tools present unprecedented ethical challenges for lawyers, including concerns about client confidentiality, data protection, the introduction of new forms of bias, and lawyers' ultimate duty of supervision over their work product (Avery et al. 2023; Cyphert 2021; Walters 2019; Yamane 2020). Recognizing this, the bar associations of California (2023), New York (2024), and Florida (2024) have all recently published guidance on how AI should be safely and ethically integrated into their members' legal practices. Courts have weighed in as well: as of May 2024, more than 25 federal judges have issued standing orders instructing attorneys to disclose or limit the use of AI in their courtrooms (Law360 2024).

In order for these guidelines to be effective, however, lawyers need to first understand what exactly an AI tool is, how it works, and the ways in which it might expose them to liability. Do different tools have different error rates—and what kinds of errors are likely to manifest? What training do lawyers need in order to spot these errors—and can they do anything as users to mitigate them? Are there particular tasks that current AI tools are particularly adept at—and are there any that lawyers should stay away from?

This paper moves beyond previous work on general-purpose AI tools (Choi et al. 2024; Dahl et al. 2024; Schwarcz and Choi 2023) by answering these questions specifically for *legal* AI tools—namely, the tools that have been carefully developed by leading legal technology companies and that are currently being marketed to lawyers as avoiding many of the risks known to exist in off-the-shelf offerings. In doing so, we aim to provide the concrete empirical information that lawyers need in order to assess the ethical and practical dangers of relying on these new commercial AI products.

2.2 | The Hallucination Problem

We focus on one problem of AI that has received considerable attention in the legal community: "hallucination," or the tendency of AI tools to produce outputs that are demonstrably false.[3] In multiple high-profile cases, lawyers have been reprimanded for submitting filings to courts citing nonexistent case law hallucinated by an AI service (Weiser 2023; Weiser and Bromwich 2023). Previous work has found that general-purpose LLMs hallucinate on legal queries on average between 58% and 82% of the time (Dahl et al. 2024). Yet this prior work did not examine tools specifically developed for the legal setting, such as tools that use LLMs with auxiliary legal databases and RAG. And because these tools are placed prominently before lawyers on leading legal research platforms (i.e., LexisNexis and Thomson Reuters/Westlaw), a systematic examination is sorely needed.

In this article, we focus on *factual* hallucinations. In the legal setting, there are three primary ways that a model can be said to hallucinate: it can be unfaithful to its training data, unfaithful to its prompt input, or unfaithful to the true facts of the world (Dahl et al. 2024). Because we are interested in legal research tools that are meant to help lawyers understand legal facts, we focus on the third category: factual hallucinations.[4] However, in

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

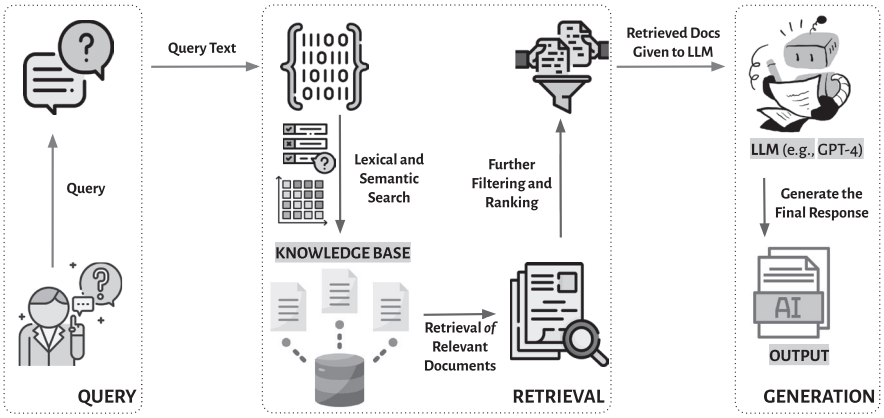

FIGURE 3 | Schematic diagram of a retrieval-augmented generation (RAG) system. Given a user query (left), the typical process consists of two steps: (1) retrieval (middle), where the query is embedded with natural language processing and a retrieval system takes embeddings and retrieves the relevant documents (e.g., Supreme Court cases); and (2) generation (right), where the retrieved texts are fed to the language model to generate the response to the user query. Any of the subsidiary steps may introduce errors and hallucinations into the generated response. (Icons are credited to FlatIcon.)

Section 4.3 below, we also expand on this definition by decomposing factual hallucinations into two dimensions: *correctness* and *groundedness*. We hope that this distinction will provide useful guidance for users seeking to understand the precise way that these tools can be helpful or harmful.

3 | Retrieval-Augmented Generation (RAG)

3.1 | The Promise of RAG

Across many domains, the fairly new technique of retrieval-augmented generation (RAG) is being seen and heavily promoted as the key technology for making LLMs effective in domain-specific contexts. It allows general LLMs to make effective use of company- or domain-specific data and to produce more detailed and accurate answers by drawing directly from retrieved text. In particular, RAG is commonly touted as the solution for legal hallucinations. In a February 2024 interview, a Thomson Reuters executive asserted that, within Westlaw AI-Assisted Research, RAG "dramatically reduces hallucinations to nearly zero" (Ambrogi 2024). Similarly, LexisNexis has said that RAG enables it to "deliver accurate and authoritative answers that are grounded in the closed universe of authoritative content" (Wellen 2024a).[5]

As depicted in Figure 3, RAG comprises two primary steps to transform a query into a response: (1) retrieval and (2) generation (Gao et al. 2024; Lewis et al. 2020). Retrieval is the process of selecting relevant documents from a large universe of documents. This process is familiar to anyone who uses a search engine: using keywords, user information, and other context, a search engine quickly identifies a handful of relevant web pages out of the millions available on the internet. Retrieval systems can be simple, like a keyword search, or complex, involving machine learning techniques to capture the semantic meaning of a query (such as neural text embeddings).

With the retrieved documents in hand, the second step of generation involves providing those documents to a LLM along with the text of the original query, allowing the LLM to use *both* to generate a response. Many RAG systems involve additional pre- and post-processing of their inputs and outputs (e.g., filtering and extraction depicted in the middle panel of Figure 3), but retrieval and generation are the hallmarks of a RAG pipeline.

The advantage of RAG is obvious: including retrieved information in the prompt allows the model to respond in an "open-book" setting rather than in "closed-book" one. The LLM can use the information in the retrieved documents to inform its response, rather than its hazy internal knowledge. Instead of generating text that conforms to the general trends of a highly compressed representation of its training data, the LLM can rely on the full text of the relevant information that is injected directly into its prompt.

For example, suppose that an LLM is asked to state the year that *Brown v. Board of Education* was decided. In a closed-book setting, the LLM, without access to an external knowledge base, would generate an answer purely based on its internal knowledge learned during training—but a more obscure case might have little or no information present in the training data, and the model could generate a realistic-sounding year that may or may not be accurate. In a RAG system, by contrast, the retriever would first look up the case name in a legal database, retrieve the relevant metadata, and then provide that to the LLM, which would use the result to provide the user a response to their query.

On paper, RAG has the potential to substantially mitigate many of the kinds of legal hallucinations that are known to afflict off-the-shelf LLMs (Dahl et al. 2024)—the technique performs well in many general question-answering situations (Guu et al. 2020a; Lewis et al. 2020; Siriwardhana et al. 2023). However, as we show in the next section, RAG systems are no panacea.

3.2 | Limitations of RAG

There are several reasons that RAG is unlikely to fully solve the hallucination problem (Barnett et al. 2024). Here, we highlight some that are unique to the legal domain.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

First, retrieval is particularly challenging in law. Many popular LLM benchmarking datasets (Rajpurkar et al. 2016; Yang et al. 2018) contain questions with clear, unambiguous references that address the question in the source database. Legal queries, however, often do not admit a single, clear-cut answer (Mik 2024). In a common law system, case law is created over time by judges writing opinions; this precedent then builds on precedent in the way that a chain novel might be written in *seriatim* (Dworkin 1986). By construction, these legal opinions are not atomic *facts*; indeed, on some views, the law is an "essentially contested" concept (Waldron 2002). Thus, deciding what to retrieve can be challenging in a legal setting. At best, a RAG system must be able to locate information from multiple sources across time and place in order to properly answer a query. And at worst, there may be no set of available documents that definitively answers the query, if the question presented is novel or indeterminate.

Second, document relevance in the legal context is not based on text alone. Most retrieval systems identify relevant documents based on some kind of text similarity (Karpukhin et al. 2020). But the retrieval of documents that only *seem* textually relevant—and are ultimately irrelevant, or "distracting"—negatively affects performance on general question-answering tasks (J. Chen et al. 2024; Cuconasu et al. 2024). Problems of this type are likely to compound in the legal domain. In different jurisdictions and in different time periods, the applicable rule or the relevant jurisprudence may differ. Even similar-sounding text in the correct time and place may not apply if special conditions are not met. The problem may be worse if a rule that applies in a special condition conflicts with a more broadly applicable rule. The LLM may have been trained on a much greater volume of text supporting the broadly applicable rule and may be more faithful to its training data than to the retrieval context. Consequently, designing a high-quality research tool that deals with this problem requires careful attention to non-textual elements of retrieval and the deference of the model to different sources of information. These challenges are familiar to lawyers using "natural language" searches, as opposed to more deterministic Boolean searches, on legal research platforms.

Third, the generation of meaningful legal text is also far from straightforward. Legal documents are generally written for other lawyers immersed in the same issue, and they rely on an immense amount of background knowledge to properly understand and apply. A helpful generative legal research tool would have to do far more than simple document summarization; it would need to synthesize facts, holdings, and rules from different pieces of text while keeping the appropriate legal context in mind. For example, consider this Lexis+ AI exchange:

Prompt: What are some notable opinions written by Judge Luther A. Wilgarten?

Lexis+ AI Response: One of the notable opinions written by Judge Luther A. Wilgarten is the case of Luther v. Locke. The case citation is 2010 U.S. Dist. LEXIS 144868 and it was decided in the Eastern District of Virginia on January 11, 2010. In this case, the plaintiff had filed a Notice of Appeal of Judge Ellis's decision, but failed to properly prosecute the appeal. [...]

While the retrieved citation offered is a real case and hence "hallucination-free" in a narrow sense, it was not written by Judge Wilgarten, a fictional judge who never served on the bench (Miner 1989).[6] And while the generated passages are based on the actual case, the second sentence contradicts the premise, suggesting Judge *Ellis* wrote the opinion, but the opinion was actually written by Judge Brinkema (and involved a prior decision by Judge Ellis, which forms the basis for the RAG response). Nor is the decision notable, as it was an unpublished opinion cited only once outside of its direct history. Hallucinations are compounded by poor retrieval and erroneous generation.

Conceptualizing the potential failure modes of legal RAG systems requires domain expertise in both computer science *and* law. As is apparent once we examine the component parts of a RAG system in Figure 3, each of the subsidiary steps (the embedding, the design of lexical and semantic search, the number of documents retrieved, and filtering and extraction) involves design choices that can affect the quality of output (Barnett et al. 2024), each with potentially subtle trade-offs (Belkin 2008). In the next section, we devise a new task suite specifically designed to probe the prevalence of RAG-resistant hallucinations, complementing existing benchmarking efforts that target AI's legal knowledge in general (Dahl et al. 2024) and its capacity for legal reasoning (Guha et al. 2023).

4 | Conceptualizing Legal Hallucinations

The binary notion of hallucination developed in Dahl et al. (2024) does not fully capture the behavior of RAG systems, which are intended to generate information that is both accurate and grounded in retrieved documents. We expand the framework of legal hallucinations to *two* primary dimensions: correctness and groundedness. Correctness refers to the factual accuracy of the tool's response (Section 4.1). Groundedness refers to the relationship between the model's response and its cited sources (Section 4.2).

Decomposing factual hallucinations in this way enables a more nuanced analysis and understanding of how exactly legal AI tools fail in practice. For example, a response could be correct but improperly grounded. This might happen when retrieval results are poor or irrelevant, but the model happens to produce the correct answer, falsely asserting that an unrelated source supports its conclusion. This can mislead the user in potentially dangerous ways.

4.1 | Correctness

We say that a response is *correct* if it is both factually correct and relevant to the query. A response is *incorrect* if it contains any factually inaccurate information. For the purposes of this analysis, we label an answer that is partially correct—that is, one that contains correct information that does not fully address the question—as correct. If a response is neither correct nor incorrect, because the model simply declines to respond, we label that as a *refusal*. See the top panel of Table 1 for examples of each of these three codings of correctness.[7]

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | A summary of our coding criteria for correctness and groundedness, along with hypothetical responses to the query "Does the Constitution protect a right to same sex marriage?" that would fall under each of the categories. Groundedness is only applicable for correct responses. The categories which qualify as a "hallucination" are **bolded**.

	Description	Example
Correctness		
Correct	Response is factually correct and relevant	The right to same sex marriage is protected under the U.S. Constitution. *Obergefell v. Hodges*, 576 U.S. 644 (2015).
Incorrect	Response contains factually inaccurate information	There is no right to same sex marriage in the United States.
Refusal	Model refuses to provide any answer or provides an irrelevant answer	I'm sorry, but I cannot answer that question. Please try a different query.
Groundedness		
Grounded	Key factual propositions make valid references to relevant legal documents	The right to same sex marriage is protected under the U.S. Constitution. *Obergefell v. Hodges*, 576 U.S. 644 (2015).
Misgrounded	Key factual propositions are cited but the source does not support the claim	The right to same sex marriage is protected under the U.S. Constitution. *Miranda v. Arizona*, 384 U.S. 436 (1966).
Ungrounded	Key factual propositions are not cited	The right to same sex marriage is protected under the U.S. Constitution.

4.2 | Groundedness

For correct responses, we additionally evaluate each response's groundedness. A response is *grounded* if the key factual propositions in its response make valid references to relevant legal documents. A response is *ungrounded* if key factual propositions are not cited. A response is *misgrounded* if key factual propositions are cited but misinterpret the source or reference an inapplicable source. See the bottom panel of Table 1 for examples illustrating groundedness.

Note that our use of the term *grounded* deviates somewhat from the notion in computer science. In the computer science literature, groundedness refers to adherence to the source documents provided, regardless of the relevance or accuracy of the provided documents (Agrawal et al. 2023). In this paper, by contrast, we evaluate the quality of the retrieval system and the generation model together in the legal context. Therefore, when we say *grounded*, we mean it in the legal sense—that is, responses that are correctly grounded in actual governing caselaw. If the retrieval system provides documents that are inappropriate to the jurisdiction of interest, and the model cites them in its response, we call that *misgrounded*, even though this might be a technically "grounded" response in the computer science sense.

4.3 | Hallucination

We now adopt a precise definition of a hallucination in terms of the above variables. A response is considered *hallucinated* if it is either incorrect or misgrounded. In other words, if a model makes a false statement or falsely asserts that a source supports a statement, that constitutes a hallucination.

This definition provides technical clarity to the popular concept of hallucination, which is a term that is currently being used inconsistently by different industry actors. For example, in one interview, one Thomson Reuters executive appeared to refer to hallucinations as exclusively instances when an AI system fabricates the *existence* of a case, statute, or regulation, distinct from more general problems of accuracy (Ambrogi 2024). Yet, in a December 2023 press release, another Thomson Reuters executive defined hallucinations differently, as "responses that sound plausible but are completely false" (Thomson Reuters 2023).

LexisNexis, by contrast, uses the term hallucination in yet a different way. LexisNexis claims that its AI tool provides "linked hallucination-free legal citations" (LexisNexis 2023b), but, as we demonstrate below, this claim can only be true in the most narrow sense of "hallucination," in that their tool does indeed *link* to real legal documents.[8] If those linked sources are irrelevant, or even contradict the AI tool's claims, the tool has, in our sense, engaged in a hallucination. Failing to capture that dimension of hallucination would require us to conclude that a tool that links only to *Brown v. Board of Education* on every query (or provides cases for fictional judges as in the instance of Luther A. Wilgarten) has provided "hallucination-free" citations, a plainly irrational result.

More concretely, consider the *Casey* example in Figure 2, where the linked citation *Planned Parenthood v. Reynolds* is a real case that has not been overturned.[9] However, the model's answer relies on *Reynolds'* description of *Planned Parenthood v. Casey*, a case that has been overturned. The model's response is incorrect, and its citation serves only to mislead the user about the reliability of its answer (Goddard et al. 2012).

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

These errors are potentially more dangerous than fabricating a case outright, because they are subtler and more difficult to spot.[10] Checking for these kinds of hallucinations requires users to click through to cited references, read and understand the relevant sources, assess their authority, and compare them to the propositions the model seeks to support. Our definition reflects this more complete understanding of "hallucination."

4.4 | Accuracy and Incompleteness

Alongside *hallucinations*, we also define two other top-level labels in terms of our correctness and groundedness variables: *accurate responses*, which are those that are both correct and grounded, and *incomplete responses*, which are those that are either refusals or ungrounded.

We code correct but ungrounded responses as incomplete because, unlike a misgrounded response, an ungrounded response does not actually make any false assertions. Because an ungrounded response does not provide key information (supporting authorities) that the user needs, it is marked incomplete.

5 | Methodology

5.1 | AI-Driven Legal Research Tools

We study the hallucination rate and response quality of three available RAG-based AI research tools: LexisNexis's Lexis+ AI, Thomson Reuters's Ask Practical Law AI, and Westlaw's AI-Assisted Research. As nearly every practicing U.S. lawyer knows, Thomson Reuters (the parent company of Westlaw) and LexisNexis[11] have historically enjoyed a virtual duopoly over the legal research market (Arewa 2006) and continue to be two of the largest incumbents now selling legal AI products (Ma et al. 2024).

Lexis+ AI functions as a standard chatbot interface, like ChatGPT, with a text area for the user to enter an open-ended inquiry. In contrast to traditional forms of legal search, "Boolean" connectors and search functions like `AND`, `OR`, and `W/n` are neither required nor supported. Instead, the user simply formulates their query in natural language, and the model responds in kind. The user then has the option to continue the chat by asking another question, which the tool will respond to with the complete context of both questions. Introduced in October 2023, Lexis+ AI states that it has access to LexisNexis's entire repository of case law, codes, rules, constitution, agency decisions, treatises, and practical guidance, all of which it presumably uses to craft its responses. While not much technical detail is published, it is known that Lexis+ AI implements a proprietary RAG system that ensures that every prompt "undergoes a minimum of five crucial checkpoints ... to produce the highest quality answer" (Wellen 2024b).[12]

Ask Practical Law AI, introduced in January 2024 and offered on the Westlaw platform, is a more limited product, but it operates in a similar way. Like Lexis+ AI, Ask Practical Law AI also functions as a chatbot, allowing the user to input their queries in natural language and responding to them in the same format. However, instead of accessing all the primary sources that Lexis+ AI uses, Ask Practical Law AI only retrieves information from Thomson Reuters's database of "practical law" documents—"expert resources ... that have been created and curated by more than 650 bar-admitted attorney editors" (Thomson Reuters 2024b) promising "90,000+ total resources across 17 practice areas" (Thomson Reuters 2024a). Thomson Reuters markets this database for general legal research: "Practical Law provides trusted, up-to-date legal know-how across all major practice areas to help attorneys deliver accurate answers quickly and confidently." Performing RAG on these materials, Thomson Reuters claims, ensures that its system "only returns information from [this] universe" (Thomson Reuters 2024b).

Westlaw's AI-Assisted Research (AI-AR), introduced in November 2023, is also a standard chatbot interface, promising "answers to a far broader array of questions than what we could anticipate with human power alone" (Thomson Reuters 2023). The RAG system retrieves information from Westlaw's databases of cases, statutes, regulations, West Key Numbers, headnotes, and KeyCite markers (Thomson Reuters 2023). While not much technical detail is provided, AI-AR appears to rely on OpenAI's GPT-4 system (Ambrogi 2023). This system was built out after a $650 million acquisition of Casetext, which had developed legal research systems on top of GPT-4 (Ambrogi 2023). RAG is prominently touted as addressing hallucinations: one Thomson Reuters official stated, "We avoid [hallucinations] by relying on the trusted content within Westlaw and building in checks and balances that ensure our answers are grounded in good law" (Thomson Reuters 2023). While AI-AR has been sold to law firms, it has not been made available generally for educational and research purposes.[13]

Both AI-AR and Ask Practical Law AI are made available via the Westlaw platform and are commonly referred to as AI products within Westlaw.[14] For shorthand, we will refer to Ask Practical Law AI as a Thomson Reuters system and AI-AR as a Westlaw system, as this appears to track the internal company product distinctions.

To provide a point of reference for the quality of these bespoke legal research tools—and because AI-AR appears to be built on top of GPT-4—we also evaluate the hallucination rate and response quality of GPT-4, a widely available LLM that has been adopted as a knowledge-work assistant (Collens et al. 2024; Dell'Acqua et al. 2023). GPT-4's responses are produced in a "closed-book" setting; that is, produced without access to an external knowledge base.

5.2 | Query Construction

We design a diverse set of legal queries to probe different aspects of a legal RAG system's performance. We develop this benchmark dataset to represent real-life legal research scenarios, without prior knowledge of whether they would succeed or fail.

For ease of interpretation, we group our queries into four broad categories:

1. **General legal research questions:** common-law doctrine questions, holding questions, or bar exam questions

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. **Jurisdiction or time-specific questions:** questions about circuit splits, overturned cases, or new developments
3. **False premise questions:** questions where the user has a mistaken understanding of the law
4. **Factual recall questions:** queries about facts of cases not requiring interpretation, such as the author of an opinion and matters of legal citation

Queries in the first category ($n = 80$) are the paradigmatic use case for these tools, asking general questions of law. For instance, such queries pose bar exam questions that have ground-truth answers, but in contrast to assessments that focus only on the accuracy of the multiple choice answer (e.g., Martínez 2024), we assess hallucinations in the fully generated response. Queries in the second category ($n = 70$) probe for jurisdictional differences or developing areas in the law, which represent precisely the kinds of active legal questions requiring up-to-date legal research. Queries in the third category ($n = 22$) probe for the tendency of LLMs to assume that premises in the query are true, even when flatly false. The last category ($n = 30$) probes the extent to which RAG systems are able to overcome known vulnerabilities about how general LLMs encode legal knowledge (Dahl et al. 2024).

Table 2 describes these categories in more depth and provides an example of a question that falls within each category. We used 20 queries from LegalBench's Rule QA task verbatim (Guha et al. 2023), and 20 BARBRI bar exam prep questions verbatim (BARBRI Inc. 2013). Each of the 162 other queries were handwritten or adapted for use in our benchmark. Appendix A provides a more granular list of the types of queries and descriptive information.

Our dataset advances AI benchmarking in five respects. First, it is expressly designed to move the evaluation of AI systems from standard question-answer settings with a discrete and known answer (e.g., multiple choice) to the generative (e.g., open-ended) setting (Li and Flanigan 2024; McIntosh et al. 2024; Raji et al. 2021). Prior work has evaluated the amount of legal information that LLMs can produce (Dahl et al. 2024), but this kind of benchmark does not capture the practical benefits and risks of everyday use cases. Legal practice is more than answering multiple choice questions. Of course, because these are not simple queries, their design and evaluation is time-intensive—all queries must be written based on external legal knowledge and submitted by hand through the providers' web interfaces, and evaluation of answers requires careful assessment of the tool's legal analysis and citations, which can be voluminous.

Second, our queries are specifically tailored to RAG-based, open-ended legal research tools. This differentiates our dataset from previously released legal benchmarks, like LegalBench (Guha et al. 2023). Most LegalBench tasks are tailored towards legal analysis of information given to the model in the prompt; tasks like contract analysis or issue spotting. Our queries are written specifically for RAG-based legal research tools; each query is an open-ended legal question that requires legal analysis supported by relevant legal documents that the model must retrieve. This provides a more realistic representation of the way that lawyers are intended to use these tools. Our goal with our dataset is to move beyond anecdotal accounts and offer a systematic

TABLE 2 | The high-level categories of the query dataset, with counts and percentages (Perc.) of queries, descriptions, and sample queries.

Category	Count	Perc.	Description	Example query
General legal research	80	39.6%	Common-law doctrine questions, previously published practice bar exam questions, holding questions	Has a habeas petitioner's claim been "adjudicated on the merits" for purposes of 28 U.S.C. § 2254(d) where the state court denied relief in an explained decision but did not expressly acknowledge a federal-law basis for the claim?
Jurisdiction or time-specific	70	34.7%	Questions about circuit splits, overturned cases, or new developments	In the Sixth Circuit, does the Americans with Disabilities Act require employers to accommodate an employee's disability that creates difficulties commuting to work?
False premise	22	10.9%	Questions where the user has a mistaken understanding of the law	I'm looking for a case that stands for the proposition that a pedestrian can be charged with theft for absorbing sunlight that would otherwise fall on solar panels, thereby depriving the owner of the panels of potential energy.
Factual recall questions	30	14.9%	Basic queries about facts not requiring interpretation, like the year a case was decided	Who wrote the majority opinion in Candela Laser Corp. v. Cynosure Inc., 862 F. Supp. 632 (D. Mass. 1994)?

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

investigation of the potential strengths and weaknesses of these tools, responding to documented challenges in evaluating AI in law (Guha et al. 2023; Kapoor et al. 2024).

Third, these queries are designed to represent the temporal and jurisdictional variation (e.g., overruled precedents and circuit splits) that is often the subject of live legal research (Beim and Rader 2019). We hypothesize that AI systems are not able to encode this type of multifaceted and dynamic knowledge at the moment, but these are precisely the kinds of inquiries requiring legal research. Due to the nature of legal authority, attorneys will inevitably have questions specific to their time, place, and facts, and even the most experienced lawyers will need to ground their understanding of the legal landscape when facing issues of first impression.

Fourth, the queries probe for "contrafactual bias," or the tendency of chat systems to assume the veracity of a premise even when false (Dahl et al. 2024). Many claim that AI systems will help to address longstanding access to justice issues (Bommasani et al. 2022; Chien and Kim 2024; Chien et al. 2024; Perlman 2023; Tan et al. 2023), but contrafactual bias poses a particular risk for *pro se* litigants and lay parties.

Last, to guard against selection bias in our results (i.e., choosing queries based on hallucination results), we modeled best practices with our dataset by preregistering our study and associated queries with the Open Science Foundation prior to performing our evaluation (Surani et al. 2024).[15]

5.3 | Query Execution

For Lexis+ AI, Thomson Reuters's Ask Practical Law AI, and Westlaw's AI-AR, we executed each query by copying and pasting it into the chat window of each product.[16] For GPT-4, we prompted the LLM via the OpenAI API (model `gpt-4-turbo-2024-04-09`) with the following instruction, appending the query afterwards:

> You are a helpful assistant that answers legal questions. Do not hedge unless absolutely necessary, and be sure to answer questions precisely and cite caselaw for propositions.

This prompt aims to ensure comparability with legal AI tools, particularly by prompting for legal citations and concrete factual assertions. We recorded the complete response that each tool gave, along with any references to case law or documents. The dataset was preregistered on March 22, 2024 and all queries on Lexis+ AI, Ask Practical Law AI, and GPT-4 were run between March 22 and April 22, 2024. Queries on Westlaw's AI-AR system were run between May 23–27, 2024.

5.4 | Inter-Rater Reliability

To code each response according to the concepts of correctness, groundedness, and hallucination, we relied on our expert domain knowledge in law to hand-score each model response according to the rubric developed in Section 4. As noted above, efficiently evaluating AI-generated text remains an unsolved problem with inevitable trade-offs between internal validity, external validity, replicability, and speed (Hashimoto et al. 2019; Liu et al. 2016; Smith et al. 2022). These problems are particularly pronounced in our legal setting, where our queries represent real legal tasks. Accordingly, techniques of letting these legal AI tools "check themselves"—which have become popular in other AI evaluation pipelines (Manakul et al. 2023; Mündler et al. 2023; Zheng et al. 2023)—are not suitable for this application. Precisely because adherence to authority is so important in legal writing and research, our tasks must be qualitatively evaluated by hand according to the definitions of correctness and groundedness that we have carefully constructed. This makes studying these legal AI tools expensive and time-consuming: this is a cost that must be reflected in future conversations about how to responsibly integrate these AI products into legal workflows.

To ensure that our queries are sufficiently well-defined and that our coding definitions are sufficiently precise, we evaluated the inter-rater reliability of different labelers on our data. Task responses were first graded by one of three different labelers. A fourth labeler then labeled a random sample of 48 responses, stratified by model and task type. We oversampled *The Bluebook* citation task slightly because it is particularly technical. The fourth labeler did not discuss anything with the first three labelers and did not have access to the initial labels. Their knowledge of the labeling process came only from our written documentation of labeling criteria, fully described in Appendix D.

With this protocol, we find a Cohen's kappa (Cohen 1960) of 0.77 and an inter-rater agreement of 85.4% on the final outcome label (correct, incomplete, or hallucinated) between the evaluation labeler and the initial labels. This is a substantial degree of agreement that suggests that our task and taxonomy of labels are well defined. Our results are comparable to similar evaluations for complex, hand-graded legal tasks (Dahl et al. 2024).[17]

6 | Results

Section 6.1 describes our findings on hallucinations and responsiveness. Section 6.2 examines the varied and sometimes insidious nature of hallucinations. Section 6.3 provides a typology of the potential causes of inaccuracies we encountered.

6.1 | Hallucinations Persist Across Query Types

Commercially-available RAG-based legal research tools still hallucinate. Over 1 in 6 of our queries caused Lexis+ AI and Ask Practical Law AI to respond with misleading or false information. Westlaw hallucinated substantially more—*one-third of its responses* contained a hallucination.

On the positive side, these systems are less prone to hallucination than GPT-4, but users of these products must remain cautious about relying on their outputs.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The left panel of Figure 4 provides a breakdown of response types across the four products. Lexis+ AI's answers are accurate (i.e., correct and grounded) for 65% of queries, compared to much lower accuracy rates of 41% and 19% by Westlaw and Practical Law AI, respectively. The right panel of Figure 4 also provides the hallucination rate when an answer is responsive, showing that Lexis+ AI appears to have a statistically significantly lower hallucination rate than Westlaw and Thomson Reuters, even conditional on a response.

Figure 5 also breaks down these statistics by query type. We observe that, while hallucination rates are slightly higher for jurisdiction and time-specific questions, they remain high for general legal research questions, such as questions posed on the bar exam. Accuracy rates are highest on "false premise" questions—in which the query contains a mistaken understanding of law—and lower on the categories that represent real-world use by attorneys.

Westlaw's high hallucination rate is driven by several kinds of errors (as discussed further in Section 6.2), but we note that it is also the system which tends to generate the longest answers. Excluding refusals to answer, Westlaw has an average word length of 350 (SD = 120), compared to 219 (SD = 114) by Lexis+ AI and 175 (SD = 67) by Ask Practical Law AI.[18] With longer answers, Westlaw contains more falsifiable propositions and therefore has a greater chance of containing at least one hallucination. Lengthier answers also require substantially more time to check, verify, and validate, as every proposition and citation has to be independently evaluated.

Responsiveness differs dramatically across systems. As shown in Figure 4, Lexis+ AI, Westlaw AI-AR, and Ask Practical Law AI provide incomplete answers 18%, 25%, and 62% of the time, respectively. The low responsiveness of Ask Practical Law AI can be explained by its more limited universe of documents. Rather than connecting its retrieval system to the general body of law (including cases, statutes, and regulations), Ask Practical Law AI draws solely from articles about legal practice written by its in-house team of lawyers.

On the other hand, the Westlaw and Lexis retrieval systems are connected to a wider body of case law and primary sources. This means that they have access to all the documents that are in principle necessary to answer any of our questions. Both systems often offer high-quality responses. In one instance, Lexis+ AI pointed to a false premise in one of our questions. The question `scalr-19` asked whether the 6 year statute of limitation applied to retaliatory discharge actions under the False Claims Act. The question was drawn from *Graham County Soil & Water Conservation District v. U.S.*, 559 U.S. 280 (2010), where the Court held that there was ambiguity. Congress moved thereafter to amend the statute to clarify the statute of limitations. Lexis+ AI explained the mistaken premise, and cited the relevant, updated code section. Similarly, when prompted about the need for specific, proven "teaching, suggestion, or motivation" (TSM) that would have led a person of ordinary skill in the art to combine the relevant prior art for a finding of obviousness, AI-AR correctly responded by discussing the Supreme Court's decision in *KSR v. Teleflex*, 550 U.S. 398 (2007), which rejected a rigid notion of the Federal Circuit's TSM test.

6.2 | Hallucinations Can Be Insidious

These systems can be quite helpful when they work. But as we now illustrate in detail, their answers are often significantly flawed. We find that these systems continue to struggle with elementary legal comprehension: describing the holding of a case (Zheng et al. 2021), distinguishing between legal actors (e.g., between the arguments of a litigant and the holding of the court), and respecting the hierarchy of legal authority. Identifying these misunderstandings often requires close analysis of cited sources. These vulnerabilities remain problematic for AI adoption in a profession that requires precision, clarity, and fidelity.

Tables 3, 4, and 5 provide examples of hallucinations in the Westlaw, Lexis, and Practical Law systems, respectively.[19] In each example, our detailed analysis of responses and cited cases reveals a serious inaccuracy and hallucination in the system

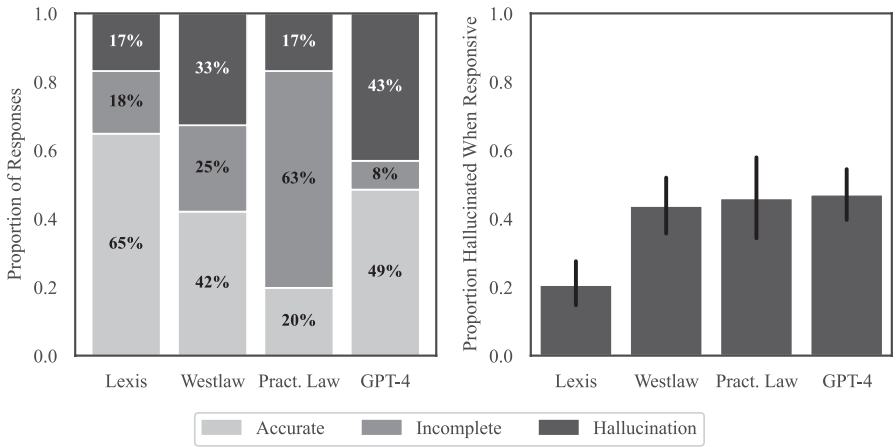

FIGURE 4 | *Left panel*: Overall percentages of accurate, incomplete, and hallucinated responses. *Right panel*: The percentage of answers that are hallucinated when a direct response is given. Westlaw AI-AR and Ask Practical Law AI respond to fewer queries than GPT-4, but the responses that they do produce are not significantly more trustworthy. Vertical bars denote 95% confidence intervals.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

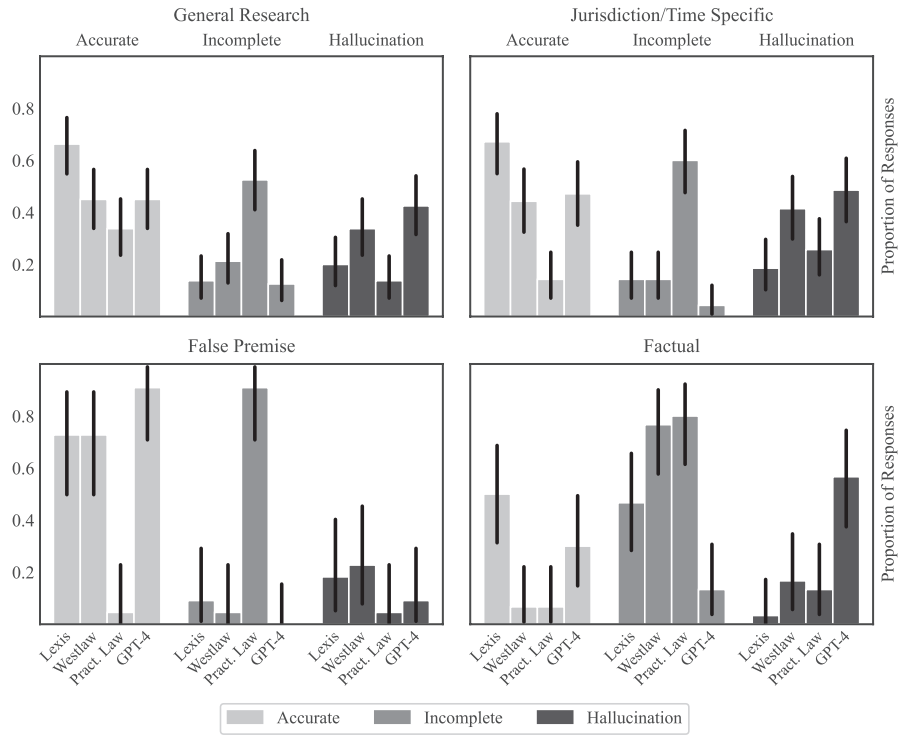

FIGURE 5 | Response evaluations broken down by question category. We show the accuracy (lightest), incompleteness (middle shade), and hallucination (darkest) rate for each question category. Vertical bars denote 95% confidence intervals. This figure shows that hallucinations are not driven by an isolated category and persist across task types and questions, such as bar exam and appellate litigation issues.

response. The following sections refer to examples in these tables to illustrate different failure modes in legal RAG systems.

6.2.1 | Misunderstanding Holdings

Systems do not seem capable of consistently making out the holding of a case. This is a serious issue, as legal research relies centrally on distinguishing the holding from other parts of the case. Table 3 rows 1, 4, 5, and 6 provide examples of when Westlaw states a summary that is the direct opposite of the actual holding of a case, including a case by the U.S. Supreme Court. For instance, Westlaw states that collateral is considered a return of "any part" of the loan, indicating that this was established by the Supreme Court in *Robers v. U.S.*, but *Robers* held the exact opposite (Table 3 row 1). In another response, Lexis+ AI recites Missouri legislation criminalizing unauthorized camping on state-owned lands. But that legislation comes from the statement of facts and analysis, and in the cited case, the Missouri Supreme Court actually held that legislation unconstitutional (Table 4, row 3).

6.2.2 | Distinguishing Between Legal Actors

Systems can fail to distinguish between arguments made by litigants and statements by the court. In one example, Westlaw attributes an action of the defendant to the court (Table 3 row 8) and in another, it stated that a provision of the U.S. Code was found unconstitutional by the 10th Circuit when, in fact, the 10th Circuit rejected that argument by the defendant (Table 3, row 10).

6.2.3 | Respecting the Order of Authority

All models strain in grasping hierarchies of legal authority. This is crucial, as courts often discuss similar propositions that may be in tension. When sources conflict, a complex system of precedence and hierarchy determines governing law. Sorting through different sources to find the authoritative ones requires legal "background knowledge" about the way that different courts interact in different jurisdictions, and even systems with direct access to case law can fail to adhere to these legal hierarchies. For example:

- Westlaw asserts that a U.S. Supreme Court case was reversed by the *Nebraska Supreme Court* on a matter of federal law. That is not possible in the U.S. legal system, and in fact, the Nebraska Supreme Court did not so much as cite the Supreme Court case in question (Table 3 row 2).
- Westlaw confuses holdings between different levels of courts (Table 3 rows 5, 9). In row 9, for instance, Westlaw properly states the holding of the Ninth Circuit panel, but improperly attributes it to the Ninth Circuit sitting en banc, which actually overruled the panel on that issue.
- Lexis+ AI fails to distinguish between district and appellate courts. In Table 4 row 1, Lexis+ AI transmogrifies a district court recitation of the trial court standard for awarding attorney's fees into a patently incorrect standard of *appellate review* of attorney's fees (incorrectly stating that an appeals court may disturb attorney's fees "as long as they provide reasoning").

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | Ten examples of hallucinations in Westlaw's AI-Assisted Research responses, with explanations for why they are coded as hallucinations.

Westlaw AI-AR Hallucination	Explanation
1. "[W]hen the lender receives the collateral that secured the fraudulent loan, this is considered a return of 'any part' of the loan money ... This was established in the Supreme Court case Robers v. U.S."	*Robers* held precisely the opposite: "the phrase 'any part of the property ... returned' refers to the property the banks lost ... and not to the collateral." 572 U.S. 639, 642 (2014).
2. "D.M. v. State ... has been overruled by Davis v. State. Also, the case *Millbrook v. U.S. was reversed* by the same case at a later date."	*Millbrook v. United States* is a U.S. Supreme Court decision that controls on federal questions. 569 U.S. 50 (2013). The Nebraska Supreme Court did not cite, much less "reverse," it in *Davis v. State*. 297 Neb. 955 (2017).
3. "[A] paragraph from the Federal Rules of Bankruptcy Procedure [FRBP], Rule 4007 states that the deadlines set by bankruptcy rules governing the filing of dischargeability complaints are jurisdictional."	There is no such paragraph in the FRBP and 4007's deadlines are unlikely to be jurisdictional under *Kontrick v. Ryan*, 540 U.S. 443, 447–48, 448 n.3 (2004).
4. "The Supreme Court of Alaska ... noted ... in the absence of physical injury, the losses for claimed emotional damage [from a misdiagnosis] should not be permitted for legal and public policy reasons (Chizmar v. Mackie, 896 P.2d 196 (1995))."	The Supreme Court of Alaska in fact reversed this trial court holding (that losses are not permitted on public policy grounds) and allowed the question to proceed to the jury. 896 P.2d at 204–05.
5. "[I]n the Doo v. Packwood case, a restrictive covenant precluding the operation of a grocery store in a building conveyed was found to be potentially inequitable and unenforceable ... (Doo v. Packwood, 265 Cal.App.2d 752 (1968))."	The *Doo* court in fact held the opposite. It reversed the trial court's finding that the covenant was unenforcable. 265 Cal.App.2d at 756, 759.
6. "[T]he FCC's new policy regarding 'fleeting expletives' is arbitrary and capricious under the Administrative Procedure Act."	The Supreme Court in fact found the opposite: the fleeting expletives policy was **not** arbitrary and capricious. FCC v. Fox, 556 U.S. 502 (2009).
7. "In the Reid v. Fargo case, there is no mention or reference made to the Irvine v. The Hesper case."	The *Reid* court expressly relied on the *Hesper* case. 241 U.S. 544, 548 (1916).
8. "In the case of Bragg v. Linden Research Inc., the **court** recognized participants' full intellectual property protection for the digital content they created or owned in the game Second Life ... Bragg v. Linden Research Inc., 487 F.Supp.2d 593 (2007)."	The court took no such position and was only describing actions by the **defendant** (Linden): "Linden announced that it would recognize participants' full intellectual property protection." 487 F.Supp.2d at 595.
9. "In Washington Legal Foundation v. Legal Foundation of Washington, the Court of Appeals for the Ninth Circuit held that the interest generated by IOLTA pooled trust accounts is the property of the clients, and the appropriation of that interest for charitable purposes was a per se 'taking' within the meaning of the Fifth Amendment."	This statement falsely credits the holding of a Ninth Circuit panel instead of the superseding holding of the Ninth Circuit en banc court, which overruled the panel. 271 F.3d 835, 857–58 (applying ad hoc, not per se takings analysis).
10. "*18 U.S.C.A. § 3583* ... was recognized as unconstitutional by United States v. Winter Rose Old Rock.	The *Winter Rose* court in fact rejected the defendant's claim that § 3583 was unconstitutional, 76 F.4th 1314, 1316 (10th Cir. 2023), and only discussed and distinguished *Haymond*, 588 U.S. 634 (2019) (4–1–4 decision with judgment that specific application of § 3583**(k)** was unconstitutional).

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Six examples of hallucinations in Lexis+ AI responses with explanations for why they are coded as hallucinations.

Lexis+ AI Hallucination	Explanation
1. "In Vadatech v. Overnight [2016 Nev. Dist. LEXIS 1619], the court awarded attorney's fees and costs to the prevailing party. It cited Nevada Supreme Court precedent that appellate courts can modify fee awards as long as they provide reasoning, without needing to find an abuse of discretion."	Vadatech, a district court opinion, said no such thing and instead recited the trial court standard for granting attorney's fees—set by the Nevada Supreme Court—not the standard of appellate modification of such fees.
2. "If a driver is unable to produce identification documents or produces documents that appear to be false, a limited search of places in the vehicle where the driver may have hidden a wallet is reasonable. In the case of Arturo D., the court upheld the warrantless vehicle search when the driver failed to produce the required documentation People v. Lopez, 8 Cal. 5th 353."	*Lopez* actually overturned *Arturo D.*.. Lexis+ AI describes *Arturo D.* as if it is good law and supported by *Lopez*.
3. "The new section 67.2300 of TAFP HB 1606, which was signed into law by Governor Parson ... makes the act of unauthorized sleeping and camping on state-owned lands a class C misdemeanor. Byrd v. State, 679 S.W.3d 492."	HB 1606 was in fact struck down by the Missouri Supreme Court in the cited case *Byrd v. State*, 679 S.W.3d 492 (Mo. 2023). This detail is not mentioned in Lexis+ AI's response.
4. "The equity clean-up doctrine is a self-imposed ordinance that closes the doors of a court of equity to one tainted with an inequitableness or bad faith relative to the matter in which he seeks relief. ... EEOC v. Recruit U.S.A. Inc., 939 F.2d 746."	This statement confuses the equity clean-up doctrine with the "clean hands" doctrine. EEOC v. Recruit U.S.A., 939 F.2d 746, 752 (9th Cir. 1991).
5. "The Telephone Consumer Protection Act of 1991 (TCPA) grants federal courts exclusive jurisdiction over actions brought by the state attorneys general, but places exclusive jurisdiction over private civil actions in state courts."	The Supreme Court in fact found concurrent state and federal court jurisdiction over private suits under the TCPA. Mims v. Arrow Financial Services, 565 U.S. 368 (2012).
6. "The Eleventh Circuit has not definitively ruled on the causation standard for Family and Medical Leave Act (FMLA) retaliation claims. Some courts have applied the "but-for" causation standard, which requires a showing that the adverse action would not have occurred but for the protected activity Lapham v. Walgreen Co., 88 F.4th 879 ... However, other courts have applied a motivating factor or negative factor standard ... Lee v. Yellowpages.Com LLC, 2019 U.S. Dist. LEXIS 250502, Nelson v. Americold Logistics LLC, 2020 U.S. Dist. LEXIS 63611."	Lexis claims the standard is not definitive, but *Lapham* is the most recent controlling case, and the other cases contesting the issue are older inferior court rulings.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 | Four examples of hallucinations in Thomson Reuters's Ask Practical Law AI response, with explanations of why they are coded as hallucinations. The Practical Law documents cited are named in square brackets.

Practical Law Hallucination	Explanation
1. "Based on 28 U.S.C § 2680(h), the United States does waive its sovereign immunity for certain intentional torts committed by federal investigative or law enforcement officers acting within the scope of their employment. However, this waiver applies only if these officers are acting in their capacity to search, seize evidence, or make arrests for federal crimes [Federal Tort Claims Act (FTCA): Exceptions and Defenses]."	The cited Practical Law resource states: "However, the tortious act need not occur while the officer is searching, seizing, or arresting for the proviso to apply (Millbrook v. United States, 569 U.S. 50, 55–57 (2013)."
2. "Yes, a person subject to domestic-violence-related restraining orders can be prohibited from purchasing a firearm in the Fifth Circuit. [State and Local Regulation of Firearms Under the Second Amendment]."	The Practical Law article actually states: "Federal law prohibits a person who is subject to a domestic violence restraining order from possessing a firearm ... However, the US Court of Appeals for the Fifth Circuit has held that this law violates the Second Amendment (United States v. Rahimi, 61 F.4th 443 (5th Cir. 2023)."
3. "A dying declaration is admissible in federal court under the Federal Rules of Evidence (FRE) 804(b)(2) This provision is commonly invoked in homicide prosecutions, where the victim's last words about the identity or description of their killer are introduced into evidence. However, it could theoretically apply to any case where these specific conditions are met."	The dying declaration exception is not just "commonly" invoked in homicide prosecutions; FRE 804(b)(2) ... specifically limits the admissibility of dying declarations to cases involving "a prosecution for homicide or in a civil case," not any case.
4. "The undue burden standard applies in abortion cases when assessing the constitutionality of state regulations on abortion before a fetus becomes viable. ... [Health Plan Coverage of Surgical and Medication Abortion and Related Services, Group Health Plan Coverage Considerations for Employers After the Overturning of Roe v. Wade, Employee Assistance Program (EAP) Compliance]."	Dobbs v. Jackson Women's Health, 597 U.S. 215 (2022) overturns *Casey* and the undue burden standard. All three Practical Law articles discuss *Roe* and *Casey*'s reversal, with one title including the phrase "Overturning of *Roe*."

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

- In Table 4 row 2, Lexis+ AI describes a rule established in *Arturo D.* as good law, with citation to the case that actually overrules *Arturo D.*

We note one additional area where systems struggle with orders of authority. In numerous instances, we observed the Westlaw system stating a proposition based on an overruled or reversed case, *without* citing the case. These errors may stem from design choices: Westlaw may be adding citations in a second pass, after generating the statement, while suppressing the citation of cases that receive a "red flag" under its KeyCite system.[20] For instance, when prompted about the equity clean-up doctrine, which allows courts of equity to decide legal and equity issues when it has jurisdiction over the equity issues, AI-AR properly cites the rule, but then notes, "However, this general rule does not apply when the facts relied on to sustain the equity jurisdiction fail of establishment." This statement is unaccompanied by an in-text citation; the language appears only in a search result below the response, in a Missouri case[21] that was overruled on that issue by the Missouri Supreme Court.[22] We believe this suppression behavior can be dangerous—it impedes verification of the claims most likely to be false.

6.2.4 | Fabrications

The systems we test occasionally generate text that is unrelated or deviates materially from retrieved documents.

- Westlaw generates provisions of law that do not exist. For instance, it asserts that the Federal Rules of Bankruptcy (FRBP) state that deadlines are jurisdictional, which is not a statement contained in the FRBP text at all (Table 3 row 3). (The hallucination seems to emanate from a retrieved 1996 bankruptcy court case, which is also likely invalid under the Supreme Court's *Kontrick* decision, which found that bankruptcy deadlines are not jurisdictional.)
- Westlaw misinterprets the Supreme Court's specific holding on a statutory *subsection* as the 10th Circuit finding the entire statutory section unconstitutional, when in fact the 10th Circuit rejected the defendant's claim of unconstitutionality (Table 3 row 10).
- Lexis+ AI attributes a description of the equity clean-up doctrine to a case that only discusses the "clean hands" doctrine (Table 4 row 4).

6.3 | A Typology of Legal RAG Errors

Interpreting why an LLM hallucinates is an open problem (Ji et al. 2023; Zou, Phan, et al. 2023). While it is possible to identify correlates of hallucination (Dahl et al. 2024), it is hard to conclusively explain why a model hallucinates on one question but does not on another, or why one model hallucinates where another does not.

RAG systems, however, are composed of multiple discrete components (Gao et al. 2024). While each piece may be a black box, due to the lack of documentation by providers, we can partially observe the way that information moves between them. Lexis+ AI, Ask Practical Law AI, and AI-AR each show the list of documents that were retrieved and given to the model (though not exactly which pieces of text are passed in). Consequently, comparing the retrieved documents and the written response allows us to develop likely explanations for the reasons for hallucination.

In this section, we present a typology of different causes of RAG-related hallucination that we observe in our dataset. Other analyses of RAG failure points identify a larger number of distinct failure points (Barnett et al. 2024; Chen et al. 2024). Our typology collapses some of these since we focus on broader causes that can be identified using the limited information we have about the systems we test. Our typology also introduces new failure points unique to the legal context that have not previously been considered in analyses of general-purpose RAG systems. Evaluations of general purpose RAG systems often assume that all retrievable documents (1) contain true information and (2) are authoritative and applicable, an assumption that is not true in the legal setting (Barnett et al. 2024; Chen et al. 2024).[23] Legal documents often contain outdated information, and their relevance varies by jurisdiction, time period, statute, and procedural posture. Determining whether a document is binding or persuasive often requires non-trivial reasoning about its content, metadata, and relationship with the user query.

This typology is intended to be useful to both legal researchers and AI developers. For legal researchers, it illustrates some pathways to incorrect outputs and highlights specific areas of caution. For developers, it highlights areas for improvement in these tools. The categories that we present are not mutually exclusive; the failures we observe are often driven by multiple causes or have unclear causes. Table 6 compares the prevalence of different hallucination causes in our typology. Because these are closed systems, we are not able to clearly identify a single point of failure for each hallucination.

6.3.1 | Naive Retrieval

Many failures in the three systems stem from poor retrieval—failing to find the most relevant sources available to address the user's query. For instance, when asked to define the "moral wrong doctrine," a doctrine pertaining to mistake-of-fact instructions in criminal prosecutions for morally wrongful acts (`doctrine-test-177`), Lexis+ AI relies on a source which defines moral turpitude, a legal term of art with a seemingly similar but actually unrelated meaning.

Part of the challenge is that retrieval itself often requires legal reasoning. As Section 3.2 discusses, legal sources are not composed of unambiguous facts. Lawyers are often taught to analyze situations with an IRAC framework—first identify the issue (I) and governing legal rule (R), then analyze (A) the facts with that rule to arrive at a conclusion (C) (Guha et al. 2023). For example, `bar-exam-96` asks whether an airline's motion to dismiss should be granted in a wrongful death suit arising out of a plane crash. Ask Practical Law AI retrieves sources discussing motions to dismiss in various contexts such as bankruptcy and patent litigation. But correctly answering this question requires

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | This table shows the prevalence of different contributing causes among all hallucinated responses for each model. Because the types are not mutually exclusive, the proportions do not sum to 1.

Contributing cause	Lexis	Westlaw	Pract. Law
Naive Retrieval	0.47	0.20	0.34
Inapplicable Authority	0.38	0.23	0.34
Reasoning Error	0.28	0.61	0.49
Sycophancy	0.06	0.00	0.03

identifying the true underlying issue as being one about tort negligence, not general procedures for motions to dismiss. Thomson Reuters's tool likely errs because it fails to perform this analytical step prior to querying its database, thereby ending up with sources pertaining to the wrong issue.

6.3.2 | Inapplicable Authority

An inapplicable authority error occurs when a model cites or discusses a document that is not legally applicable to the query. This can be because the authority is for the wrong jurisdiction, wrong statute, wrong court, or has been overruled. This kind of error is uniquely important and prevalent in the legal setting, and has not been explored as thoroughly in prior literature (Barnett et al. 2024; Gao et al. 2024). One example is Lexis+ AI's response to `scalr-15`. This question asks about certain deadlines under Bankruptcy Rule 4004, but the model describes and cites a case about tax court deadlines under 26 U.S.C.S. § 6213(a) instead. This could be because the excerpt of the case that is given to the model does not include key information, or because the model was given that information and ignored it. Because it is not possible to see exactly what information is available to the model, it is not possible to say precisely where the error occurs.

6.3.3 | Sycophancy

LLM assistants have been found to display "sycophancy," a tendency to agree with the user even when the user is mistaken (Sharma et al. 2023). While sycophancy can cause hallucinations (Dahl et al. 2024), we found that Lexis+ AI, AI-AR, and GPT-4 were quite capable of navigating our false premise queries, and often corrected the false premise without hallucination. For example, `false-holding-statements-108` asks for a case showing that due process rights can be violated by negligent government action. Lexis+ AI steers the user towards the correct answer, stating that intentional interference can violate due process, and that negligent interference cannot, supporting these propositions with case law. Ask Practical Law AI also seldom hallucinated in this category, but refused to answer at all in the overwhelming majority of queries.

6.3.4 | Reasoning Errors

In addition to the more complex behaviors described above, LLM-based systems also tend to make elementary errors of reasoning and fact. The legal research systems we test are no exception. We observe such errors most frequently in Westlaw; though retrieved results often seemed relevant and helpful, the model would not always correctly reason through the text to arrive at the correct conclusion. In one instance (Table 3 row 8), AI-AR describes a district court decision as "recogniz[ing] participant's full intellectual property protection for the digital content they created or owned in the game Second Life..." But as the passage cited by the model makes clear, the court held no such thing. It was describing the statements of the defendant, and the language model made a simple factual error in describing the passage given to it.

7 | Limitations

While our study provides critical information about widely deployed AI tools in legal practice, it comes with certain limitations.

First, our evaluation is limited to three specific products by LexisNexis, Thomson Reuters, and Westlaw. The legal AI product space is growing rapidly with many startups (e.g., Harvey, Vincent AI) (Ma et al. 2024). Access to these emerging systems is even more restricted than to the services offered in LexisNexis and Westlaw, making evaluation exceptionally challenging.[24] That said, our approach provides a common benchmark that can be deployed for similar systems as they become available.

Second, our evaluation only captures a point in time. Even over the course of our study, we noticed the responses of these systems—particularly Lexis+ AI—evolve over time. While these changes may improve responses, we note that benchmarking, evaluation, and supervision remain difficult when a model changes over time (Chen et al. 2023).[25] This is compounded by uncertainty over whether such differences are driven by changes in the base model (e.g., GPT-4) or by engineering by the legal technology provider. More generally, a fundamental concern for the evaluation of LLMs lies in test leakage—because language models are trained on all available data, they may memorize data that is used for evaluation (Deng et al. 2023; Li and Flanigan 2024; Oren et al. 2023). That is a particularly challenging concern when the only mechanism for accessing legal AI tools is by sending test prompts to providers themselves. Even if providers fix the discrete errors noted above, that may not mean that the problems we identify have been solved in general.[26]

Third, while we have been able to design an effective evaluation framework for chat-based interfaces, the evaluation for more specified generative tasks is still evolving. LegalBench (Guha et al. 2023), for instance, still requires manual evaluation of certain generative outputs, and we do not here assess Casetext CoCounsel's effectiveness at drafting open-ended legal memoranda. Developing benchmarks for the full range of legal tasks—for example, deposition summaries, legal memoranda, contract review—remains an important open challenge for the field (Kapoor et al. 2024).

Fourth, although we designed the first benchmark dataset, the sample size of 202 queries remains small in comparison to other evaluations such as Dahl et al. (2024). There are two

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reasons for this. In contrast to general-purpose LLMs, which have open models or API access, LexisNexis, Thomson Reuters, and Westlaw restrict access to their interfaces.[27] In addition, extensive manual work is required to evaluate the results of each query, making it harder to scale automated evaluations. The trend toward LLM-based evaluations may address the latter obstacle, but the fact remains that the legal AI product space remains quite closed.[28]

Fifth, while we managed to develop a measurement protocol that yielded substantial agreement between human raters, we acknowledge that groundedness may exist on a spectrum. A citation, for instance, might point to a case that has been overruled, but that case might still be helpful to an attorney in starting the research process. In our setting, we coded such instances as misgrounded, but whether the model is helpful will still fundamentally have to be determined by use cases and evaluations that involve human interactions with the system. The range of failure points documented in Section 6.3 provides a more granular sense of the limitations of current AI systems.

Sixth, some might argue that our benchmark dataset does not represent the natural distribution of queries. We designed our benchmark to reflect a wide range of query types and to constitute a challenging real-world dataset. Questions are ones that arise on the bar exam, that arise in appellate litigation, that present circuit splits, that present issues that are dynamically changing, and that were contributed by the legal community (Guha et al. 2023). The benchmark was designed to be challenging precisely because (a) those are the settings where legal research is needed the most, and (b) it responds to the marketing claims by providers. It is true that these may not represent all tasks for which lawyers turn to generative AI. Our estimate of the hallucination rate is not meant to be an unbiased estimate of the (unknown) population-level rate of hallucinations in legal AI queries, but rather to assess whether hallucinations have in fact been solved by RAG, as claimed. We show that hallucinations persist across the wide range of task types (see Figure 1) and the full natural distribution of such queries is (a) only known to legal technology providers, (b) highly in flux given uncertainty about the appropriate use of AI in law, and (c) itself endogenous to assessments of reliability and marketing claims.

Last, our primary goal is limited to assessing the hallucination rate, accuracy, and groundedness on emerging legal technology. These are central concepts to the trustworthiness of AI tools, but they are not the sole criteria for the quality and value of a legal research system. For instance, notwithstanding the many hidden hallucinations, the overall output of Lexis+ AI and AI-AR may still be quite valuable for distinct use cases (e.g., starting on a research thread). But evaluations like the one we designed here are critical to understanding these appropriate use cases.

8 | Implications

Excitement over the potential for AI to transform the practice of law is at an all-time high (Magnanelli 2024; Markelius et al. 2024). On the demand side, lawyers fear missing out on the real gains in efficiency and thoroughness that new AI tools can offer. On the supply side, the companies developing these tools continue to market them as more and more powerful. We agree that these tools are hugely promising (Chien and Kim 2024; Choi et al. 2024), but our research has important implications for both the lawyers using these products and the myriad of companies now marketing them.

8.1 | Implications for Legal Practice

In the United States, all lawyers are required to abide by certain professional and ethical rules. Most jurisdictions have adopted a version of the Model Rules of Professional Conduct, which are issued by the American Bar Association (2018). Two of these rules bear directly on the integration of AI into law: Rule 1.1's duty of competence and Rule 5.3's duty of supervision (Cyphert 2021; Walters 2019; Yamane 2020). Competence requires "legal knowledge, skill, thoroughness and preparation" (Rule 1.1); supervision requires "reasonable efforts to ensure that the [non-lawyer's] conduct is compatible with the professional obligations of the lawyer" (Rule 5.3).

In addition to these rules, the bar associations of New York (2024), California (2023), and Florida (2024) have all recently published more detailed guidance on how lawyers' ethical responsibilities intersect with their use of AI. For example, the New York State Bar Association's AI Task Force states that lawyers "have a duty to understand the benefits, risks and ethical implications" associated with the tools that they use (2024, 57); similarly, the State Bar of California's Standing Committee on Professional Responsibility and Conduct implores lawyers to "understand the risks and benefits of the technology used in connection with providing legal services" (2023, 1).

In other words, lawyers' ability to comply with their professional duties in both of these jurisdictions is contingent on access to *specific* information about empirical risks and benefits of legal AI. Yet, so far, no legal AI company has provided this information. The New York State Bar Association points its members to a list of publications and fora that discuss matters related to AI in general (2024, 76–77), but general knowledge is not the same as understanding the trade-offs of specific tools.

Indeed, our work shows that the risks and benefits associated with AI-driven legal research tools are different from those associated with general-purpose chatbots like GPT-4. As we discuss in Section 6, the tools we study in this article differ in responsiveness and accuracy, and these differences may even change over time within the same tool. The closed nature of these tools, however, makes it difficult for lawyers to assess when it is safe to trust them. Official documentation does not clearly illustrate what they can do for lawyers and in which areas lawyers should exercise caution. Thus, given the high rate of hallucinations that we uncover in this article, lawyers are faced with a difficult choice: either verify by hand each and every proposition and citation produced by these tools (thereby undercutting the efficiency gains that AI is promised to provide), or risk using these tools without full information about their specific risks and benefits (thereby neglecting their core duties of competency and supervision).

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

8.2 | Implications for Legal AI Companies

Legal AI developers face dilemmas as well. On the one hand, these companies are subject to economic pressures to compete in an increasingly crowded market (Ma et al. 2024), pressures made more acute by the recent entry of previously copyrighted and proprietary data into the public domain (Henderson et al. 2022; Östling et al. 2024; The Library Innovation Lab 2024). On the other hand, like all businesses, they are also constrained by laws and regulations limiting the products they can lawfully offer and advertise. We flag two of these potential restrictions here.

First, companies must be careful not to overclaim or misrepresent the abilities of their AI products. As we discuss in Section 1, a number of legal AI providers are currently making claims about their products' ability to "eliminat[e]" (Casetext 2023) or "avoid" hallucinations (Thomson Reuters 2023), yet, as we note in Section 4.3, these same companies are inconsistently using the term "hallucination" in ways that may not conform to users' expectations. Without additional precision about the exact mistakes that their tools purportedly avoid, companies may find themselves exposed to civil liability for unfair competition or false, misleading, or unsubstantiated claims. For instance, under Section 43(a) of the Lanham Act, 15 U.S.C. § 1125, both customers and competitors alike may seek to recover for damages caused by such practices. The Securities and Exchange Commission has charged investment advisers with false and misleading claims about AI (Securities and Exchange Commission 2024), expressing concerns about "AI washing" by public companies (Grewal 2024), and the Federal Trade Commission, too, has warned about deceptive AI claims lacking scientific support (Atleson 2023).

Second, legal AI providers must also be cautious about emerging theories of tort liability for AI-inflicted harms. This territory is less well-charted, but a developing scholarly literature suggests that developers who negligently release AI products with known defects may also face legal exposure (van der Merwe et al. 2024; Wills 2025). For example, one airline company in Canada has already been held liable for negligent misrepresentation based on output produced by its AI chatbot (Rivers 2024). From theories of vicarious liability (Diamantis 2023) to products liability (Brown 2023) to defamation (Salib 2024; Volokh 2023), legal AI providers must carefully weigh the potential tort risks of releasing products with known hallucination problems.

Third, while we have limited information on black-box systems, our typology of legal RAG errors suggests methodological paths for improving systems, such as query expansion prior to retrieval (Carpineto and Romano 2012), improved representation of case metadata (e.g., precedents (Mentzingen et al. 2023), holdings (Zheng et al. 2021), and entities (Leitner et al. 2019)), deepening interaction between the retriever and the generator through joint training (Guu et al. 2020b; Izacard et al. 2022) and with domain- or task-adaptation (Asai et al. 2022; Siriwardhana et al. 2023). We stress, however, that the improvement of RAG systems is an active area of ongoing research, without silver bullets. Ultimately, what is challenging is that current systems present users with open-ended chat bots that accept any kind of legal question. The universe of legal queries is vast and unbounded. Providers should be transparent around the distribution of queries they face, and the evaluation of benchmarking of AI systems may be far easier for tools that accomplish specific, well-defined legal tasks.

9 | Conclusion

AI tools for legal research have not eliminated hallucinations. Users of these tools must continue to verify that key propositions are accurately supported by citations.

The most important implication of our results is the need for rigorous, transparent benchmarking and public evaluations of AI tools in law. In other AI domains, benchmarks such as the Massive Multitask Language Understanding (Hendrycks et al. 2020) and BIG Bench Hard (BIG-bench Authors 2023; Suzgun et al. 2023) have been central to developing a common understanding of progress and limitations in the field. But in contrast to even GPT-4—not to mention open-source systems like Llama and Mistral—legal AI tools provide no systematic access, publish few details about models, and report no benchmarking results at all. This stands in marked contrast to the general AI field (Liang et al. 2023) and makes responsible integration, supervision, and oversight acutely difficult.

Vendor-driven benchmarking efforts are a step in the right direction, but they are far from adequate. In other domains, such as facial recognition software and AI hiring software, vendor-driven audits and evaluations have had compromised independence and have failed to identify serious accuracy and equity issues (Raji et al. 2022b). Designing and executing credible benchmarks requires the leadership of professional organizations and third parties who are not incentivized to compromise the integrity of the evaluation. For example, In the facial recognition context, NIST's Facial Recognition Vendor Test (FRVT) program plays this role (NIST 2020). No clear arbiter or trustworthy institutional mechanism has yet emerged in the legal AI space.

We note that some well-resourced firms have conducted internal evaluations of products. Paul Weiss, a firm with over $2B in annual revenue, for instance, has conducted an internal evaluation of Harvey, albeit with no published results or quantitative benchmarks (Gottlieb 2024). This itself has distributive implications on AI and the legal profession, as "businesses are looking to well-resourced firms ... to get some understanding of how to use and evaluate the new software" (Gottlieb 2024). If only well-heeled actors can even evaluate the risks of AI systems, claims of functionality (Raji et al. 2022a) and that AI can improve access to justice may be quite overstated (Bommasani et al. 2022; Chien et al. 2024; Perlman 2023; Tan et al. 2023).

That said, even in their current form, these products can offer considerable value to legal researchers compared to traditional keyword search methods or general-purpose AI systems, particularly when used as the first step of legal research rather than the last word. Semantic, meaning-based retrieval of legal documents may be of substantial value independent of how these systems then use those documents to generate statements about the law. The reduction we find in the hallucination rate of legal RAG systems compared to general purpose

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

LLMs is also promising, as is their ability to question faulty premises.

But until vendors provide hard evidence of reliability, claims of hallucination-free legal AI systems will remain, at best, ungrounded.

Acknowledgments

We thank Pablo Arredondo, Bobby Bartlett, Josh Cohen, Mike Dahn, Joe Grundfest, Neel Guha, Sandy Handan- Nader, John Hawkinson, Peter Henderson, Pamela Karlan, Aniket Kesari, Mark Lemley, Michelle Mello, Larry Moore, Julian Morimoto, Arvind Narayanan, Lisa Ouellette, Nate Persily, Deb Raji, Dilara Soylu, Andrea Vallebueno, Diego Zambrano, Lucia Zheng, and participants at the Schwartz Reisman Institute Seminar at the University of Toronto, the Cyperpolicy Center Seminar at Stanford University, the Tech Lunch "n" Learn seminar at Fordham Law School, the Conference on Empirical Legal Studies, the Berkeley Law, Economics, and Social Science Workshop, the speaker series at Apple University, and the faculty workshop at Stanford Law School for helpful comments. Authors have no conflicts to disclose. For transparency, C.D.M. is an advisor to various LLM-related companies both individually and through being an investment advisor at AIX Ventures.

Data Availability Statement

Our dataset of queries, tool outputs, and labels are made available at huggingface.com/reglab. Consistent with emerging understanding of AI benchmarking and leaderboards, we reserve a random sample of 50% of the dataset to guard against potential model memorization and ensure that a proper test set remains available for future evaluation (Haimes et al. 2024; Li and Flanigan 2024)

Endnotes

[1] The following are official statements from Lexis, Casetext, and Thomson Reuters; however, none of them has provided any clear evidence so far to support their claims about the capabilities of their AI-based legal research tools:

Lexis: "Unlike other vendors, however, *Lexis+AI delivers 100% hallucination-free linked legal citations* connected to source documents, grounding those responses in authoritative resources that can be relied upon with confidence." (Wellen 2024b) (emphasis added).

Casetext: "Unlike even the most advanced LLMs, *CoCounsel does not make up facts, or 'hallucinate,'* because we've implemented controls to limit CoCounsel to answering from known, reliable data sources—such as our comprehensive, up-to-date database of case law, statutes, regulations, and codes—or not to answer at all." (Casetext 2023) (emphasis added).

Thomson Reuters: "*We avoid [hallucinations] by relying on the trusted content within Westlaw* and building in checks and balances that ensure our answers are grounded in good law." (Thomson Reuters 2023) (emphasis added). "We've all heard horror stories where generative AI just makes things up. That doesn't work for the legal industry. They have to trust the content that AI serves up. With Ask Practical Law AI, *all the responses are based on the expert resources* of Practical Law." (Thomson Reuters 2024b) (emphasis added)

[2] We ran the queries for Lexis+ AI and Thomson Reuters Ask Practical Law AI in Figure 2 as a test prior to the creation of our benchmark dataset; because our queries for the evaluation presented in this article were preregistered, these two examples are not included in our results below.

[3] Theoretical work has shown that hallucinations must occur at a certain rate for calibrated generative language models, regardless of their architecture, training data quality, or size (Kalai and Vempala 2023).

[4] Other definitions of hallucination could be more relevant in other contexts. For example, future research should examine AI tools for contract analysis or document summarization. For that analysis, it would be more important to study hallucinations with respect to the tool's input prompt, rather than with respect to the general facts of the world. Evaluation standards for such generative AI output, however, are still in flux.

[5] In Section 4.3 below, we discuss how different companies may be using definitions of "hallucination" different from the ones more commonly accepted in the literature or in popular discourse.

[6] This retrieval error likely reflects the similarity in the embedding space between "Judge Luther A. Wilgarten" and the terms "judge" (mentioned 9 times in the 900-some word order) and "William Luther," the plaintiff in the case.

[7] Note that for our false premise questions, the desired behavior is for the model to refute and state the false assumption in the user's prompt. A gold-standard response to such a question would therefore be a statement that the assumption may be incorrect, with a case law citation to the opposite proposition. However, for these false premise questions alone, we also label a refusal which mentions the fact that no pertinent sources were found as correct.

[8] Of course, there is some evidence that Lexis+ AI does not succeed even by this metric. McGreel (2024) reports instances of Lexis+ AI citing cases decided in 2025.

[9] *Reynolds* even appears in the citation list with a positive Shepardization symbol.

[10] As Gottlieb (2024) reports in one the assessment by law firms of generative AI products, "The importance of reviewing and verifying the accuracy of the output, including checking the AI's answers against other sources, makes any efficiency gains difficult to measure."

[11] LexisNexis is owned by the RELX Group.

[12] Since the completion of our evaluation for this paper in April 2024, LexisNexis has released a "second generation" version of its tool. Our results do not speak to the performance of this second generation product, if different. Accompanying this release, LexisNexis noted, "our promise is not perfection, but that all linked legal citations are hallucination-free" (LexisNexis 2024).

[13] Thomson Reuters denied three requests for access by our team at the time we conducted our initial evaluation. The company provided access after the initial release of our results.

[14] The home page of Practical Law is titled "Practical Law US—Westlaw" and is located on a subdomain of westlaw.com (Google 2024). See also, for example, Berkeley Law School (2024) (noting that "Ask Practical Law AI" is now available on "Westlaw"); Yale Law School (2024) (describing "Ask Practical Law AI" as a Westlaw product); University of Washington (2024) (describing "Practic[al] Law [a]s a database within Westlaw"); Suffolk University (2023) (noting "Ask Practical Law AI (Westlaw)"); Campbell (2024) (writing that "Westlaw released Ask Practical Law AI to academic accounts").

[15] We did not run any preregistered query against any tool prior to registration, with one exception, `changes-in-law-73` ("When does the undue burden standard apply in abortion cases?"). Some queries were slightly rephrased during evaluation to better elicit an answer with factual content (a prospect explicitly contemplated by the preregistration). Our hallucination rates are substantially the same when excluding the small number of queries that were edited from the preregistered prompts. Those queries are marked as such in our released dataset. For more detail, see our documentation in Appendix B.1.

[16] We created a new "conversation" for each query.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[17] In updating results to include AI-AR, we also conducted another round of validation of every hallucination coding. This validation led to nearly identical results—for instance, the accuracy rate of Ask Practical Law AI in Figure 1 increased from 19% to 20%, which is of course within the bounds of inter-rater reliability.

[18] This is based on a simple word count separating based on space.

[19] The number of examples reported are roughly proportional to the relative hallucination rates between tools.

[20] Per Westlaw, a red flag indicates that a case "is no longer good law for at least one of the points of law it contains" (Thomson Reuters 2019). In our labelled sample, we were not able to observe such cases being cited, though they were sometimes discussed without citation.

[21] State ex rel. Leonardi v. Sherry, No. ED 82789, 2003 WL 21384384, at *1 (Mo. Ct. App. June 17, 2003).

[22] See State ex rel. Leonardi v. Sherry, 137 S.W.3d 462, 472 (Mo. 2004) ("The dissenting opinion apparently would cling to the inefficient and wasteful need for a second trial at law if equity 'fails of establishment' in the initial request for equitable relief.").

[23] Chen et al. (2024) consider the possibility of retrievable documents that contain false information. However, its evaluation focuses on a significantly simplified setting that is not applicable to the complexity of legal use cases.

[24] Even AI-Assisted Research was exclusively available to law firms when we initially conducted the evaluation of Lexis+ AI and Ask Practical Law AI (Thomson Reuters 2023).

[25] Indeed, even presenting the same query to these models may yield different answers each time, as the text decoding process may not be set to be deterministic (e.g., via the temperature parameter). GPT-4, for instance, is known not to be deterministic. It is also not clear what retrieval parameters (e.g., similarity threshold or top-k value) are used, impeding consistent analysis of the model.

[26] For instance, OpenAI appeared to patch its system to prevent adversarial attacks with specific suffixes discovered in Zou, Wang, et al. (2023), but the underlying vulnerability may still persist. As one of the authors of that study noted, "Companies like OpenAI have just patched the suffixes in the paper, but numerous other prompts acquired during training remain effective. Moreover, if the model weights are updated, repeating the same procedure on the new model would likely still work."

[27] See, for example, §2.2 of the LexisNexis Terms of Service (LexisNexis 2023a), which prohibits programmatic access.

[28] In contrast to Dahl et al. (2024), which relied on an external knowledge base to validate the answer (not groundedness), our setting is not amenable to automated evaluation for two reasons. First, groundedness depends on a legally informed judgment about the weight of the underlying text. As we note above, LLMs are not currently capable of distinguishing between the holding and dicta of a case, for instance, leading it to erroneously deem a proposition as "grounded" in dicta. Second, in many circumstances, the reason for the error was not in the cited document. Another case may have superseded the cited one; the underlying law may have changed; and the cited document may not be controlling. As a result, we deemed it important to conduct a manual verification.

References

Agrawal, A., M. Suzgun, L. Mackey, and A. T. Kalai. 2023. "Do Language Models Know When They're Hallucinating References?" https://doi.org/10.48550/arXiv.2305.18248.

Ambrogi, B. 2023. "Major Thomson Reuters News: Westlaw Gets Generative AI Research Plus Integration With Casetext CoCounsel." Gen AI Coming Soon to Practical Law. https://www.lawnext.com/2023/11/major-thomson-reuters-news-westlaw-gets-generative-ai-research-plus-casetext-integration-gen-ai-coming-soon-to-practical-law.html.

Ambrogi, B. 2024. "LawNext: Thomson Reuters' AI Strategy for Legal, With Mike Dahn, Head of Westlaw, and Joel Hron." Head of AI. https://www.lawnext.com/2024/02/lawnext-thomson-reuters-ai-strategy-for-legal-with-mike-dahn-head-of-westlaw-and-joel-hron-head-of-ai.html.

American Bar Association. 2018. "Alphabetical List of Jurisdictions Adopting Model Rules." https://www.americanbar.org/groups/professional%5C_responsibility/publications/model%5C_rules%5C_of%5C_professional%5C_conduct/alpha%5C_list%5C_state%5C_adopting%5C_model%5C_rules/.

Arewa, O. B. 2006. "Open Access in a Closed Universe: Lexis, Westlaw, Law Schools, and the Legal Information Market." *Lewis and Clark Law Review* 10, no. 4: 797–840.

Asai, A., T. Schick, P. Lewis, et al. 2022. "Task-Aware Retrieval With Instructions." *arXiv preprint arXiv:2211.09260*.

Atleson, M. 2023. "Keep Your AI Claims in Check." https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check.

Avery, J. J., P. S. Abril, and A. del Riego. 2023. "ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI." *Yale Journal of Law and Technology* 26, no. 1: 64–129.

BARBRI Inc. 2013. Multistate Testing Practice Questions.

Barnett, S., S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. 2024. "Seven Failure Points When Engineering a Retrieval Augmented Generation System." https://doi.org/10.48550/arXiv.2401.05856.

Beim, D., and K. Rader. 2019. "Legal Uniformity in American Courts." *Journal of Empirical Legal Studies* 16, no. 3: 448–478. https://doi.org/10.1111/jels.12224.

Belkin, N. J. 2008. "Some (What) Grand Challenges for Information Retrieval." *ACM SIGIR Forum* 42, no. 1: 47–54.

Berkeley Law School. 2024. "Generative AI Resources for Berkeley Law Faculty & Staff." https://www.law.berkeley.edu/library/legal-research/chatgpt/.

Berring, R. C. 1986. "Full-Text Databases and Legal Research: Backing Into the Future." *High Technology Law Journal* 1, no. 1: 27–60.

BIG-bench Authors. 2023. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj.

Black, R. C., and J. F. Spriggs II. 2013. "The Citation and Depreciation of U.S. Supreme Court Precedent." *Journal of Empirical Legal Studies* 10, no. 2: 325–358. https://doi.org/10.1111/jels.12012.

Bommasani, R., D. A. Hudson, E. Adeli, et al. 2022. "On the Opportunities and Risks of Foundation Models." https://doi.org/10.48550/arXiv.2108.07258.

Brown, N. 2023. "Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation." *Journal of Free Speech Law* 3, no. 2: 389–424.

Campbell, E. 2024. "Resources for Exploring the Benefits and Drawbacks of AI." https://library.law.unc.edu/2024/02/resources-for-exploring-the-benefits-and-drawbacks-of-ai/.

Carpineto, C., and G. Romano. 2012. "A Survey of Automatic Query Expansion in Information Retrieval." *Acm Computing Surveys (CSUR)* 44, no. 1: 1–50.

Casetext. 2023. "GPT-4 Alone Is Not a Reliable Legal Solution—but It Does Enable One: CoCounsel Harnesses GPT-4's Power to Deliver Results That Legal Professionals Can Rely On." https://casetext.com/blog/cocounsel-harnesses-gpt-4s-power-to-deliver-results-that-legal-professionals-can-rely-on/.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Chen, J., H. Lin, X. Han, and L. Sun. 2024. "Benchmarking Large Language Models in Retrieval-Augmented Generation." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 1616: 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728.

Chen, L., M. Zaharia, and J. Zou. 2023. "How Is Chatgpt's Behavior Changing Over Time?" *arXiv preprint arXiv:2307.09009*.

Chien, C. V., and M. Kim. 2024. "Generative AI and Legal Aid: Results From a Field Study and 100 Use Cases to Bridge the Access to Justice Gap." *Loyola of Los Angeles Law Review* 57: 903.

Chien, C. V., M. Kim, R. Akhil, and R. Rathish. 2024. *How Generative AI Can Help Address the Access to Justice Gap Through the Courts*. Loyola of Los Angeles Law Review.

Choi, J. H., A. Monahan, and D. Schwarcz. 2024. "Lawyering in the Age of Artificial Intelligence." *Minnesota Law Review* 109: 147. https://doi.org/10.2139/ssrn.4626276.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20, no. 1: 37–46. https://doi.org/10.1177/001316446002000104.

Collens, J., R. Reimer, G. Schifman, and P. Wilkinson. 2024. "AI Survey: Where Artificial Intelligence Stands in the Legal Industry." https://assets.law360news.com/1826000/1826128/law360_pulse-ai_survey.pdf.

Cuconasu, F., G. Trappolini, F. Siciliano, et al. 2024. "The Power of Noise: Redefining Retrieval for RAG Systems." https://doi.org/10.1145/3626772.3657834.

Cyphert, A. B. 2021. "A Human Being Wrote This Law Review Article: GPT-3 and the Practice of Law." *UC Davis Law Review* 55, no. 1: 401–444.

Dahl, M., V. Magesh, M. Suzgun, and D. E. Ho. 2024. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." *Journal of Legal Analaysis* 16, no. 1: 64–93.

Dell'Acqua, F., E. McFowland, E. R. Mollick, et al. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." https://doi.org/10.2139/ssrn.4573321.

Deng, C., Y. Zhao, X. Tang, M. Gerstein, and A. Cohan. 2023. "Investigating Data Contamination in Modern Benchmarks for Large Language Models." *arXiv preprint arXiv:2311.09783*.

Diamantis, M. E. 2023. "Vicarious Liability for AI." *Indiana Law Journal* 99, no. 1: 317–334.

Dworkin, R. 1986. *Law's Empire*. Harvard University Press.

Fowler, J. H., T. R. Johnson, J. F. Spriggs II, S. Jeon, and P. J. Wahlbeck. 2007. "Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court." *Political Analysis* 15, no. 3: 324–346. https://doi.org/10.1093/pan/mpm011.

Free Law Project. 2024. "Courtlistener." courtlistener.com.

Gao, Y., Y. Xiong, X. Gao, et al. 2024. "Retrieval-Augmented Generation for Large Language Models: A Survey." https://doi.org/10.48550/arXiv.2312.10997.

Goddard, K., A. Roudsari, and J. C. Wyatt. 2012. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association: JAMIA* 19, no. 1: 121–127. https://doi.org/10.1136/amiajnl-2011-000089.

Google. 2024. "Google Search Results for 'Practical Law'." Accessed May 05, 2024. https://web.archive.org/web/20240522191224/https://www.google.com/search?q=practical+law.

Gottlieb, I. 2024. *Paul Weiss Assessing Value of AI*. But Not Yet on Bottom Line.

Greenhill, S. 2024. "Lawyers Cross Into the New Era of Generative AI." https://www.lexisnexis.co.uk/insights/lawyers-cross-into-the-new-era-of-generative-ai/index.html.

Grewal, G. S. 2024. "Remarks at Program on Corporate Compliance and Enforcement Spring Conference 2024." https://www.sec.gov/news/speech/gurbir-remarks-pcce-041524.

Guha, N., J. Nyarko, D. E. Ho, et al. 2023. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." https://doi.org/10.48550/arXiv.2308.11462.

Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang. 2020b. "Retrieval Augmented Language Model Pre-Training." In *International Conference on Machine Learning*, 3929–3938. PMLR.

Guu, K., K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. 2020a. "REALM: Retrieval-Augmented Language Model Pre-Training." In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 3929–3938. PMLR.

Haimes, J., C. Wenner, K. Thaman, et al. 2024. "Benchmark Inflation: Revealing Llm Performance Gaps Using Retro-Holdouts." *arXiv:2410.09247*. https://doi.org/10.48550/arXiv.2410.09247.

Hashimoto, T. B., H. Zhang, and P. Liang. 2019. "Unifying Human and Statistical Evaluation for Natural Language Generation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio, 1689–1701. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1169.

Henderson, P., M. S. Krass, L. Zheng, et al. 2022. "Pile of Law: Learning Responsible Data Filtering From the Law and a 256GB." Open-Source Legal Dataset. https://doi.org/10.48550/arXiv.2207.00220.

Hendrycks, D., C. Burns, S. Basart, et al. 2020. "Measuring Massive Multitask Language Understanding." In *International Conference on Learning Representations*. https://openreview.net/forum?id=uyTL5Bvosj.

Henry, J. 2024. "We Asked Every am Law 100 Law Firm How They're Using Gen AI. Here's What we Learned." The American Lawyer. https://www.law.com/americanlawyer/2024/01/29/we-asked-every-am-law-100-firm-how-theyre-using-gen-ai-heres-what-we-learned/.

Izacard, G., P. Lewis, M. Lomeli, et al. 2022. "Few-Shot Learning With Retrieval Augmented Language Models." *arXiv preprint arXiv:2208.03299*, 2 (3).

Ji, Z., N. Lee, R. Frieske, et al. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no. 12: 248:1–248:38. https://doi.org/10.1145/3571730.

Kalai, A. T., and S. S. Vempala. 2023. "Calibrated Language Models Must Hallucinate." https://doi.org/10.48550/arXiv.2311.14648.

Kapoor, S., P. Henderson, and A. Narayanan. 2024. "Promises and Pitfalls of Artificial Intelligence for Legal Applications." *Journal of Cross-Disciplinary Research in Computational Law* 2, no. 2. https://journalcrcl.org/crcl/article/view/62.

Karpukhin, V., B. Oguz, S. Min, et al. 2020. "Dense Passage Retrieval for Open-Domain Question Answering." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by B. Webber, T. Cohn, Y. He, and Y. Liu, 6769–6781. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550.

Kite-Jackson, D. W. 2023. *2023 Artificial Intelligence (AI) TechReport (Tech. Rep.)*. American Bar Association. https://www.americanbar.org/groups/law_practice/resources/tech-report/2023/2023-artificial-intelligence-ai-techreport/.

Krieger, S. H., and K. F. Kuh. 2014. "Accessing Law: An Empirical Study Exploring the Influence of Legal Research Medium." (2456679). https://papers.ssrn.com/abstract=2456679.

Law360. 2024. "Tracking Federal Judge Orders on Artificial Intelligence." https://www.law360.com/pulse/ai-tracker.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Leitner, E., G. Rehm, and J. Moreno-Schneider. 2019. "Fine-Grained Named Entity Recognition in Legal Documents." In *Semantic Systems. the Power of Ai and Knowledge Graphs*, edited by M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, and Y. Sure-Vetter, 272–287. Springer International Publishing.

Lewis, P., E. Perez, A. Piktus, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33: 9459–9474.

LexisNexis. 2023a. "General Terms and Conditions." https://www.lexisnexis.com/en-us/terms/general/default.page.

LexisNexis. 2023b. "LexisNexis Launches Lexis+ AI, a Generative AI Solution With Linked Hallucination-Free Legal Citations." https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-launches-lexis-ai-a-generative-ai-solution-with-hallucination-free-linked-legal-citations.

LexisNexis. 2024. "LexisNexis Launches Second-Generation Legal AI Assistant on Lexis+ AI." https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-launches-second-generation-legal-ai-assistant-on-lexis-ai.

Li, C., and J. Flanigan. 2024. "Task Contamination: Language Models May Not Be Few-Shot Anymore." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 16: 18471–18480. https://doi.org/10.1609/aaai.v38i16.29808.

Liang, P., R. Bommasani, T. Lee, et al. 2023. "Holistic Evaluation of Language Models." https://doi.org/10.48550/arXiv.2211.09110.

Lissner, M. 2022. "Important Opinions on Courtlistener Are Now Summarized by the Top Experts—Judges." https://free.law/2022/03/17/summarizing-important-cases.

Liu, C.-W., R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 2016. "How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, edited by J. Su, K. Duh, and X. Carreras, 2122–2132. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1230.

Ma, M., A. Sinha, A. Tandon, and J. Richards. 2024. "Generative AI Legal Landscape 2024." (Tech. Rep.). https://law.stanford.edu/publications/generative-ai-legal-landscape-2024-2/.

Magnanelli, N. 2024. "The Legal Tech Bro Blues: Generative AI, Legal Indeterminacy, and the Future of Legal Research and Writing." *Georgetown Law Technology Review* 8, no. 2: 300–315.

Manakul, P., A. Liusie, and M. J. F. Gales. 2023. "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." https://doi.org/10.48550/arXiv.2303.08896.

Markelius, A., C. Wright, J. Kuiper, N. Delille, and Y.-T. Kuo. 2024. "The Mechanisms of AI Hype and Its Planetary and Social Costs." *AI and Ethics* 4: 727–742. https://doi.org/10.1007/s43681-024-00461-2.

Mart, S. 2018. "Results May Vary." ABA Journal. https://scholar.law.colorado.edu/faculty-articles/964.

Martínez, E. 2024. "Re-Evaluating GPT-4's Bar Exam Performance." *Artificial Intelligence and Law* 1–24. https://doi.org/10.1007/s10506-024-09396-9.

McGreel, P. 2024. "I Asked Lexas+ AI [Sic] a Simple Question: "What Cases Have Applied Students for Fair Admissions, Inc. v. Harvard College to the Use of Race in Government Decisionmaking?" This Screenshot Has the Answer I Received. Here Are Some of the (Serious) Problems With This Answer." [Twitter]. https://x.com/conlawgeek/status/1762561111725899859.

McIntosh, T. R., T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge. 2024. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence.

Mentzingen, H., N. António, and F. Bacao. 2023. "Automation of Legal Precedents Retrieval: Findings From a Literature Review." *International Journal of Intelligent Systems* 2023, no. 1: 6660983. https://doi.org/10.1155/2023/6660983.

Mik, E. 2024. "Caveat Lector: Large Language Models in Legal Practice." https://papers.ssrn.com/abstract=4757452.

Miner, R. J. 1989. *Remarks: Clerks of Judge Luther A*. Wilgarten, Jr.

Mündler, N., J. He, S. Jenko, and M. Vechev. 2023. "Self-Contradictory Hallucinations of Lage Language Models: Evaluation." Detection and Mitigation. https://doi.org/10.48550/arXiv.2305.15852.

NIST. 2020. "Face Recognition Vendor Test (FRVT)." https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt.

Oren, Y., N. Meister, N. S. Chatterji, F. Ladhak, and T. Hashimoto. 2023. "Proving Test Set Contamination for Black-Box Language Models." In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=KS8mIvetg2.

Östling, A., H. Sargeant, H. Xie, et al. 2024. "The Cambridge Law Corpus: A Dataset for Legal AI Research." https://doi.org/10.48550/arXiv.2309.12269.

Perlman, A. 2023. "The Implications of ChatGPT for Legal Services and Society." The Practice, (March/April).

Raji, I. D., E. Denton, E. M. Bender, A. Hanna, and A. Paullada. 2021. "AI and the Everything in the Whole Wide World Benchmark." Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Raji, I. D., I. E. Kumar, A. Horowitz, and A. Selbst. 2022a. "The Fallacy of Ai Functionality. Proceedings of the 2022 ACM Conference on Fairness." *Accountability, and Transparency, Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, June 20, 2022, 959–972. Association for Computing Machinery. https://doi.org/10.1145/3531146.3533158.

Raji, I. D., P. Xu, C. Honigsberg, and D. E. Ho. 2022b. "Outsider Oversight: Designing a Third Party Audit Ecosystem for Ai Governance." *arXiv:2206.04737* [cs]. http://arxiv.org/abs/2206.04737.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." https://doi.org/10.48550/arXiv.1606.05250.

Reuters, T. 2019. "Westlaw Tip of the Week: Checking Cases With Keycite." https://legal.thomsonreuters.com/blog/westlaw-tip-of-the-week-checking-cases-with-keycite/.

Rivers, C. C. 2024. "Moffatt v. Air Canada." https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html.

Roberts, J. G. 2023. "2023 Year-End Report on the Federal Judiciary." (tech. rep.). https://www.uscourts.gov/news/2023/12/31/chief-justice-roberts-issues-2023-year-end-report.

Salib, P. 2024. *AI Outputs Are Not Protected Speech*. Washington University Law Review. https://doi.org/10.2139/ssrn.4687558.

Schwarcz, D., and J. H. Choi. 2023. "AI Tools for Lawyers: A Practical Guide." *Minnesota Law Review Headnotes* 108: 1–39. https://scholarship.law.umn.edu/faculty_articles/1048/.

Securities and Exchange Commission. 2024. "Sec Charges Two Investment Advisers With Making False and Misleading Statements About Their Use of Artificial Intelligence." https://www.sec.gov/news/press-release/2024-36.

Sharma, M., M. Tong, T. Korbak, et al. 2023. "Towards Understanding Sycophancy in Language Models." https://doi.org/10.48550/arXiv.2310.13548.

Siriwardhana, S., R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara. 2023. "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Question Answering." *Transactions of the Association for Computational Linguistics* 11: 1–17. https://doi.org/10.1162/tacl_a_00530.

Smith, E., O. Hsu, R. Qian, S. Roller, Y.-L. Boureau, and J. Weston. 2022. "Human Evaluation of Conversations Is an Open Problem: Comparing the Sensitivity of Various Methods for Evaluating Dialogue Agents." In *Proceedings of the 4th Workshop on NLP for Conversational AI (Pp. 77–97)*, edited by B. Liu, A. Papangelis, S. Ultes, et al. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nlp4convai-1.8.

Suffolk University. 2023. "Law Faculty Guide to Artificial Intelligence: Practical Law AI (Westlaw)." https://lawguides.suffolk.edu/AIFac/PracticalLaw.

Surani, F., M. Dahl, and V. Magesh. 2024. "Legal RAG Hallucinations." https://osf.io/etzp2.

Suzgun, M., N. Scales, N. Schärli, et al. 2023. "Challenging BIG-Bench Tasks and Whether Chain-Of-Thought Can Solve Them." In *Findings of the Association for Computational Linguistics: Acl 2023*, edited by A. Rogers, J. Boyd-Graber, and N. Okazaki, 13003–13051. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.824.

Tan, J., H. Westermann, and K. Benyekhlef. 2023. "ChatGPT as an Artificial Lawyer?" Proceedings of the ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice.

Task Force on Artificial Intelligence. 2024. "Report and Recommendations of the New York State Bar Association Task Force on Artificial Intelligence" (tech. rep. New York State Bar Association). https://nysba.org/app/uploads/2024/04/Task-Force-on-AI-Report-draft-2024-04-02-FINAL.pdf.

The Florida Bar. 2024. "Florida Bar Ethics Opinion," (tech. rep. No. 24–1. The Florida Bar). https://www.floridabar.org/etopinions/opinion-24-1/.

The Library Innovation Lab. 2024. Transitions for the Caselaw Access Project.

The State Bar of California. 2023. "Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law" (Tech. Rep. The State Bar of California). https://www.calbar.ca.gov/Portals/0/documents/ethics/Generative-AI-Practical-Guidance.pdf.

Thomson Reuters. 2023. "Introducing AI-Assisted Research: Legal Research Meets Generative AI." https://legal.thomsonreuters.com/blog/legal-research-meets-generative-ai/.

Thomson Reuters. 2024a. "Accelerate How You Find Answers so You Can Practice With Confidence." https://legal.thomsonreuters.com/en/products/practical-law.

Thomson Reuters. 2024b. "Introducing Ask Practical Law AI on Practical Law: Generative AI Meets Legal How-To." https://legal.thomsonreuters.com/blog/legal-how-to-meets-generative-ai/.

University of Washington. 2024. "Artificial Intelligence." University of Washington Gallagher law Library. https://lib.law.uw.edu/c.php?g=1387996&p=10265405.

van der Merwe, M., K. Ramakrishnan, and M. Anderljung. 2024. "Tort Law and Frontier AI Governance" (tech. rep. Lawfare).

Volokh, E. 2023. "Large Libel Models? Liability for AI Output." *Journal of Free Speech Law* 3, no. 2: 489–558.

Waldron, J. 2002. "Is the Rule of Law an Essentially Contested Concept (In Florida)?" *Law and Philosophy* 21, no. 2: 137–164. https://doi.org/10.1023/A:1014513930336.

Walters, E. 2019. "The Model Rules of Autonomous Conduct: Ethical Responsibilities of Lawyers and Artificial Intelligence." *Georgia State University Law Review* 35, no. 4: 1073–1092.

Weiser, B. 2023. "Here's What Happens When Your Lawyer Uses ChatGPT." *New York Times*. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

Weiser, B., and J. E. Bromwich. 2023. "Michael Cohen Used Artificial Intelligence in Feeding Lawyer Bogus Cases." The New York Times. https://www.nytimes.com/2023/12/29/nyregion/michael-cohen-ai-fake-cases.html.

Wellen, S. 2024a. "Tech Innovation With LLMs Producing More Secure and Reliable Gen AI Results." https://www.lexisnexis.com/community/insights/legal/b/thought-leadership/posts/tech-innovation-with-llms-producing-more-secure-and-reliable-gen-ai-results.

Wellen, S. 2024b. "How Lexis+ AI Delivers Hallucination-Free Linked Legal Citations." https://www.lexisnexis.com/community/insights/legal/b/product-features/posts/how-lexis-ai-delivers-hallucination-free-linked-legal-citations.

Wills, P. 2025. "Care for Chatbots." UBC Law Review. https://www.ubclawreview.ca/special-issue-ai-the-legal-profession.

Yale Law School. 2024. "Lexis and Westlaw Generative AI Products." Yale Law Library, Lillian Goldman Law Library. https://library.law.yale.edu/news/lexis-and-westlaw-generative-ai-products.

Yamane, N. 2020. "Artificial Intelligence in the Legal Field and the Indispensable Human Element Legal Ethics Demands." *Georgetown Journal of Legal Ethics* 33, no. 3: 877–890.

Yang, Z., P. Qi, S. Zhang, et al. 2018. "HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering." https://doi.org/10.48550/arXiv.1809.09600.

Zheng, L., W.-L. Chiang, Y. Sheng, et al. 2023. "Judging LLM-As-a-Judge With MT-Bench and Chatbot Arena." Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Zheng, L., N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho. 2021. "When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the Casehold Dataset of 53,000+ Legal Holdings." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, July 27, 2021, 159–168. International Conference on Artificial Intelligence and Law ACM Digital Library.

Zou, A., L. Phan, S. Chen, et al. 2023a. Representation Engineering: A Top-Down Approach to AI Transparency.

Zou, A., Z. Wang, J. Z. Kolter, and M. Fredrikson. 2023b. "Universal and Transferable Adversarial Attacks on Aligned Language Models." *arXiv preprint arXiv:2307.15043*.

Appendix A

Complete Query Descriptions

General Legal Research

Multistate Bar Exam

Description Questions from the multiple-choice multistate bar exam, reformatted as open-ended questions (i.e., no response choices given).

# of Queries in Dataset 20

Example Arnold decided to destroy an old warehouse that he owned because the taxes on the structure exceeded the income that he could receive from it. He crept into the building in the middle of the night with a can of gasoline and a fuse and set the fuse timer for 30 min. He then left the building. The fuse failed to ignite, and the building was not harmed. Arson is defined in this jurisdiction as "The intentional burning of any building or structure of another, without the consent of the owner." Arnold believed, however, that burning one's own building was arson, having been so advised by his lawyer. Has Arnold committed attempted arson?

Source BARBRI practice bar exam questions (BARBRI Inc. 2013).

Evaluation Reference BARBRI answer key.

Rule QA

Description Questions asking the model to describe a well-established legal rule. These rules sometimes represent the kind of legal "background knowledge" that does not always require a citation to a specific

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

case. Other rules are tied to a specific civil or criminal statute. They are also the kind of question that a lawyer may ask when learning about a new area of the law, and the kind of question that is not easy to keyword-search.

# of Queries in Dataset 20

Example What are the four fair use factors?

Source Rule QA task in LegalBench (Guha et al. 2023).

Evaluation Reference LegalBench answer key.

Treatment (Doctrinal Agreement)

Description Questions about how one Supreme Court case treated another Supreme Court case that it cites.

# of Queries in Dataset 20

Example How did Nassau Smelting & Refining Works Ltd., v. United States, 266 U.S. 101 (1924) treat United States v. Pfirsch, 256 U.S. 547 (1924)?

Source Entries in a Shepard's Citations dataset for the Supreme Court (Black and Spriggs II. 2013; Fowler et al. 2007).

Evaluation Reference Whether the model correctly characterizes the treatment of the cited case, for example, as "followed," "distinguished," "overruled," and so forth.

Doctrine Test

Description Questions asking the model to define a well-known legal doctrine taught in standard black-letter courses like contracts, evidence, procedure, or statutory interpretation.

# of Queries in Dataset 10

Example What is the near miss doctrine?

Source Hand-curated.

Evaluation Reference Our own domain knowledge.

Question With Irrelevant Context

Description The Doctrine Test questions, but with some irrelevant context prepended, which is not related to the questions and which the model is expected to ignore.

# of Queries in Dataset 10

Example Escheat is the passing of an interest in land to the state when a decedent has no will, no heirs, or devisees. In the United States, escheat rights are governed by the laws of each state. Probate is usually used to determine escheat rights. What is the near miss doctrine?

Source We selected arbitrary definitions from Black's Law Dictionary and appended them to our doctrine test questions.

Evaluation Reference Our own domain knowledge.

Jurisdiction or Time-Specific

SCALR

Description Questions presented in Supreme Court cases decided between 2000 and 2019. The questions are slightly rephrased to be suitable to ask an LLM. The task measures whether the AI system correctly identifies legal standards after recent changes in law (which typically take place when a Supreme Court case is decided). Unlike the LegalBench version of this task, which is multiple-choice for easier evaluation, this is presented as an open-ended task.

# of Queries in Dataset 30

Example Did Congress divest the federal district courts of their federal-question jurisdiction under 28 U.S.C. § 1331 over private actions brought under the Telephone Consumer Protection Act?

Source SCALR task in LegalBench (derived from the questions presented hosted on the Supreme Court's website) (Guha et al. 2023).

Evaluation Reference LegalBench answer key containing a holding statement describing the relevant SCOTUS case. Evaluators may also refer to Oyez, or check for any overruled cases if relevant.

Circuit Splits

Description Questions testing whether the model correctly identifies the law in a specific circuit on a legal question that circuits disagree on.

# of Queries in Dataset 19

Example To prove the "haboring" of undocument immigrants in the Sixth Circuit under 8 U.S.C. § 1324, must the prosecution prove that the defendant acted intentionally?

Source Bloomberg Law articles on current circuit splits.

Evaluation Reference Bloomberg Law articles' descriptions of the rule in the specified court, with checks for overruling cases if any are cited.

Changes in Law

Description Questions about legal issues whose correct answer recently changed because of a recent circuit court ruling.

# of Queries in Dataset 11

Example In the Eighth Circuit, can a private litigant sue under Section 2 of the Voting Rights Act?

Source A variety of news articles on recent changes in law in circuit courts.

Evaluation Reference Circuit opinions and news articles' descriptions of different sides in circuit splits.

Local Rules

Description Questions about the Local Rules of Civil Procedure in various United States District Courts, limited to issues where the Local Rules diverge from the Federal Rules of Civil Procedure.

# of Queries in Dataset 20

Example In the U.S. District Court for the Southern District of Indiana, how many days before serving a Rule 45 subpoena to a non-party must a party first serve notice of the proposed subpoena to the other parties?

Source Hand-curated from websites of federal district courts.

Evaluation Reference Local rules sourced from the websites of federal district courts.

False Premise

False Holding Statements

Description Questions seeking a case in a particular jurisdiction standing for a proposition that is the precise opposite of what a case in that jurisdiction has held.

# of Queries in Dataset 10

Example I'm looking for a case in Colorado which holds that pro se litigants can be awarded attorney's fees.

Source We randomly selected legal documents from CourtListener. We excluded cases that did not have clearly written parenthetical holding statements (Lissner 2022); brief, judge-written summaries of the key holding of the case. We hand-constructed a question by negating the holding statement. We asked the model to identify a case in the relevant jurisdiction that held the opposite of what this real case held.

Evaluation Reference Our query asks for support for a proposition that has been contradicted in a particular jurisdiction. Acceptable responses are that (1) no such case exists, (2) a case to the contrary exists (citing a case

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

similar to the one that our question negates), or (3) such a case does exist, and our evaluation finds that the case in fact does supersede the opinion that our question was drawn from. We did not observe any instances of (3).

Whimsical Holding Statements

Description Questions seeking a case standing for a proposition that is legally preposterous. (The model should respond that no such case exists.)

# of Queries in Dataset 10

Example I'm looking for a case that stands for the proposition that a pedestrian can be charged with theft for absorbing sunlight that would otherwise fall on solar panels, thereby depriving the owner of the panels of potential energy.

Source Hand-written.

Evaluation Reference We ask to support a clearly false proposition. Acceptable responses are that (1) no such cases exists, (2) a case to the contrary exists, or (3) a description of the case supporting the closest reasonable proposition, while explaining that the original proposition is false.

False Judicial Contributions

Description Questions asking about the legal career of Judge Luther A. Wilgarten, a fictional jurist (Miner 1989).

# of Queries in Dataset 2

Example What contributions has Judge Luther A. Wilgarten made to the legal field?

Source Hand-curated, with thanks to Professor Pam Karlan for inspiration.

Evaluation Reference Any described contribution is a hallucination; Judge Wilgarten is fictional.

Factual Recall

Metadata: Citation

Description Questions seeking the well-formatted Bluebook citation of a given case.

# of Queries in Dataset 10

Example Give me a Bluebook-formatted citation for Sears, Roebuck & Co. v. Blade, decided by the California Court of Appeals in 1956.

Source Randomly selected legal documents from CourtListener.

Evaluation Reference Publicly available case data from CourtListener (Free Law Project 2024).

Metadata: Year

Description Questions seeking the year that a given case was decided.

# of Queries in Dataset 10

Example What year was Riley v. Albany Savings Bank, 103 N.Y. 669 (NY) decided?

Source Randomly selected legal documents from CourtListener.

Evaluation Reference Publicly available case data from CourtListener (Free Law Project 2024).

Metadata: Author

Description Questions seeking the author of the majority opinion in a given case.

# of Queries in Dataset 10

Example Who wrote the majority opinion in In Re Bebar, 315 F. Supp. 841 (E.D.N.Y 1970)?

Source Randomly selected legal documents from CourtListener.

Evaluation Reference Publicly available case data from CourtListener (Free Law Project 2024).

Appendix B

Running Queries

We ran queries against Lexis+ AI and Thomson Reuters Practical Law AI by pasting the complete text of each query into the chat box, without system message or other text. We started a new conversation for each query, so no state was preserved. We copied the complete text of each response and pasted it into our records. In-text citations were included in our copy, and we made an effort to copy the list of materials presented after the response, but these were not consistently captured.

Queries Modified After Pre-Registration

During the pre-registration process, we noted that we retain the flexibility to make minor, non-substantive edits to our questions. Any changes that we made to our queries after pre-registration are enumerated here.

`scalr-2` We inserted the word "specific" in the question to more accurately describe the legal distinction drawn by the Supreme Court in the case.

`scalr-9` We inserted the phrase "reasonable probability" in the question to more accurately describe the legal distinction drawn by the Supreme Court in the case.

`changes-in-law-74` We replaced "midwife" with "nurse practitioner" to more accurately capture the effect of the relevant change in law.

`bar-exam-90` The original query was formatted as a fill-in-the-blank ("the defendant's testimony is"), and we rephrased it to be a proper question ("is the defendant's testimony admissible?").

`metadata-citation-130` The original query was mistakenly truncated, and we corrected it to include the court and year, as all the other citation queries do.

`local-rules-191` *to* `local-rules-200` The original questions said, for example, "the Southern District of Indiana," which could be interpreted to refer to state courts in Indiana. The questions were about federal courts, so we edited all of these to say, e.g., "the *U.S. District Court for the* Southern District of Indiana."

The edits we make are minor, made only to clarify a question or to correct an error or a typo in the underlying question. The substantive content of all 202 of our questions remains the same, and our hallucination rates are substantially the same when excluding the small number of questions that were edited from the pre-registered prompt (Table B1).

Appendix C

Per-Task Breakdown

Table C1 reports the number of hallucinations and incomplete responses each model produced for a specific task.

Appendix D

Query Evaluation

The below materials reproduce the annotation criteria we adhered to during the evaluation of queries.

TABLE B1 | Hallucination rates with and without edited questions.

Model	Proportion hallucinated	Proportion hallucinated (w/o Edited Q's)
GPT-4	0.43	0.39
Lexis	0.17	0.16
Pract. Law	0.17	0.15
Westlaw	0.33	0.31

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE C1 | Number of hallucinations and incomplete responses on each task. These subcategories are too small to make meaningful calculation of standard errors. Hallucinations counts are bolded and labeled "Hal.," Incomplete responses are printed in lighter gray and labeled "Inc."

			GPT-4		Lexis		Pract. Law		Westlaw	
Category	**Task**	**N**	**Hal.**	**Inc.**	**Hal.**	**Inc.**	**Hal.**	**Inc.**	**Hal.**	**Inc.**
General legal research	Bar Exam	20	9	2	6	6	9	5	7	2
	Rule QA	20	1	3	0	0	2	1	9	0
	Treatment	20	16	0	8	4	0	20	5	13
	Doctrine Test	10	4	2	1	0	0	7	3	1
	Q. w/Irrelevant Context	10	4	3	1	1	0	9	3	1
Jurisdiction or time specific	SCALR	30	7	2	7	5	5	18	14	1
	Circuit Splits	19	12	1	3	3	7	11	6	2
	Changes in Law	11	9	0	3	0	3	6	6	1
	Local Rules	10	6	0	0	2	3	7	3	6
False premise	False Holdings	10	2	0	2	0	1	8	3	0
	Whimsical Holdings	10	0	0	0	2	0	10	2	1
	False Judicial Contribution	2	0	0	2	0	0	2	0	0
Factual recall	Metadata: Author	10	9	0	0	7	1	9	0	9
	Metadata: Citation	10	6	4	1	4	3	5	4	5
	Metadata: Year	10	2	0	0	3	0	10	1	9
	Total	202	87	17	34	37	34	128	66	51

Instructions

Every response must be coded with two values: correctness and groundedness. Note that Partially Correct answers were collapsed into Correct answers during our final analysis, and Irrelevant/Unhelpful and Stock Refusal answers were collapsed into the category Refusal.

1. Evaluate whether the response is correct, referring to the correctness rubric.
   a. Groundedness is only judged for correct or partially correct responses.
   b. If the correctness is Stock Refusal, Irrelevant/Unhelpful, or Incorrect, then groundedness must be N/A.
   c. If the question is a Bluebook citation question, then groundedness must be Grounded.
   d. If none of the criteria above are met, proceed to step 2.
2. Evaluate groundedness, referring to the groundedness rubric.

Correctness Rubric

For the below labels, we provide example responses to the hypothetical query: *Do law enforcement officers in California have to inform drivers why they are being pulled over?*

Correct

The answer provides accurate information that is fully responsive to the query.

Example: Yes. AB 1542 goes into effect in 2024, which requires California police officers to inform drivers about the reason for the stop ...

Partially Correct

The answer contains no false propositions, but it does not address the substance of the question, or it fails to include a piece of information relevant to the question.

Example: Yes, law enforcement officers in California are generally required to inform drivers why they are being pulled over. This requirement is part of the procedural norms that ensure transparency and fairness... (there is no mention of the relevant CA law).

Irrelevant/Unhelpful

The response contains irrelevant or unhelpful information, not answering the question that is asked. However, it does not contain any false information or statements.

Example: The Fourth Amendment requires law enforcement officers to obtain a warrant prior to entering a suspect's home ...

Stock Refusal

The system provides a rote refusal to answer the question.

Example: The sources provided contain no information relevant to the query.

Incorrect

The response makes any false statement, whether material to the response or not.

Notes on Correctness

Coding False Premise Questions

For false premise questions, a response indicating that no relevant authority could be located is coded as Correct, and not Irrelevant/Unhelpful. However, a stock refusal without any such indication is coded as a Refusal.

- "I cannot provide you with any information on this topic." (Refusal)
- "I cannot find any information on this topic." (Correct)
- "X case held the opposite to the premise presented." (Correct)

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Coding Bluebook Citation Responses

- We are strict Bluebookers. Accept only entirely compliant definitions; missing years, courts, or any information in the Bluebook standard citation is **incorrect**.
- For example, if the parenthetical contains the year but not the court (where the court is required by *The Bluebook*), that is incorrect.
- A citation in which the year is off by one is incorrect

Groundedness Rubric

Grounded

Every legal proposition which is material (i.e., relevant and non-trivial) to the query is supported by an applicable legal source. Indirect support is acceptable; that is, a citation to a document which then cites an applicable document is grounded.

Ungrounded

Every legal proposition which is material (i.e., relevant and non-trivial) to the query requires a citation to a source. If any material proposition is not supported by a citation, the response is ungrounded.

Misgrounded

The system supports a proposition with a source that does not in reality support the proposition.

Fabricated

The answer cites a source that does not exist.

Not Applicable

Only coded when no factual propositions are present; only selected for Irrelevant/Unhelpful and Stock Refusal responses.

Notes on Groundedness

Multiple Propositions, Single Source

- A model may sometimes assert two distinct propositions and cite a single source at the end. If the single source supports both propositions, we consider that **grounded**. However, if both propositions are material to the user's query and only the latter proposition is supported by the source, the response is **ungrounded**.
  - "The Constitution protects the right to interracial marriage. It also protects the right to same-sex marriage. *Obergefell v. Hodges*..."—Grounded, because *Obergefell* includes discussion of *Loving v. Virginia* and its recognition of a right to interracial marriage
  - "The exclusionary rule prevents the admission of unlawfully obtained evidence. The Constitution protects the right to same-sex marriage. *Obergefell v. Hodges* ..."—Ungrounded, because the source supports only the second proposition
- A response can be both ungrounded and misgrounded, for example, if Proposition 1 contains no support and Proposition 2 is incorrectly supported. In this case, the response is labeled with the most serious offense: Misgrounded.

Miscellaneous

- If the primary ("correctness") label of an example is irrelevant or unhelpful, then its secondary ("groundedness") label should be N/A.
- If the primary label of an example is incorrect, then the secondary label should be N/A.

17401461, 2025, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jels.12413 by Columbia University Libraries, Wiley Online Library on [09/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License