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

Varun Magesh* Stanford University

Faiz Surani* Stanford University

Matthew Dahl Yale University

Mirac Suzgun Stanford University

Christopher D. Manning Stanford University

Daniel E. Ho† Stanford University

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. But 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" (Casetext, 2023) or "avoid[ing]" hallucinations (Thomson Reuters, 2023), or guaranteeing "hallucination-free" legal citations (LexisNexis, 2023). 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 and Thomson Reuters each hallucinate more than 17% 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]

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

*Equal contribution.

†Corresponding author: deho@stanford.edu.

[1]Our dataset, tool outputs, and labels will be made available upon publication.

Preprint. Under review.

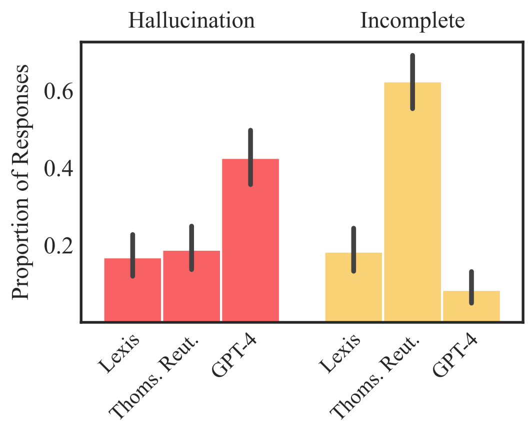

Figure 1: Comparison of hallucinated and incomplete answers across generative legal research tools. Hallucinated responses are those that include false statements or falsely assert a source supports a statement. Incomplete responses are those that fail to either address the user's query or provide proper citations for factual claims.

citing ChatGPT-invented fictional cases in a legal brief (Weiser, 2023); many similar incidents have since been 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 (which owns Westlaw) have claimed to mitigate, if not entirely solve, hallucination risk (LexisNexis, 2023; Casetext, 2023; 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.[2] (We provide details on RAG systems in Section 3.1 below.)

But none of these bold proclamations have been accompanied by empirical evidence which substantiates these claims. 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 and Thomson Reuters, and, for comparison, GPT-4.

Our findings, as 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 left panel of Figure 2, when prompted with a false premise ("Why did Justice Ginsburg dissent in Obergefell?"), Thomson Reuters's system both fails to contradict the mistaken premise and offers an answer based on a statutory provision nowhere implicated in the case. Justice Ginsburg did *not* dissent in *Obergefell v. Hodges*, 576 U.S. 644 (2015) (finding a constitutionally protected right

[2]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, 2024a) (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, 2024) (emphasis added)

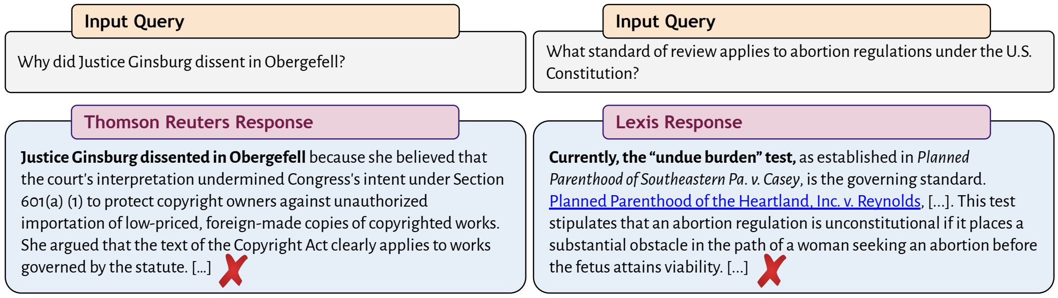

Figure 2: Left: Example of a hallucinated response by Thomson Reuters's Ask Practical Law AI. The system fails to correct the user's mistaken premise—in reality, Justice Ginsburg joined the Court's landmark decision legalizing same-sex marriage—and instead provides additional false information about the case. 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.

to same-sex marriage), and certainly not based on her views of international copyright.[3] We also document substantial differences in the systems: Thomson Reuters's answers are incomplete more than 60% of the time, leading LexisNexis to provide accurate responses at more than three times the rate.

Our article makes four key contributions. First, it conducts 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, Thomson Reuters, 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, our article not only uncovers limitations of current technologies, but also characterizes 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 (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 and wide-ranging. 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 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

[3] We ran the queries 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, the particular examples are not included in our results below.

(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 1,200 lawyers practicing in the United Kingdom, 14% say that they are using generative AI tools weekly or more often (Greenhill, 2024).

However, adoption of these tools is not without risk. 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; 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. 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 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

[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.

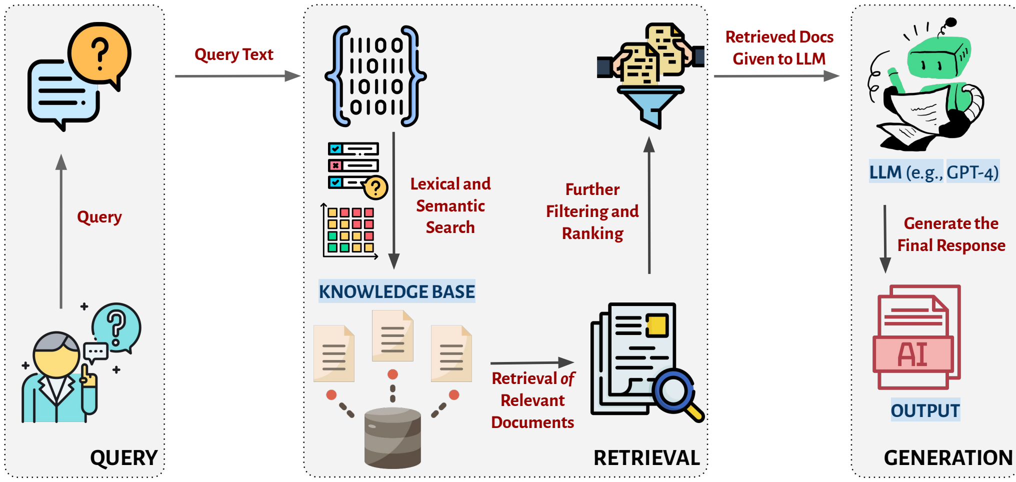

Figure 3: Schematic diagram of a retrieval-augmented generation (RAG) system. Given a user query (left), the typical process consists of two steps: (1) retrieval (middle), where the query is embedded with natural language processing and a retrieval system takes embeddings and retrieves the relevant documents (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 error and hallucinations into the generated response. (Icons are credited to FlatIcon.)

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 their system, RAG "dramatically reduces hallucinations to nearly zero" (Ambrogi, 2024).[5] 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, 2024b).

As depicted in Figure 3, RAG comprises two primary steps to transform a query into a response: (1) retrieval and (2) generation (Lewis et al., 2020; Gao et al., 2024). 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.

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 would generate the answer based on its training data

[5] In that interview, the 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. 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). This conceptual confusion is a problem for those seeking to assess legal AI tools, which is why we offer a clarification of the relationship between hallucinations, groundedness, and accuracy in Section 4.

alone—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., 2020; 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.

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 (Cuconasu et al., 2024; Chen 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.

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 and rules from different pieces of text while keeping the appropriate legal context in mind. For example, consider:

> **Prompt:** What are some notable opinions written by Judge Luther A. Wilgarten?
>
> **Lexis 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]

[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.

	Description	Example
Correctness		
Correct	Response is factually correct and relevant	The right to same sex marriage is protected under the U.S. Constitution. *Obergfell 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. *Obergfell v. Hodges*, 576 U.S. 644 (2015).
Misgrounded	Key factual propositions are cited but misinterpret the source or reference an inapplicable source	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.

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 highlighted in red.

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 misleads the user in 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]

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 they 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 caselaw. If the retrieval system provides documents that are inappropriate to the correct jurisdiction, 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 clarity to the concept of "hallucination," which has been used in different ways in the legal discourse. For example, LexisNexis claims that its AI tool provides "linked hallucination-free legal citations" (LexisNexis, 2023), 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 which links only to *Brown v. Board of Education* on every query has provided "hallucination-free" citations, a plainly irrational result.

More concretely, returning to the *Casey* example in Figure 2, the linked citation *Planned Parenthood v. Reynolds* is a real case that has not been overturned. However, the model's answer relies on *Reynolds*' description of *Planned Parenthood v. Casey*, a case that has been overturned. *Reynolds* even appears in the citation list with a positive Shepardization symbol. 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).

In fact, these errors are potentially more dangerous than fabricating a case outright, because they are subtler and more difficult to spot.[9] Checking for these kinds of hallucinations requires a user to click through to cited references, read and understand the relevant source, assess its authority, and compare

[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]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."

it against the propositions the model seeks to support. Our definition reflects this more complete understanding of "hallucination."

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 two publicly available, RAG-based AI research tools: LexisNexis's Lexis+ AI and Thomson Reuters's Ask Practical Law AI. As nearly every practicing U.S. lawyer knows, Thomson Reuters (the parent company of Westlaw) and LexisNexis 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. 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, 2024a).[10]

Thomson Reuters's Ask Practical Law AI, 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, 2024). 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, 2024), reducing "hallucinations to nearly zero" (Ambrogi, 2024). Thomson Reuters also offers a product called "AI-Assisted Research" that appears to have access to additional primary source material as well (Thomson Reuters, 2023). However, this product is not yet generally available, and multiple requests for access were denied by the company at the time we conducted the evaluation. Both products are made available via the Westlaw platform and are commonly also referred to as AI products within Westlaw.[11] For shorthand, we will refer to it as Thomson Reuters.[12]

[10]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).

[11]The home page of Practical Law is titled "Practical Law US - Westlaw" and located on a subdomain of `westlaw.com`. Google (2024). See also, e.g., Berkeley Law School (2024) (noting that "Ask Practical Law AI" is now available on Westlaw"); Yale Law School (2024) (describing "Ask Practical 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 AI (Westlaw)"); Campbell (2024) (writing that "Westlaw released Ask Practical Law AI to academic accounts").

[12]For transparency, we made this clarification upon request by Westlaw and Thomson Reuters representatives, who consider Westlaw and Practical Law distinct product lines.

Category	Count	Description	Example Query
General legal research	80	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	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	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	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)?

Table 2: The high-level categories of the query dataset.

To provide a point of reference for the quality of these bespoke legal AI tools, 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 (Dell'Acqua et al., 2023; Collens et al., 2024). GPT-4's responses are produced in a "closed-book" setting; that is, produced without RAG.

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
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

Queries in the first category are the paradigmatic use case for these tools, asking general questions of law. Queries in the second and third categories are intended to highlight areas that we expect to be more challenging for legal RAG systems. The last category 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 hand-written 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 known answer (e.g., multiple choice) to the generative setting (Raji et al., 2021; Li and Flanigan, 2024; McIntosh et al., 2024). 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 assessment of the tool's legal analysis and citations.

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

Third, these queries are designed to represent the temporal and jurisdictional variation (e.g., overruled precedents, circuit splits) that is often the subject of live legal research (Beim and Rader, 2019). Our hypothesis is that AI systems may not encode this type of knowledge effectively, 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 et al., 2024; Chien and Kim, 2024; Perlman, 2023; Tan et al., 2023), but contrafactual bias poses 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).[13]

5.3 Query Execution

For Lexis+ AI and Thomson Reuters's Ask Practical Law AI, we executed each query by copying and pasting it into the chat window of each product. 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 were run between March 22 and April 22.

[13] 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 pre-registration); those queries are marked as such in our released dataset and documented in Appendix B.1.

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 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 (Liu et al., 2016; Hashimoto et al., 2019; 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 C.

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).

6 Results

Commercially-available RAG-based legal research tools still hallucinate. Nearly 1 in 5 of our queries caused the tools we tested to respond with misleading or false information. Lexis+ AI and Thomson Reuters's Ask Practical Law AI are less prone to hallucination than GPT-4, but users of these products must remain cautious about relying on their outputs. Below, we show that each tool hallucinates in distinct ways (Section 6.1) and describe the observable RAG failures that we find are driving these top-level results (Section 6.2).

6.1 LexisNexis and Thomson Reuters Hallucinate in Different Ways

Overall hallucination rates are similar between Lexis+ AI and Thomson Reuters's Ask Practical Law AI (Figure 1, page one above), but these top-line results obscure dramatic differences in responsiveness. As shown in Figure 4, Lexis+ AI provides accurate (i.e., correct and grounded) responses on 65% of queries, while Ask Practical Law AI refuses to answer queries 62% of the time and responds accurately just 18% of the time. When looking solely at responsive answers, Thomson Reuters's system hallucinates at a similar rate to GPT-4, and more than twice as often as Lexis+ AI.

Some of these disparities can be explained by the Thomson Reuters system's 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. This limits the system's coverage, as Practical Law documents cover a small fraction of legal topics.

We observe that the limited set of documents causes the model to frequently fail to find relevant information, and sometimes even to hallucinate. These hallucinations often happen when Thomson Reuters's retrieval systems identify text that is tangentially relevant to a query. Take, for example, question `bar-exam-95` in our dataset: a bar exam question on the relationship between a person's knowledge and a charge of accessory to bigamy. The retrieval system finds a document about an unrelated criminal statute (the Foreign Corrupt Practices Act), leading the generation model to apply the wrong legal rule to the fact pattern.

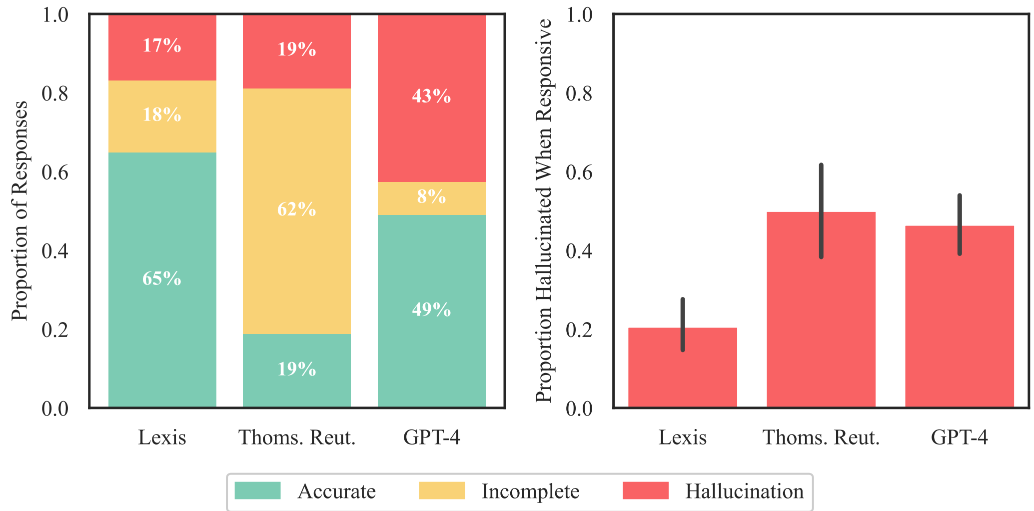

Figure 4: Left panel: overall percentages of accurate, incomplete, and hallucinated responses. Right panel: the percentage of answers that are hallucinated when a direct response is given. Thomson Reuters is the tool most likely to refuse to answer, but the responses that it does produce are not any more trustworthy than those of GPT-4.

On the other hand, Lexis+ AI's retrieval system is connected to a wider body of case law and primary sources. This means that Lexis+ AI has access to all the documents that would be in principle necessary to answer any of our questions. Indeed, Lexis+ AI often produces high quality results. In one instance, it pointed to a false premise in one of our questions. The question `scalr-19` asked whether the six 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.

However, LexisNexis's product is still far from "hallucination-free." Its larger universe of cases sometimes yields irrelevant or less-relevant results, or cases which have facially similar fact patterns but different underlying legal issues. For example, on `scalr-15`, which asks about discharge rules, Lexis cites a case a case which discusses the discharge rules under a different part of the law — text that seems highly similar, on its face, but is actually inapplicable.

Another issue introduced by its larger universe of cases is *indirect* citation, where the model cites to lower courts or different jurisdictions that in turn describe a relevant opinion. While this may be helpful for researchers, who can follow citations and quickly identify the canonical source, these references would be inappropriate for a final, polished legal work. We coded indirect citations to a valid case as correctly grounded because they do help researchers, but it is important to note that more research and reading is generally required in these instances.

6.2 A Typology of Legal RAG Errors

Interpreting why an LLM hallucinates is an open problem (Ji et al., 2023; Zou 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 not 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. Both Lexis+ AI and Thomson Reuters's Ask Practical Law AI show the list of documents which 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.

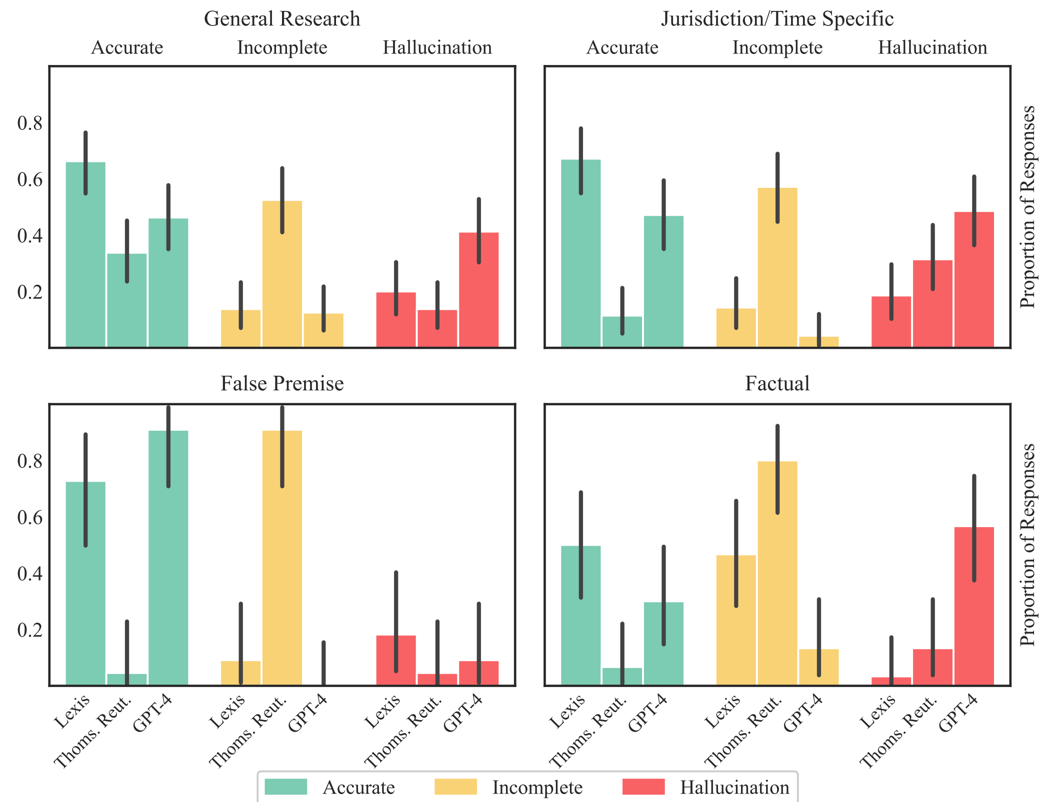

Figure 5: Response evaluations broken down by question category.

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 & applicable, an assumption that is not true in the legal setting (Barnett et al., 2024; Chen et al., 2024).[14] 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 3 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.

Naive retrieval. Many failures in the LexisNexis and Thomson Reuters 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

[14] Chen et al. (2024) considers 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.

Contributing Cause	**Lexis**	**Westlaw**
Naive Retrieval	0.47	0.40
Inapplicable Authority	0.38	0.30
Reasoning Error	0.28	0.50
Sycophancy	0.06	0.00

Table 3: This table shows 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.

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. Thomson Reuters's system retrieves sources discussing motions to dismiss in various contexts such as bankruptcy and patent litigation. But correctly answering this question requires identifying the true underlying issue as being one about *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.

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 (retrieval error), or because the model was given that information and ignored it (reasoning error). 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.

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 and GPT-4 were quite capable at 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. Thomson Reuters's system also hallucinated infrequently in this category, but refused to answer at all in the overwhelming majority of queries.

Reasoning errors. In addition to the more complex behaviors described above, LLM-based systems also tend to make elementary errors of reasoning and fact. Lexis+ AI and Thomson Reuters Ask Practical Law AI are no exception. In one case, Lexis+ AI provided an incorrect answer to a question about class-wide arbitration and then immediately contradicted itself: "The Federal Arbitration Act (FAA) does not prohibit class-action procedures from being superimposed onto an arbitration agreement that does not provide for class-action arbitration. However, the U.S. Supreme Court has ruled that the FAA prohibits arbitrators from imposing class arbitration on parties unless an arbitration agreement specifically provides for it." These reasoning errors are notable in light of open questions around emergent capabilities of LLMs (see, e.g., Bubeck et al., 2023; Schaeffer et al., 2023)

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 two specific products by LexisNexis and Thomson Reuters. The legal AI product space is growing rapidly with many startups (e.g., Harvey, Vincent AI) (Ma et al., 2024). Indeed, Thomson Reuters itself has developed an additional AI product ("AI-Assisted Research") which was exclusively available to law firms at the time we conducted this evaluation (Thomson Reuters, 2023). (Our team's multiple efforts to gain access were turned down by Thomson Reuters at the time we conducted this study.) 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 are likely to improve responses, benchmarking, evaluation, and supervision remain difficult when a model changes over time (Chen et al., 2023), but points to the need for transparency.

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—e.g., 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 reasons for this. In contrast to general-purpose LLMs, which have open models or API access, LexisNexis and Thomson Reuters restrict access to their interfaces.[15] 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, but the fact remains that the AI product space remains quite closed in law.

Last, 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 — and the ecological validity of benchmark tasks – 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 6.2 provides a more granular sense of the limitations of current AI systems.

8 Conclusion

AI tools for legal research have not eliminated hallucinations. Users of these tools must continue verify that key propositions are accurately supported by citations. That said, 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. The reduction we find in the hallucination rate of legal RAG systems compared to general purpose LLMs, for instance, is promising, as is their ability to be able to question faulty premises.

While AI tools show great potential for producing higher-quality legal work faster, their adoption must nonetheless come with caution and principles for ensuring safe use. 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 (Task Force on Artificial Intelligence, 2024, 57). They point lawyers to a list of general-purpose publications and fora that discuss matters related to AI.

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

While it is certainly helpful to learn more about AI, general knowledge is not the same as understanding the trade-offs of specific tools, such as the substantial differences in the responsiveness and accuracy between LexisNexis and Thomson Reuters. Our work shows that the risks and benefits associated with AI-driven legal research tools are different from those associated with chatbots, and that different tools have important differences between them (and even over time within the same tool). Documentation for these tools does not clearly illustrate what they can do for lawyers and in which areas lawyers should exercise caution.

The duty to understand the benefits, risks, and ethical implications of legal research tools requires access to information about the quality of results *of specific tools* in their area of intended use. One commentator, for instance, notes that "[l]awyers must sufficiently understand the AI tools they and others use, including AI's limitations [and] [e]nsure the accuracy and integrity of data and algorithms" (Mack, 2023). The closed nature of these systems, however, makes such supervision difficult in practice. Given the high rate of hallucinations, lawyers may find themselves having to verify each and every proposition and citation, undercutting the claims for efficiency-enhancing use of AI in legal practice.

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., 2022) 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).

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 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.

Until progress on these fronts is made, claims of hallucination-free legal AI systems are, at best, ungrounded.

9 Acknowledgments

We thank Pablo Arredondo, Mike Dahn, Neel Guha, Sandy Handan-Nader, Peter Henderson, Julian Morimoto, Pamela Karlan, Larry Moore, Dilara Soylu, Andrea Vallebueno, and Lucia Zheng for helpful comments.

Authors have no conflicts to disclose. For transparency, CDM is an advisor to various LLM-related companies both individually and through being an investment advisor at AIX Ventures.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do Language Models Know When They're Hallucinating References? *arXiv preprint*.

Bob Ambrogi. 2024. LawNext: Thomson Reuters' AI Strategy for Legal, with Mike Dahn, Head of Westlaw, and Joel Hron, Head of AI.

Olufunmilayo B. Arewa. 2006. Open Access in a Closed Universe: Lexis, Westlaw, Law Schools, and the Legal Information Market. *Lewis & Clark Law Review*, 10(4):797–840.

Joseph J. Avery, Patricia Sánchez Abril, and Alissa del Riego. 2023. ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI. *Yale Journal of Law & Technology*, 26(1):64–129.

BARBRI, Inc. 2013. *Multistate Testing Practice Questions*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven Failure Points When Engineering a Retrieval Augmented Generation System. *arXiv preprint*.

Deborah Beim and Kelly Rader. 2019. Legal Uniformity in American Courts. *Journal of Empirical Legal Studies*, 16(3):448–478.

Nicholas J Belkin. 2008. Some (what) grand challenges for information retrieval. In *ACM SIGIR Forum*, volume 42, pages 47–54. ACM New York, NY, USA.

Berkeley Law School. 2024. Generative AI Resources for Berkeley Law Faculty & Staff.

BIG-bench Authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ryan C. Black and James F. Spriggs, II. 2013. The Citation and Depreciation of U.S. Supreme Court Precedent. *Journal of Empirical Legal Studies*, 10(2):325–358.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. *arXiv preprint*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ellie Campbell. 2024. Resources for Exploring the Benefits and Drawbacks of AI.

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.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1616):17754–17762.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Colleen V. Chien and Miriam 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*, forthcoming.

Colleen V. Chien, Miriam Kim, Raj Akhil, and Rohit Rathish. 2024. How Generative AI Can Help Address the Access to Justice Gap Through the Courts. *Loyola of Los Angeles Law Review*, forthcoming.

Jonathan H. Choi, Amy Monahan, and Daniel Schwarcz. 2024. Lawyering in the Age of Artificial Intelligence. *Minnesota Law Review*, forthcoming.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jack Collens, Rachel Reimer, Gerald Schifman, and Pamela Wilkinson. 2024. AI Survey: Where Artificial Intelligence Stands in the Legal Industry.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. *arXiv preprint*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analaysis*, forthcoming.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

Ronald Dworkin. 1986. *Law's Empire*. Harvard University Press, Cambridge, MA.

James H. Fowler, Timothy R. Johnson, James F. Spriggs, II, Sangick Jeon, and Paul J. Wahlbeck. 2007. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15(3):324–346.

Free Law Project. 2024. Courtlistener.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*.

K. Goddard, 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(1):121–127.

Google. 2024. Google search results for "practical law". Accessed: 2024-05-22.

Isabel Gottlieb. 2024. Paul weiss assessing value of ai, but not yet on bottom line. *Bloomberg Law*.

Stuart Greenhill. 2024. Lawyers Cross into the New Era of Generative AI.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv preprint*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 3929–3938. JMLR.org.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy 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)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Justin Henry. 2024. We Asked Every Am Law 100 Law Firm How They're Using Gen AI. Here's What We Learned. *The American Lawyer*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and Pitfalls of Artificial Intelligence for Legal Applications. *Journal of Cross-disciplinary Research in Computational Law*, forthcoming.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Darla Wynon Kite-Jackson. 2023. 2023 Artificial Intelligence (AI) TechReport. Technical report, American Bar Association.

Law360. 2024. Tracking Federal Judge Orders On Artificial Intelligence.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

LexisNexis. 2023. General terms and conditions.

LexisNexis. 2023. LexisNexis Launches Lexis+ AI, a Generative AI Solution with Linked Hallucination-Free Legal Citations.

LexisNexis. 2024. LexisNexis Launches Second-Generation Legal AI Assistant on Lexis+ AI.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models.

Michael Lissner. 2022. Important opinions on courtlistener are now summarized by the top experts — judges.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle 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*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Megan Ma, Aparna Sinha, Ankit Tandon, and Jennifer Richards. 2024. Generative AI Legal Landscape 2024. Technical report.

Olga Mack. 2023. How lawyers can embrace the challenge to supervise AI. *Legal Dive*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *arXiv preprint*.

Paul McGreel. 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.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *Preprint*, arXiv:2402.09880.

Eliza Mik. 2024. Caveat Lector: Large Language Models in Legal Practice.

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

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *arXiv preprint*.

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

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 959–972, New York, NY, USA. Association for Computing Machinery.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint*.

John G. Roberts. 2023. 2023 Year-End Report on the Federal Judiciary. Technical report.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Daniel Schwarcz and Jonathan H. Choi. 2023. AI Tools for Lawyers: A Practical Guide. *Minnesota Law Review Headnotes*, 108:1–39.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason 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*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Suffolk University. 2023. Law Faculty Guide to Artificial Intelligence: Practical Law AI (Westlaw).

Faiz Surani, Matthew Dahl, and Varun Magesh. 2024. Legal RAG hallucinations.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. 2023. ChatGPT as an Artificial Lawyer? In *Proceedings of the ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice*, Braga, Portugal. CEUR Workshop Proceedings.

Task Force on Artificial Intelligence. 2024. Report and Recommendations of the New York State Bar Association Task Force on Artificial Intelligence. Technical report, New York State Bar Association.

The Florida Bar. 2024. Florida Bar Ethics Opinion. Technical Report 24-1, The Florida Bar.

The State Bar of California. 2023. Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law. Technical report, The State Bar of California.

Thomson Reuters. 2023. Introducing AI-Assisted Research: Legal research meets generative AI.

Thomson Reuters. 2024. Introducing Ask Practical Law AI on Practical Law: Generative AI meets legal how-to.

University of Washington. 2024. Artificial Intelligence.

Jeremy Waldron. 2002. Is the Rule of Law an Essentially Contested Concept (in Florida)? *Law and Philosophy*, 21(2):137–164.

Ed Walters. 2019. The Model Rules of Autonomous Conduct: Ethical Responsibilities of Lawyers and Artificial Intelligence. *Georgia State University Law Review*, 35(4):1073–1092.

Benjamin Weiser. 2023. Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*.

Benjamin Weiser and Jonah E. Bromwich. 2023. Michael Cohen Used Artificial Intelligence in Feeding Lawyer Bogus Cases. *The New York Times*.

Serena Wellen. 2024a. How Lexis+ AI Delivers Hallucination-Free Linked Legal Citations.

Serena Wellen. 2024b. Tech Innovation with LLMs Producing More Secure and Reliable Gen AI Results.

Yale Law School. 2024. Lexis and Westlaw Generative AI Products.

Nicole Yamane. 2020. Artificial Intelligence in the Legal Field and the Indispensable Human Element Legal Ethics Demands. *Georgetown Journal of Legal Ethics*, 33(3):877–890.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *arXiv preprint*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Complete Query Descriptions

A.1 General Legal Research

A.1.1 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 minutes. 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.

A.1.2 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 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.

A.1.3 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 (Fowler et al., 2007; Black and Spriggs, 2013).

Evaluation Reference Whether the model correctly characterizes the treatment of the cited case, e.g., as "followed", "distinguished", "overruled," etc.

A.1.4 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.

A.1.5 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.

A.2 Jurisdiction or Time-specific

A.2.1 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.

A.2.2 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 10

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.

A.2.3 Changes in Law

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

# of Queries in Dataset 10

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.

A.2.4 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.

A.3 False Premise

A.3.1 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 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).

A.3.2 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.

A.3.3 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.

A.4 Factual Recall

A.4.1 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).

A.4.2 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).

A.4.3 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).

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.

B.1 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.

`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."

C Query Evaluation

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

C.1 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.

1. Evaluate groundedness, referring to the groundedness rubric.

C.2 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 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.

- “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)

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.
- Off-by-one year is incorrect

C.3 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; i.e. 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 which does not in reality support the proposition.

Fabricated

The answer cites a source which 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*/the 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, e.g. 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.

Checking Fabrications on Thomson Reuters/Lexis Responses

- Lexis and Thomson Reuters both sometimes cite to their own expert-written documents rather than publicly available legal text. If this seems to be the case, start by searching for the document title on the relevant service. If the document seems unavailable, reach out to the original annotator for hyperlinks to the precise reference.

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.