Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models

Matthew Dahl[*], Varun Magesh[†], Mirac Suzgun[‡], and Daniel E. Ho[§]

Abstract

Do large language models (LLMs) know the law? LLMs are increasingly being used to augment legal practice, education, and research, yet their revolutionary potential is threatened by the presence of "hallucinations"—textual output that is not consistent with legal facts. We present the first systematic evidence of these hallucinations in public-facing LLMs, documenting trends across jurisdictions, courts, time periods, and cases. Using OpenAI's ChatGPT 4 and other public models, we show that LLMs hallucinate at least 58% of the time, struggle to predict their own hallucinations, and often uncritically accept users' incorrect legal assumptions. We conclude by cautioning against the rapid and unsupervised integration of popular LLMs into legal tasks, and we develop a typology of legal hallucinations to guide future research in this area.

1 INTRODUCTION

How well do large language models (LLMs) know American case law? Modern LLMs such as OpenAI's ChatGPT—tools trained on vast amounts of textual data to predict the next token in a sequence—are driving a transformation in the legal world, from legal education (Choi and Schwarcz 2024), to legal research (Livermore, Herron, and Rockmore 2024), to legal practice itself (Rodgers, Armour, and Sako 2023). Indeed, recent versions of these artificial intelligence (AI) models seem to excel at law-related tasks, such as first-year law school exams (Choi et al. 2022), the uniform bar exam (Katz et al. 2023), statutory reasoning (Blair-Stanek, Holzenberger, and Van Durme 2023), and issue-rule-application-conclusion (IRAC) analysis (Guha et al. 2023). But despite the revolutionary potential of these models, a key challenge remains: the issue of "hallucinations." LLMs are liable to generate language that is inconsistent with current legal doctrine and case law, and, in the legal field, where adherence to authorities is paramount, unfaithful or imprecise interpretations of the law can lead to nonsensical—or worse, harmful and inaccurate—legal advice or decisions.

In this work, we present the first evidence documenting the nature, frequency, and correlates of these hallucinations. In doing so, we shed systematic, empirical light on a phenomenon that has so far only received anecdotal treatment in the literature. For example, much media attention has been directed toward a Manhattan lawyer who faced sanctions for using ChatGPT to generate fictional

[*] J.D. Candidate, Yale Law School, and Ph.D. Student, Yale University Department of Political Science, USA. E-mail: matthew.dahl@yale.edu.
[†] Research Fellow, Regulation, Evaluation, and Governance Lab, Stanford University, USA.
[‡] J.D. Candidate, Stanford Law School, and Ph.D. Candidate, Stanford University Department of Computer Science, USA.
[§] William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Senior Fellow at Stanford Institute for Economic Policy Research, and Faculty Director of the Regulation, Evaluation, and Governance Lab, Stanford University, USA.

case citations for a brief (Weiser 2023), or another instance where ChatGPT produced a supposed dissent authored by Justice Ruth Bader Ginsburg in the landmark gay rights case *Obergefell v. Hodges* (Romoser 2023). Even Chief Justice John Roberts, the chief justice of the U.S. Supreme Court, has weighed in on the problem, highlighting hallucinations in his 2023 report on the state of the federal judiciary and arguing that, as yet, "machines cannot fully replace key actors in court" (Roberts 2023, 6).

These impressionistic accounts, however, leave unanswered the deeper questions that legal scholars must confront as LLMs continue to grow in popularity. How much legal knowledge is actually embedded in an off-the-shelf LLM? Are LLMs equally familiar with different dimensions of the American common law system—where legal doctrine varies across courts, jurisdictions, and over time—or do they tend to hallucinate more in certain areas than others? When LLMs do hallucinate, do they disproportionately produce false information favoring certain judges or cases? And besides hallucination itself, are there other features of LLMs that legal scholars need to consider—other latent biases or behavioral tendencies that threaten to spill over into downstream applications of these models? Our study seeks to answer these questions, providing insights that are essential for evaluating LLMs' effectiveness in general legal settings.

This research contributes to several literatures. First, there has recently been an explosion of interest in the intersection of law and technology, with a particular focus on the emergence of AI. Much of this work focuses on how lawmakers and administrative agencies ought to govern the deployment of these tools (Engstrom, Ho 2020; Engstrom et al. 2020; Solow-Niederman 2020), given that they are already being used by public (Engel, Grgić-Hlača 2021) and private (Barocas and Selbst 2016) actors alike, producing novel privacy concerns (Ben-Shahar 2023; King et al. 2023) and giving rise to new forms of liability (Henderson, Hashimoto, and Lemley 2023; Lemley and Casey 2019; Volokh 2023). As one highly influential but still maturing species of AI, LLMs stand in need of a concrete empirical evaluation of their legal abilities and their legal risks, of which hallucination is certainly one. We supply that information here.

We also contribute to a growing literature regarding the implications of AI for access to justice. Many members of the legal community rightly regard LLMs as a promising solution to the longstanding barriers to adequate legal representation that millions of *pro se* and under-resourced litigants encounter (Chien et al. 2024; Perlman 2023; Tan, Westermann, and Benyekhlef 2023). Because they are relatively cheap, easy, and quick to use, LLMs might finally be able to deliver on the federal rules' guarantee of a "just, speedy, and inexpensive" resolution of disputes (Fed. R. Civ. P. 1; Roberts 2023). This potential can only be realized, however, if LLMs actually know the law. Additionally, if the legal knowledge embedded in LLMs is not evenly distributed, the widespread adoption of LLMs might unintentionally worsen rather than eliminate current disparities in the availability of legal services (Draper and Gillibrand 2023; Simshaw 2022). We therefore approach our study of LLMs with an eye toward assessing their ability to truly close the justice gap, examining both their raw hallucination rates as well as any other emergent behaviors that threaten this potential.

Finally, we also contribute to the pressing algorithmic harm literature, which is motivated by the concern that inscrutable algorithms often produce predictions, recommendations, or outputs that are not fairly distributed among individuals or groups (Bar-Gill, Sunstein, and Talgam-Cohen 2023; Gillis and Spiess 2019; Kleinberg et al. 2018; Mayson 2019). In our legal setting, the specific danger is that if LLMs do not properly internalize knowledge about some dimension of the law—if LLMs know California law better than Wyoming law, for example, or decisions by Justice Ketanji Brown Jackson worse than decisions by Justice Antonin Scalia, for another—they will regurgitate a falsely homogeneous sense of the legal landscape to their users, collapsing important legal nuances and perpetuating representational harms. Worse, because LLMs are so-called "foundation" models, their distributional biases, if they exist, may permeate and afflict *every* downstream version of these models (Bommasani et al. 2022), producing a kind of algorithmic "monoculture" by entrenching one particular notion of the law across a wide range of applications (Creel and Hellman 2022; Kleinberg and Raghavan 2021). Accordingly, it is important for legal scholars to obtain a sense of what the correlates of LLMs' hallucinations are, in order to address this new and profound opportunity for cascading algorithmic harms.

Our article proceeds as follows. In Section 2, we provide a brief background on LLMs for the non-technical reader and theorize a typology of legal hallucinations. In Section 3, we develop a set of legal knowledge queries that we use to elicit an LLM's understanding of the law, from simple queries

like whether or not a case exists to more complex queries like asking for a statement of a case's holding or its precedential relationship to another case. In Section 4, we describe our methodological approach, which entails asking these queries for a random sample of cases across each level of the federal judiciary—the US District Courts (USDC), the US Courts of Appeals (USCOA), and the US Supreme Court (SCOTUS)—and evaluating them using four popular LLMs: OpenAI's ChatGPT 4, OpenAI's ChatGPT 3.5, Google's PaLM 2, and Meta's Llama 2.

In Section 5, we present our results. Our findings reveal the widespread occurrence of legal hallucinations: when asked a direct, verifiable question about a randomly selected federal court case, LLMs hallucinate between 58% (ChatGPT 4) and 88% (Llama 2) of the time. However, we also find that LLMs perform better on cases that are newer, more salient, and from more prominent legal jurisdictions, suggesting that the risks of legal monoculture are real. We then investigate two additional potential failure points for LLMs, beyond their raw hallucination rates: (i) their susceptibility to contra-factual bias, i.e., their ability to respond to queries anchored in erroneous legal premises (Sharma et al. 2023; Wei et al. 2023), and (ii) their certainty in their responses, i.e., their self-awareness of their propensity to hallucinate (Azaria and Mitchell 2023; Kadavath et al. 2022; Tian, Mitchell, Zhou, et al. 2023; Xiong et al. 2023; Yin et al. 2023). Our results indicate that not only do LLMs often provide seemingly legitimate but incorrect answers to contrafactual legal questions, they also struggle to accurately gauge their own level of certainty without post-hoc recalibration. Accordingly, in Section 6 we conclude that while LLMs appear to offer a way to make legal information and services more accessible and affordable to all, their present shortcomings—particularly in terms of generating accurate and reliable statements of the law—significantly hinder this objective.

2 BACKGROUND AND THEORY

2.1 What Is a Language Model?

We first provide a brief overview of language models (LMs) for readers who may not necessarily have a deep technical background. LMs can be viewed as functions that map text to text: When a user provides a text input (known as a "prompt"), the model produces a text output (referred to as a "response"). If the prompt takes the form of a question, the response can be understood as an answer to that question. An LM generates its response by selecting the most probable sequence of tokens that follow the prompt's tokens; therefore, it essentially functions as a probability distribution over these tokens.

In this work, we focus on *large* language models (LLMs). The largeness of a language model is a dual reference to its parameter count and the scope of its training corpus: LLMs are models that contain billions of parameters and are trained on vast corpora bordering on the size of the Internet. Because of their incredible size, LLMs can be considered general purpose technologies, with the apparent ability to understand and generate human-like text across a wide range of topics, including medicine, finance, education, retail, and law (Eloundou et al. 2023). In contrast to previous forms of machine learning, however, they seem to excel at these tasks despite not being explicitly trained to perform them (Brown et al. 2020); the "jagged frontier" of their emergent abilities is still being mapped (Dell'Acqua et al. 2023).

We also set forth a more formal definition of an LLM, in order to provide the foundation for the typology of legal hallucinations that we develop in the next subsection. We let an LLM be a function $f_\tau : \text{prompt} \mapsto \text{response}$, where f_τ operates by sampling responses from a conditional probability distribution that is learned by optimizing over a training corpus hopefully reflective of facts about the world.[1] We use the symbol τ to designate a user-configurable "temperature" parameter that controls the shape of the probability distribution at inference time. When $\tau = 0$, the distribution becomes degenerate and the model's response is theoretically deterministic—the model must always return the most likely response.[2] Following convention, we refer to this deterministic response as the model's "greedy" response. As τ increases, the distribution becomes more uniform and the model's response becomes more

---

[1] In reality, language generation in LMs actually happens at the level of tokens, not responses themselves; full model responses are constructed autoregressively by sampling n tokens, one at a time, from a distribution $Pr[x_n \mid x_1, \cdots, x_{n-1}]$. We abstract from these details in this article without loss of generality.

[2] Non-determinism may persist in practice due to a model's implementation details, e.g., the "mixture of experts" architecture (Chann 2023).

stochastic—the model is free to choose from a variety of candidate responses, all of which become more equally likely to be chosen the higher the temperature is. Thus, increasing the temperature of an LLM is one way to potentially increase its hallucination frequency (Lee 2023). In this article, however, we generally perform our experiments at $\tau = 0$, showing that LLMs hallucinate even under the most conservative sampling conditions.

2.2 The Nature of Legal Hallucinations

LLMs are showing promise on a number of legal research and analysis tasks (Ash et al. 2024; Blair-Stanek, Holzenberger, and Van Durme 2023; Choi et al. 2022; Fei et al. 2023; Guha et al. 2023; Katz et al. 2023; Trozze, Davies, and Kleinberg 2023), but the problem of legal hallucination has so far only been studied in closed-domain applications, such as when a model is used to summarize the content of a given judicial opinion (Deroy, Ghosh, and Ghosh 2023; Feijo and Moreira 2023) or to synthesize provided legal text (Savelka et al. 2023). In this article, by contrast, we examine hallucination in an open-domain setting, i.e., when a model is tasked with providing an accurate answer to an open-ended legal query. This setting approximates the situation of a lawyer or a *pro se* litigant seeking advice from a legal chat interface.

In the context of such question-answering (QA) scenarios, the study of hallucinations in LMs is still in its infancy, even outside the legal field. There is no universally accepted definition or classification of LM hallucinations (Ji et al. 2023; Zhang et al. 2023; van Deemter 2024). However, as Kalai and Vempala (2023) show, LMs that assign a positive probability to every response token *must* hallucinate at least some of the time. Xu, Jain, and Kankanhalli (2024) agree, arguing that "hallucination is inevitable for any computable LLM, regardless of model architecture, learning algorithms, prompting techniques, or training data." Therefore, if hallucinations are here to stay, we believe that it is essential for legal scholars to begin to recognize that there are several different ways in which an LLM can generate false information, as not all modes of hallucination are equally concerning for legal professionals. For example, since hallucinations seem likely to give rise to new forms of tort liability (Henderson, Hashimoto, and Lemley 2023), it will be important to differentiate between different types of hallucinations to properly assess the predicate elements of such torts. We supply those theoretical resources here, summarizing our typology of legal hallucinations in Table 1.

First, a model might hallucinate by producing a response that is either unfaithful to or in conflict with the input prompt, a phenomenon canonically referred to as *closed-domain* or *intrinsic* hallucination. This is a major concern in tasks requiring a high degree of accuracy between the response and a long-form input, such as machine translation (Xu et al. 2023) or summarization (Cao et al. 2018). In legal contexts, such inaccuracies would be particularly problematic in activities like summarizing judicial opinions, synthesizing client intake information, drafting legal documents, or extracting key points from an opposing counsel's brief.

Second, an LLM might also hallucinate by producing a response that either contradicts or does not directly derive from its training corpus. Following Agrawal et al. (2023), we conceptualize this kind of hallucination as one form of *open-domain* or *extrinsic* hallucination. In general, the output of a language model should be logically derivable from the content of its training corpus, regardless of whether the content of the corpus is factually or objectively true.[3] In legal settings, this kind of hallucination poses a special challenge to those aiming to fine-tune the kind of general-purpose foundation models that we study in this article with proprietary, in-house work product.[4] For example, firms might have a

Table 1 Typology of legal hallucinations

Domain	Type of hallucination	Legal example
Closed	Response inconsistency with the prompt	Mischaracterization of an opinion
Open	Response inconsistency with the training corpus	Creative argumentation
	Response inconsistency with the facts of the world	Misstatement of the law

---

[3] For example, if a training corpus consisted of J. K. Rowling's *Harry Potter* series, we would expect an LLM to produce the sentence "Tom Marvolo Riddle" in response to a query about Voldemort's real name. However, if the training corpus consisted solely of Jane Austen's *Pride and Prejudice* (for instance), we would consider this LLM output to be a hallucination—because there would be no basis in the training data for making such a claim about Voldemort.

[4] For example, this kind of firm-specific fine-tuning is the business model of a prominent legal tech startup, Harvey.ai (Ambrogi 2023).

catalogue of internal research memos, style guides, and so forth, that they want to ensure is reflected in their bespoke LLM's output. At the same time, however, insofar as *creativity* is valued, certain legal tasks—such as persuasive argumentation—might actually benefit from some lack of strict fidelity to the training corpus; after all, a model that simply parrots exactly the text that it has been trained on could itself be undesirable. As mentioned, creativity can be induced by raising the temperature of the LLM, but responses that are more unpredictable are also those that are more likely to be hallucinations (Lee 2023). Thus, defining the contours of what counts as an *unwanted* hallucination in this specific sense requires value judgements about the balance between fidelity and spontaneity.

Finally, the third way that an LLM can hallucinate is by producing a response that lacks fidelity to the facts of the world, irrespective of how the LLM is trained or prompted (Maynez et al. 2020). We consider this to be another type of open-domain hallucination, with the key concern being "factuality" in relation to the facts of the world (cf. Wittgenstein, 1998 [1921]). In our context, this is perhaps the most alarming type of hallucination, as it can undermine the accuracy required in any legal context where a correct statement of the law is necessary.

2.3 Hallucination Trade-offs

In this article, we investigate only the last kind of hallucination. As mentioned, the first two modes of hallucination are not always problematic in the legal setting: these kinds of hallucinations could actually be somewhat desirable to lawyers if they resulted in generated language that, for example, removed unnecessary information from a given argument (at the expense of being faithful to it) or invented a novel analogy never yet proposed (at the expense of being grounded in the lexicon) (Cao, Dong, and Cheung 2022). However, what a lawyer cannot tolerate is the third kind of hallucination, or factual infidelity between an LLM's response and the controlling legal landscape. In a common law system, where *stare decisis* requires attachment to the "chain" of historical case law (Dworkin 1986), any misstatement of the binding content of that law would make an LLM quickly lose any professional or analytical utility.

Focusing on non-factual hallucinations alone, however, comes with certain trade-offs. One of the advantages of our typology is that it makes clear that it may not always be possible to minimize all modes of hallucination simultaneously; reducing hallucinations of one kind may increase hallucinations of another. For example, if a given prompt contains information that does not conform to facts

Table 2 Hallucination QA task list. Tasks are sorted in order of increasing complexity. Query wording is paraphrased; see the Online Appendix for exact prompt used. Method column describes the inferential strategy that we use to estimate a hallucination rate for each task: reference-based tasks use known metadata to assess hallucinations, and reference-free tasks use emergent contradictions to assess hallucinations (see Section 4).

Complexity	Task	Query	Method
Low	Existence	Is {case} a real case?	Reference-based
	Court	What court decided {case}?	Reference-based
	Citation	What is the citation for {case}?	Reference-based
	Author	Who wrote the majority opinion in {case}?	Reference-based
Moderate	Disposition	Did {case} affirm or reverse?	Reference-based
	Quotation	What is a quotation from {case}?	Reference-based
	Authority	What is an authority cited in {case}?	Reference-based
	Overruling year	What year was {case} overruled?	Reference-based
High	Doctrinal agreement	Does {case1} agree with {case2}?	Reference-based
	Factual background	What is the factual background of {case}?	Reference-free
	Procedural posture	What is the procedural posture of {case}?	Reference-free
	Subsequent history	What is the subsequent history of {case}?	Reference-free
	Core legal question	What is the core legal question in {case}?	Reference-free
	Central holding	What is the central holding in {case}?	Reference-free

about the world, then ensuring response fidelity with respect to the former would by definition produce infidelity—i.e., hallucination—with respect to the latter. More generally, although fidelity to the prompt is necessary for avoiding *closed-domain* hallucination, there is an important sense in which prioritizing such behavior might actually induce the kind of *open-domain* hallucination that we center in this article.

These trade-offs present unavoidable challenges for prospective users of legal LLMs. When responding to a query, should an LLM be skeptical of its prompt or sycophantic to it? If it has been trained on case law from one jurisdiction, should it enforce adherence to that training corpus even when responding about the law in another jurisdiction? If facts about the world conflict with each other—as legal rules often do—should the LLM preserve that nuance or refrain from introducing information outside the scope of a query? Questions like these are ultimately questions about which kinds of legal hallucinations are more and less preferable, and they are questions whose answers require *both* empirical evidence *and* normative arguments. For example, minimizing fact and training corpus hallucinations (at the expense of prompt hallucinations) might be best for avoiding harm to *pro se* litigants, but the calculus might be reversed for sophisticated lawyers who might be less vulnerable to such behavior. We supply some of the empirics that speak to these dilemmas (see Sections 5.1.6 and 5.2), but stress that the normative considerations are crucial and should be a topic of continued legal hallucination research.

3 PROFILING HALLUCINATIONS USING LEGAL KNOWLEDGE QUERIES

To empirically assess the incidence and correlates of non-factual hallucinations, we adopt a QA framework where the goal is to test an LLM's ability to produce accurate information in response to different kinds of legal queries. We develop fourteen tasks representative of such queries, which we group into three categories in order of increasing complexity and list in Table 2.

3.1 Low Complexity Tasks

In the low complexity category, we ask for information that we consider relatively easy for an LLM to reproduce. The information in this category does not derive from the actual content of a case itself, so it does not require higher-order legal reasoning skills to internalize. Instead, this information is readily available in a case's caption or its syllabus—standard textual locations whose patterns even non-specialized LLMs should be able to recover. We therefore expect LLMs to perform best on these tasks:

3.1.1 Existence

Given the name and citation of a case, state whether the case actually exists or not. This basic evaluation provides preliminary insights into an LLM's knowledge of actual legal cases: if it cannot distinguish real cases from non-existent ones, it probably cannot offer detailed case insights. We use only real cases in our prompts, so affirming their existence is the correct answer.[5]

3.1.2 Court

Given the name and citation of a case, supply the name of the court that ruled on it. This task assesses an LLM's knowledge about legal jurisdictions, an important building block of a case's precedential value. We perform this task across the three different levels of the federal judiciary. Importantly, we note that each level of the judiciary has a different reporter, or the series of volumes that opinions are published in. This is relevant because the reporter is included in the citation that we provide to the LLM, essentially revealing the level of the hierarchy that an opinion is from. All and only SCOTUS cases are published in the *US Reports*. Opinions from the USCOA are published in the *Federal Reporter*, and USDC cases are published in the *Federal* Supplement. Because of this, we expect this task to be more difficult as we descend the hierarchy of courts. There is only one court associated with the US reporter, but 13 associated with the *Federal Reporter*, and 94 associated with the *Federal* Supplement. For USCOA cases, we require the name of the specific circuit court, and for USDC cases, we require the name of the specific district court.

3.1.3 Citation

Given a case name, supply the Bluebook citation of the case. This query tests an LLM's ability to associate a given dispute with its official record in a reporter volume at a particular page, which is the key way in

---

[5] In the Online Appendix, we experiment with using fake cases as well.

which different opinions reference and link to each other. For USCOA cases, we further specify that we want the citation for the circuit court opinion, and for USDC cases, we further specify that we want the citation for the district court opinion. We test for citation equality using *eyecite* (Cushman, Dahl, and Lissner 2021).

3.1.4 Author

Given the name and citation of a case, supply the name of the opinion author. This query tests an LLM's ability to associate a given case with a particular judge, which is important for contextualizing a case in the broader jurisprudential landscape. For SCOTUS and USCOA cases, we further specify that we want the name of the *majority* opinion author. We accept a fuzzy match of the opinion author's name as accurate.

3.2 Moderate Complexity Tasks

Next, in the moderate complexity category, we start to require an LLM to evince knowledge of actual legal opinions themselves. To answer the queries in this category, an LLM must know something about a case's substantive content; these queries seek information that must be collated from idiosyncratic portions of its text. Of course, a database-augmented LLM might still be able to retrieve some of this information without ever actually internalizing the content of a case, but we expect this text-based knowledge to be less available than the information described in the low complexity category. Specifically, we ask for the following information:

3.2.1 Disposition

Given a case name and its citation, state whether the court affirmed or reversed the lower court. This query tests an LLM's knowledge of how the court resolved the instant appeal confronting the parties in the case, which is the first step for determining the holding that is created by the case. Though this is essentially a binary classification task where we accept correct "affirm" or "reverse" labels as accurate, we consider this task to still be probative of hallucinations because producing the wrong label is still a misstatement of the law. We filter out all ambiguous dispositions (e.g., reversals in part) and we do not ask this query of USDC cases because district courts are courts of original jurisdiction.[6]

3.2.2 Quotation

Given a case name and its citation, supply any quotation from the opinion. This query tests an LLM's ability to produce some portion of an opinion's text verbatim, which is an important feature for lawyers seeking to use a case to stand for a specific proposition. Normally, such memorization is considered an undesirable property of LLMs (Carlini et al. 2022), but in this legal application it is actually desirable behavior. We accept any fuzzy string of characters appearing in the majority opinion as accurate.

3.2.3 Authority

Given a case name and its citation, supply a case that is cited in the opinion. This query probes an LLM's understanding of the chain of precedential authority that supports a given opinion. We do not distinguish between positive and negative citations for this task; we accept any precedent cited in any way in the text of the majority opinion as accurate. We extract and match citations on their volumes, reporters, and pages using *eyecite* (Cushman, Dahl, and Lissner 2021).

3.2.4 Overruling Year

Given a case name and its citation, supply the year that it was overruled. This query tests an LLM's ability to recognize when a given case has been subsequently altered, which is crucial information for lawyer seeking to determine whether a given precedent is still good law or not. This task is the most complicated in this category because it requires the LLM to draw connections between multiple areas of the case space. We accept only the exact year of overruling as accurate, and we limit this task to only those SCOTUS cases that have been explicitly overruled ($n = 279$).[7]

---

[6] While it is possible for some administrative agency decisions to be appealed to a district court, this occurs infrequently enough that we choose not to ask for case disposition at the district court level.

[7] In Section 5.2, we experiment with cases that have never been overruled as well.

3.3 High Complexity Tasks

Finally, in the high complexity category, we seek answers to tasks that *both* presuppose legal reasoning skills (unlike the low complexity tasks) *and* are not readily available in existing legal databases like WestLaw or Lexis (unlike the moderate complexity tasks). These tasks all require an LLM to synthesize core legal information out of unstructured legal prose—information that is frequently the topic of deeper legal research. In Section 4.3, we explain how we test LLMs' knowledge of some of these more complex facts without necessarily having access to the ground- truth answers ourselves:

3.3.1 Doctrinal Agreement

Given two case names and their citations, state whether they agree or disagree with each other. This query requires an LLM to show knowledge of the precedential relationship between two different cases, information that is essential for higher-order legal reasoning. We use Shepard's treatment codes as a basis for constructing this task, filtering out all ambiguous citation treatments (e.g., neutral treatments) and coarsening the unambiguous codes into "agree" and "disagree" labels that we accept as accurate. For this task, we use a relatively balanced dataset of 2,839 citing-cited case pairs coded as "agree," and 2,161 citing-cited case pairs coded as "disagree." This task is limited to SCOTUS cases, as our underlying dataset only contains thorough Shepard's data for citations to the Supreme Court.

3.3.2 Factual Background

Given a case name and its citation, supply its factual background. This query tests an LLM's understanding of the concrete fact pattern underlying a case, which is helpful in assessing the relevance of the case to current research and in drawing parallels with other cases.

3.3.3 Procedural Posture

Given a case name and its citation, supply its procedural posture. This query tests an LLM's understanding of how and why a case has arrived at a particular court, which aids in understanding the precise question presented and standard of review applicable.

3.3.4 Subsequent History

Given a case name and its citation, supply its subsequent procedural history, if any. This query tests an LLM's knowledge of any other related proceedings that concern the given case after a particular decision, which is information that can change or clarify the legal significance of the case.

3.3.5 Core Legal Question

Given a case name and its citation, supply the core legal question at issue. This query tests an LLM's ability to pinpoint the main issue or issues that a court is addressing in a case, which is the most important factor in assessing whether a case is apposite or not.

3.3.6 Central Holding

Given a case name and its citation, supply its central holding. This query tests an LLM's knowledge of the legal principle that a given case stands for, i.e., the precedent that future cases will rely upon or distinguish from. Articulating the holding of a case is crucial for legal analysis and argumentation and is the most complex task that we evaluate.

4 EXPERIMENTAL DESIGN

4.1 Data Construction

We aim to profile hallucination rates across several legally salient dimensions, including hierarchy, jurisdiction, time, and case prominence. Thus, we construct our test data with an eye toward making statistical inferences on these covariates.

We begin with the universe of case law from each level of the federal judicial hierarchy—namely, SCOTUS, USCOA, and USDC—that has been published in the volumes of the *U.S. Reports*, the *Federal Reporter*, and the *Federal Supplement*. To ensure balance over time and place, we then perform stratified random sampling using year strata for the SCOTUS cases, circuit-year strata for the USCOA

cases, and state-year strata for the USDC cases. We draw 5,000 cases from each level of the judiciary. Finally, to generate ground-truth answers for our reference-based queries (Section 4.2), we merge these units with metadata obtained from the Caselaw Access Project (2023), the Supreme Court Database (Spaeth et al. 2022), the Appeals Courts Database Project (Songer 2008; Kuersten and Haire 2011), the Library of Congress (Congress.gov 2023), and Shepard's Citations (Fowler et al. 2007; Black and Spriggs 2013).[8]

4.2 Reference-Based Querying

The most straightforward way to study hallucinations in the open-domain setting is to use a test oracle—or an external *reference*—to detect and adjudge non-factual responses (Lin, Hilton, and Evans 2022; Lee et al. 2023; J. Li et al. 2023). Such oracles are usually difficult and costly to construct (Krishna, Roy, and Iyyer 2021), but we use the tabular metadata described in Section 4.1 to develop ours. Our design exploits the fact that while LLMs are known to have been trained on the raw text of American case law, which is in the public domain (Henderson et al. 2022), they have likely *not* been trained on these cases' attendant metadata, which exist separately from the cases' textual content and which we have aggregated from disparate sources.

These metadata enable us to construct reference-based queries for the first nine of our tasks (Table 2). These queries take the form of N question-and-answer triples (prompt, response, response′), where prompt is a case-specific question, response is the LLM's greedy answer retrieved from calling , and response′ is the known ground-truth answer.[9] Our estimand of interest for each task is the population-level hallucination rate π, which we estimate by averaging over the N sampled queries:

$$\pi = \hat{\pi} = \frac{1}{N} \sum 1[response \neq response'] \tag{1}$$

Occasionally, an LLM will produce a response that is neither a hallucination nor a correct answer, but rather an explicit *abstention* from answering the question. For example, the LLM might admit that it does not know the answer or demur that it is unable to provide the answer for some reason, perhaps due to safety concerns. In these instances, we nevertheless count the response as a non-hallucination, on the theory that an LLM cannot hallucinate when it is affirmatively abstaining from responding (Feng et al. 2024). We document the frequency of these abstentions in the Online Appendix, but they are generally rare and do not substantively affect our findings.

4.3 Reference-Free Querying

Reference-based querying lets us directly recover our population parameter of interest, but two problems limit the effectiveness of the approach. First, we are restricted to asking questions for which digestible metadata exist and a clear answer has been recorded, which rules out many more complex inquires. Second, precisely because these queries can be answered with tabular data, legal database-augmented LLMs (Cui et al. 2023; Savelka et al. 2023) are likely to soon solve or at least mask hallucinated responses to these queries (Peng et al. 2023; Shuster et al. 2021).

To test the tasks that cannot be easily verified against an external legal database, we employ reference-free querying instead, which detects hallucinations by exploiting the stochastic behavior of LLMs at higher temperatures (Agrawal et al. 2023; Manakul, Liusie, and Gales 2023; Min et al. 2023). This approach is rooted in the theory that hallucinations are more likely to originate in flat probability distributions with higher next-token uncertainties, whereas factual answers should always have a high probability of being the generated response given a prompt. Thus, by repeatedly querying an LLM at a non-greedy temperature, we can estimate the model's hallucination rate by examining its self-consistency—factual responses should not change, but hallucinated ones will.

Most reference-free approaches implicitly assume that the LLM is calibrated, i.e., that there is indeed some correlation between its self-consistency and its propensity to hallucinate. For reasons that we discuss in Section 5.3, we are unwilling to make this assumption in our legal setting. We therefore adopt a slightly different implementation that is still reference-free, but only requires *contradiction*, not *consistency* (Mündler et al. 2023). Specifically, for our final five tasks (Table 2), we construct reference-free

---

[8] More information about how we use these metadata to construct each query is available in the Online Appendix.
[9] Recall from Section 2.1 that the f_0 notation represents performing inference with the LLM at temperature zero—i.e., under its deterministic behavior.

queries in the form of N question-and-answer triples (prompt, response$^{(1)}$, response$^{(2)}$), where prompt is the question, response$^{(1)}$ is one LLM answer retrieved by calling , once, and response$^{(2)}$ is another LLM answer retrieved by calling again.[10] Detecting a hallucination then amounts to detecting a logical contradiction between the two stochastic answers: any such contradiction guarantees non-factuality, because two contradictory answers cannot both be correct.

To identify these contradictions at scale, we feed both answers into GPT 4 and ask it for its assessment. This technique does not assume anything about $f_1(\cdot)$'s calibration—it just requires that GPT 4 possess logical reasoning skills sufficient to compare $f_1(\cdot)$'s two responses and accurately label them as contradictory as not. To justify this reliance on GPT 4, we manually label a portion of the reference-free responses ourselves and conduct an intercoder reliability analysis to ensure that GPT 4 is indeed able to perform this task. Full information about our procedure and a validity check is provided in the Online Appendix. (We find that GPT 4's reliability is comparable to human labeling of contradictions.)

An important caveat of this approach is that it only allows us to establish a *lower bound* on the hallucination rate for our reference-free queries:

$$\pi \geq \hat{\pi} = \frac{1}{N} \sum 1[response^{(1)} \neq response^{(2)}]$$

(2)

Although self-contradiction guarantees hallucination, the inverse does not hold: two answers may be logically consonant but still lack fidelity to the law. Because we are unwilling to assume calibration, we accept this inferential limitation, but, as we show below, even the lower bounds on hallucination rates are quite high and informative.

4.4 Models

We perform our experiments using four popular, state-of-the-art, off-the-shelf LLMs:

1. OpenAI's ChatGPT 4 (gpt-4-1106-preview, OpenAI 2023a),
2. OpenAI's ChatGPT 3.5 (gpt-3.5-turbo-0613, 2023b),
3. Google's PaLM 2 (text-bison-001, Anil et al. 2023), and
4. Meta's Llama 2 (Llama-2-13b-chat-hf, Touvron et al. 2023).

We run each query under both "zero-shot" and "three-shot" prompting setups. In the zero-shot setup, we simply ask the LLM about the given case directly, whereas in the three-shot setup, we prepend several example questions and responses to give the LLM an opportunity to perform in- context learning (Brown et al. 2020). We provide the full text of the prompts we use for each query, along with the few-shot examples, in the Online Appendix. In total, we execute more than 800,000 queries—200,000+ per LLM—and we share our raw API calls and model responses in the replication materials accompanying this article.

5 Results

We begin by presenting our main results profiling LLMs' hallucination rates, which cut to the core of popular concerns over LLMs' suitability for legal applications (Section 5.1). Then, after showing that hallucinations are generally widespread, and highlighting the correlates of LLMs' hallucination rates, we turn to two additional challenges that threaten LLMs' utility for legal adoption: (i) their susceptibility to contra-factual bias, i.e., their ability to handle queries based on mistaken legal premises (Section 5.2), and (ii) their certainty in their responses, i.e., their self-awareness of their propensity to hallucinate (Section 5.3).

5.1 Hallucination Rates and Heterogeneity

Tables 3–5 report our estimated hallucination rates and their standard errors for each category of our tasks. We find that hallucinations vary with the substantive complexity of the task (Section 5.1.1), the hierarchical level of the court (Section 5.1.2), the jurisdictional location of the court (Section 5.1.3), the

---

[10] Recall from Section 2.1 that the $f_1(\cdot)$ notation represents performing inference with the LLM at temperature one—i.e., with some degree of stochasticity.

prominence of the case (Section 5.1.4), the year the case was decided (Section 5.1.5), and the LLM queried (Section 5.1.6). We do not find substantial differences be- tween zero-shot and few-shot prompting, so we focus our discussion on the few-shot results alone.

5.1.1 Hallucinations Vary by Task Complexity

As we hypothesized in Section 3, we first observe that hallucinations increase with the complexity of the legal research task at issue, which we visualize in Figure 1. Starting with the low complexity category (Table 3), the LLMs perform best on the simple **Existence** task, though this is in part driven by their tendency to always answer "yes" when asked about the existence of *any* case. (In the Online Appendix we demonstrate this problem by asking about the existence of fake cases instead.) The models begin to
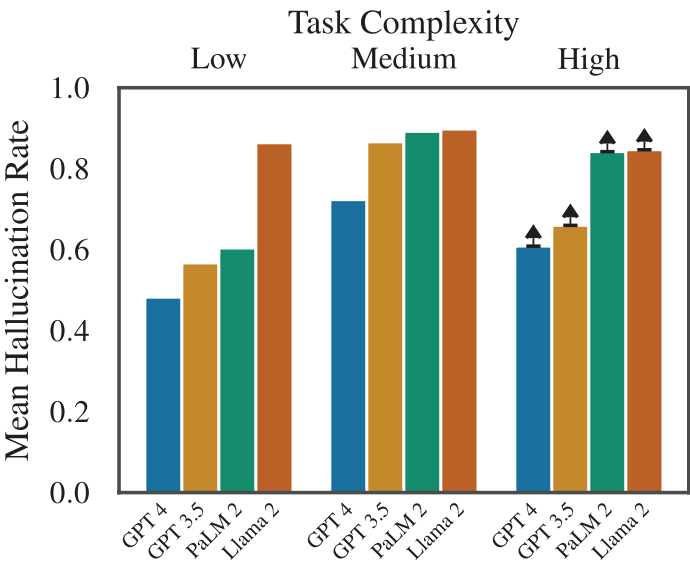
Figure 1. Relationship between task complexity and mean hallucination rate. Higher values indicate a greater likelihood of factually incorrect LLM responses. High complexity tasks include several reference-free tasks, so those reported hallucination rates are lower bounds on the true rates. Contra-factual tasks and the doctrinal agreement high complexity task are excluded from this comparison.
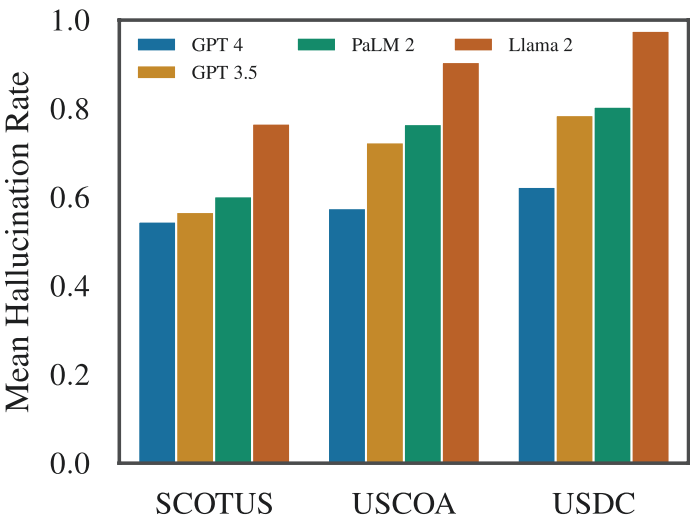
Figure 2. Relationship between judicial hierarchy and mean hallucination rate, all reference-based tasks pooled. Hallucination rates are higher for lower levels of the federal judiciary.

Table 3 Hallucination rates across levels of the federal judiciary (low complexity tasks)

Task	Prompt	SCOTUS (1794–2015; n = 5,000)				USCOA (1895–2019; n = 5,000)				USDC (1932–2019; n = 5,000)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
Existence	Zero-shot	0.204 (0.006)	0.04 (0.001)	0.054 (0.003)	0.303 (0.006)	0.237 (0.006)	0.003 (0.001)	0.025 (0.002)	0.157 (0.005)	0.174 (0.005)	0.001 (0.001)	0.016 (0.002)	0.24 (0.006)
	Few-shot	0.181 (0.005)	0.029 (0.002)	0.029 (0.002)	1.000 (0.000)	0.129 (0.005)	0.018 (0.002)	0.005 (0.001)	1.000 (0.000)	0.048 (0.003)	0.004 (0.001)	0.006 (0.001)	1.000 (0.000)
Court	Zero-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.003 (0.001)	0.490 (0.007)	0.645 (0.007)	0.703 (0.006)	0.700 (0.006)	0.829 (0.005)	0.815 (0.005)	0.839 (0.005)	0.815 (0.005)
	Few-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.005 (0.001)	0.491 (0.007)	0.641 (0.007)	0.703 (0.006)	0.679 (0.007)	0.831 (0.005)	0.87 (0.005)	0.842 (0.005)	0.87 (0.005)
Citation	Zero-shot	0.621 (0.007)	0.684 (0.007)	0.906 (0.004)	0.941 (0.003)	0.727 (0.006)	0.754 (0.006)	1.000 (0.000)	0.999 (0.000)	0.61 (0.007)	0.702 (0.006)	1.000 (0.000)	1.000 (0.000)
	Few-shot	0.689 (0.007)	0.641 (0.007)	0.833 (0.005)	0.953 (0.003)	0.468 (0.007)	0.92 (0.004)	0.999 (0.001)	0.000 (0.000)	0.270 (0.006)	0.942 (0.003)	0.999 (0.000)	1.000 (0.000)
Author	Zero-shot	0.799 (0.006)	0.796 (0.006)	0.816 (0.005)	0.884 (0.005)	0.954 (0.003)	0.965 (0.003)	0.988 (0.002)	0.991 (0.001)	0.922 (0.004)	0.911 (0.004)	0.988 (0.002)	0.987 (0.002)
	Few-shot	0.799 (0.006)	0.83 (0.005)	0.859 (0.005)	0.881 (0.005)	0.962 (0.003)	0.967 (0.003)	0.988 (0.002)	0.993 (0.001)	0.921 (0.004)	0.941 (0.003)	0.984 (0.002)	0.987 (0.002)

Note: Table reports estimated hallucination rates. Standard errors are shown in parentheses.

struggle more when prompted for information about a case's **Court**, **Citation**, or **Author**. Hallucinations then surge among the moderate complexity tasks (Table 4), all of which require the LLMs to evince knowledge of the actual content of a legal opinion. We note that these results are not just a product of different evaluation metrics: although the **Quotation** task, for example, requires near-word reproduction of particular sentences and phrases to be judged correctly, the **Disposition** task simply asks for binary responses from the model. Yet, the LLMs hallucinate widely in both setups.

The results for the high complexity tasks (Table 5) confirm this general pattern of poor performance. Starting with **Doctrinal agreement**, recall that this query asks the LLM to make an analogical judgment about the precedential relationship between two given cases, for which we have ground-truth labels from Shepard's treatment codes. Because this is another binary classification task, the LLMs' hallucination rates on this task—near 0.5—represent little improvement over random guessing, and are actually sometimes worse. This suggests that LLMs know little about substantive legal doctrine, calling into question their ability to accurately assist lawyers in more realistic, applied settings.

The remaining tasks in the high complexity category amplify these concerns, but it is important to keep in mind that the hallucination rates that we report for these tasks are only *lower bounds* on the true rates, as these tasks are evaluated using our reference-free method (Section 4.3). To provide some context for these bounds, we note that in a similar self-contradiction setup, Mündler et al. (2023) found that GPT 3.5 hallucinated about 14.3% of the time on general QA queries. On our legal QA queries, GPT 3.5 and our other LLMs far surpass this baseline rate—and it is possible that the true hallucination rate is even higher.

For example, we find that even on the easier reference-free tasks—**Factual background** and **Procedural posture**—our LLMs hallucinate at least 49% of the time. Performance degrades fur- ther on the most complex **Core legal question** and **Central holding** tasks, with hallucinations arising in response to at least 59% and 63% of our queries, respectively. Hallucinations are lowest among GPT 4 responses to the **Subsequent history** task at the SCOTUS level, but this is because the model simply tends to state that the litigation concluded with the Supreme Court decision. This may not actually be correct—many Supreme Court cases result in a remand and have additional procedural history in lower courts. However, we are unable to capture this kind of mistake, as our methodology only permits us to identify hallucinations where the model contradicts itself. We are not able to capture repeated incorrect answers as instances of hallucination, meaning that our estimate of hallucination in the SCOTUS **Subsequent history** task is likely to understate the rate of hallucination by a larger margin that other tasks.

Taken together, these results invite skepticism about LLMs' true knowledge of the law. Our reference-free tasks, in particular, raise serious doubts about LLMs' knowledge of substantive aspects of American case law—the very knowledge that attorneys must often synthesize themselves, instead of merely looking up in a database.

5.1.2 Hallucinations Vary by Court

We next examine trends by hierarchy, exploring LLMs' abilities to restate the case law of the three different levels of the federal judiciary. We find that across all tasks and all LLMs, hallucinations are lowest in the highest levels of the judiciary, and vice-versa (Figure 2). Thus, our LLMs perform best on tasks at the SCOTUS level, worse on tasks at the USCOA level, and worst on tasks at the USDC level. These results are encouraging insofar as it is important for LLMs to be knowledgeable about the most authoritative and wide-ranging precedents, but discouraging insofar as they suggest that LLMs are not well attuned to localized legal knowledge. After all, the vast majority of litigants do not appear before the Supreme Court and may benefit more from knowledge that is tailored to their home district court— their court of first appearance.

5.1.3 Hallucinations Vary by Jurisdiction

To better understand the relationship between different courts and hallucinations, we next zoom in on the middle level of the judicial hierarchy—the Courts of Appeals—and examine horizontal hetero- geneity across the circuits.[11] Figure 3 depicts these results geographically, showing lower hallucination

---

[11] Because not all Courts of Appeals were created at the same time, for parity in comparison here we exclude from our results cases decided before 1982, the year the youngest circuit—the Federal Circuit—was created. We report the full, non-truncated results in the Online Appendix, which are largely consistent with these post-1981 results.

Table 4 Hallucination rates across levels of the federal judiciary (moderate complexity tasks)

Task	Prompt	SCOTUS (1794–2015; n = 5,000)				USCOA (1895–2019; n = 5,000)				USDC (1932–2019; n = 5,000)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
Disposition	Zero-shot	0.399	0.499	0.500	0.536	0.494	0.500	0.500	0.493				
		(0.007)	(0.001)	(0.000)	(0.007)	(0.005)	(0.000)	(0.000)	(0.008)				
	Few-shot	0.452	0.496	0.501	0.502	0.498	0.489	0.501	0.501				
		(0.007)	(0.002)	(0.002)	(0.004)	(0.006)	(0.007)	(0.001)	(0.002)				
Quotation	Zero-shot	0.312	0.229	0.993	0.92	0.001	0.000	0.999	0.993	0.001	0.000	0.994	0.964
		(0.007)	(0.006)	(0.001)	(0.004)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.003)
	Few-shot	0.854	1.000	0.993	0.992	0.637	1.000	0.997	1.000	0.743	1.000	1.000	1.000
		(0.000)	(0.000)	(0.001)	(0.001)	(0.007)	(0.000)	(0.001)	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)
Authority	Zero-shot	0.922	0.937	0.985	0.991	0.934	0.978	0.997	0.999	0.956	0.87	0.995	0.999
		(0.004)	(0.003)	(0.002)	(0.001)	(0.004)	(0.002)	(0.001)	(0.001)	(0.003)	(0.005)	(0.001)	(0.000)
	Few-shot	0.828	0.916	0.953	0.993	0.958	0.976	0.993	0.999	0.928	0.956	0.994	0.997
		(0.005)	(0.004)	(0.003)	(0.001)	(0.003)	(0.002)	(0.001)	(0.000)	(0.004)	(0.003)	(0.001)	(0.001)
Overruling year[a]	Zero-shot	0.81	0.919	0.858	0.972								
		(0.025)	(0.017)	(0.022)	(0.011)								
	Few-shot	0.725	0.976	0.87	0.984								
		(0.028)	(0.010)	(0.021)	(0.008)								

[a] 1810–2022 (n = 279).
Note: Table reports estimated hallucination rates. Standard errors are shown in parentheses.

Table 5 Hallucination rates across levels of the federal judiciary (high complexity tasks)

Task	Prompt	SCOTUS (1794–2015; n = 100)				USCOA (1895–2019; n = 100)				USDC (1932–2019; n = 100)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
Doctrinal agreement	Zero-shot	0.461	0.500	0.466	0.500								
		(0.005)	(0.000)	(0.005)	(0.000)								
	Few-shot	0.449	0.458	0.453	0.500								
		(0.007)	(0.004)	(0.006)	(0.000)								
Factual background	Zero-shot	0.46	0.70	0.95	0.83	0.75	0.81	0.88	0.88	0.71	0.67	0.88	0.88
		(0.05)	(0.046)	(0.022)	(0.038)	(0.043)	(0.039)	(0.032)	(0.032)	(0.045)	(0.047)	(0.032)	(0.032)
Procedural posture	Zero-shot	0.47	0.65	0.74	0.85	0.66	0.6	0.68	0.83	0.73	0.55	0.81	0.89
		(0.05)	(0.048)	(0.044)	(0.036)	(0.047)	(0.049)	(0.047)	(0.038)	(0.044)	(0.05)	(0.039)	(0.031)
Subsequent history	Zero-shot	0.08	0.22	0.81	0.77	0.45	0.36	0.75	0.62	0.48	0.3	0.73	0.67
		(0.027)	(0.041)	(0.039)	(0.042)	(0.050)	(0.048)	(0.043)	(0.049)	(0.05)	(0.046)	(0.044)	(0.047)
Core legal question	Zero-shot	0.57	0.86	0.87	0.92	0.81	0.88	0.95	0.96	0.76	0.72	0.92	0.89
		(0.05)	(0.035)	(0.034)	(0.027)	(0.039)	(0.032)	(0.022)	(0.02)	(0.043)	(0.045)	(0.027)	(0.031)
Central holding	Zero-shot	0.6	0.73	0.78	0.92	0.83	0.87	0.95	0.95	0.77	0.73	0.93	0.84
		(0.049)	(0.044)	(0.041)	(0.027)	(0.038)	(0.034)	(0.022)	(0.022)	(0.042)	(0.044)	(0.026)	(0.037)

[a] 1796–2005 (n = 5,000).

Note: Table reports estimated hallucination rates. For all tasks except doctrinal agreement, this rate is only a lower bound on the true population rate. Standard errors are shown in parentheses.

rates in lighter colors and higher rates in darker colors. Pooling our tasks and models together, we see the best performance in the Ninth Circuit (comprising California and adjacent states in yellow), the Second Circuit (comprising New York and adjacent states in soft green), the Third Circuit (comprising Pennsylvania and adjacent states in soft green), and the First Circuit (comprising Maine and adjacent states in soft green). By contrast, performance tends to be worst in the circuits in the geographic center of the country.

These results confirm popular intuitions about the influential role that the Second, Third, and Ninth Circuits play in the American legal system. Because it encompasses New York City, the Second Circuit has traditionally had a significant impact on financial and corporate law, and many landmark decisions in securities law, antitrust, and business litigation have come from this court. The Third Circuit enjoys similar influence in the corporate law domain owing to Delaware's status as the legal home for many corporations. Finally, the Ninth Circuit handles more cases than any other federal appellate court, and often issues rulings that advance progressive positions that lead to disproportionate review by the Supreme Court.

Perhaps surprisingly, however, our results stand in tension with received wisdom about the D.C. Circuit, which is generally thought to be the *most* influential appellate division. In our tasks, our LLMs actually perform worst on this circuit. This counterintuitive finding is one example of the way that unanticipated biases might trouble the reliance on LLMs in practice.

5.1.4 Hallucinations Vary by Case Prominence

To probe the role of legal prominence more directly, we move to SCOTUS-level results next, examining the relationship between *case* importance and hallucinations. To measure case prominence within this single level of the judiciary, we use the Caselaw Access Project's PageRank percentile scores, a metric of citation network centrality that captures the general legal and political prominence of a case.

We find that case prominence is negatively correlated with hallucination, reaffirming our results from above (Figure 4). However, we also note that a sharp slope change occurs around the 90th prominence percentile in the GPT 4, GPT 3.5, and PaLM 2 models. This suggests that the bias of these LLMs—but not Llama 2—may be skewed even more toward the most well-known decisions of the American legal system, even within the SCOTUS level.

5.1.5 Hallucinations Vary by Case Year

Because case law develops in virtue of new decisions building on old ones over time, the age of a case may be another useful predictor of hallucination. Examining this relationship at the SCOTUS level in Figure 5, we find a non-linear correlation between hallucination and age: hallucinations are most common among the Supreme Court's oldest and newest cases, and least common among its post-war Warren Court cases (1953–1969). This result suggests another important limitation on LLMs' legal knowledge that users should be aware of: LLMs' peak performance may lag several years behind the current state of the doctrine, and LLMs may fail to internalize case law that is very old but still applicable and relevant law.
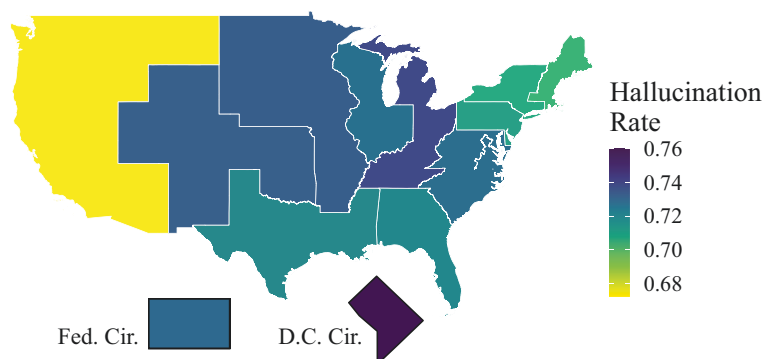

Figure 3. Relationship between USCOA jurisdiction and mean hallucination rate, all reference-based US-COA tasks and models pooled, post-1981 cases only. LLM performance is strongest in jurisdictions that are commonly perceived to play a more influential role.

5.1.6 Hallucinations Vary by LLM

Finally, we also partition our results by the LLM itself and compare across models. We find that not all LLMs are equal: as expected, GPT 4 performs best overall, followed by GPT 3.5, followed by PaLM 2, followed by Llama 2 (Figure 6).

We also discover tendencies towards different inductive biases, or the predisposition of an LLM to generate certain outputs more frequently than others. In Figure 7, we highlight one of these biases for our SCOTUS-level **Author** task, which asks the LLM to supply the name of the justice who authored the majority opinion in the given case. Each LLM we test has slightly different inductive preferences;
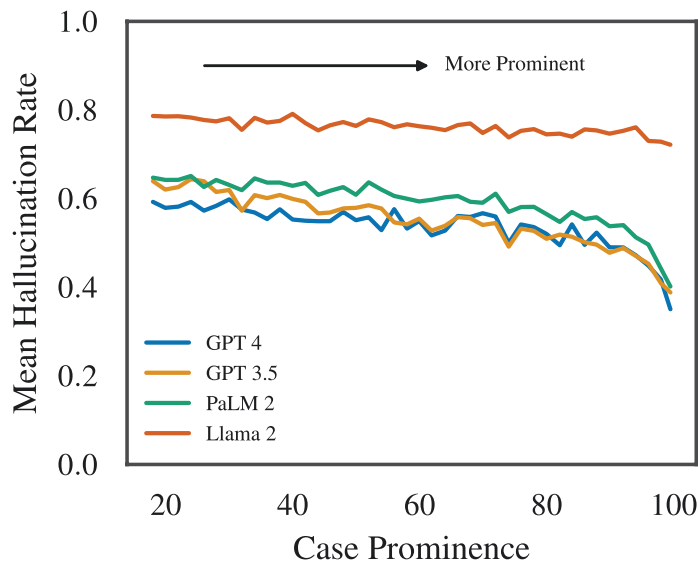
Figure 4. Relationship between SCOTUS case prominence (measured by PageRank percentile) and mean hallucination rate, all SCOTUS tasks pooled. Hallucinations decline sharply as case prominence passes the 90th percentile, meaning that LLMs are more likely to respond with accurate information about prominent cases.
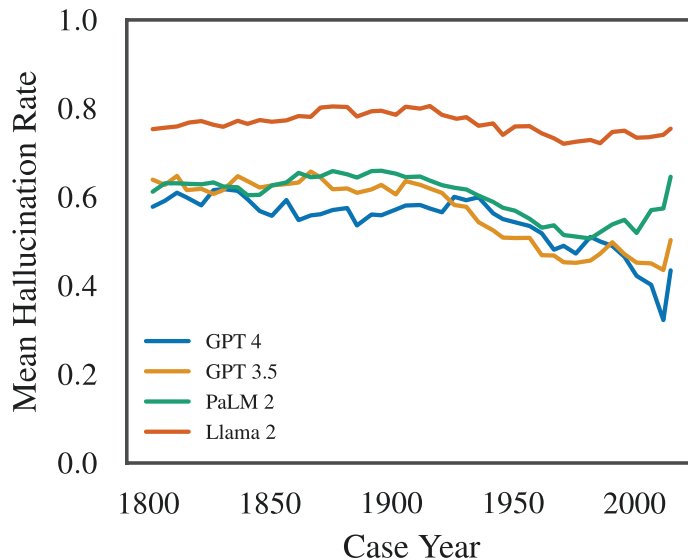
Figure 5. Relationship between SCOTUS case decision year and mean hallucination rate, all SCOTUS tasks pooled. LLMs are most likely to respond with accurate information in cases from the latter half of the 20th century, struggling on very old and very new cases.

some err towards the most recognizable justices, but others are a little more difficult to explain. For example, Llama 2 disproportionately favors Justice Story—an influential jurist who authored the famous *Amistad* opinion, among others—whereas PaLM 2 prefers Justice McLean—also an important jurist, but one more known for his dissents than his majority opinions, such as his dissent in the infamous *Dred Scott* case. Across the board, all our LLMs tend to overstate the true prevalence of justices at a higher magnitude than they understate them, as indicated by the greater dispersion of the points above the y = x line in Figure 7.

These biases demonstrate one way that LLMs inevitably encounter the kind of hallucination trade-off that we discuss in Section 2.3. If the inductive bias that an LLM learns from its training corpus is not
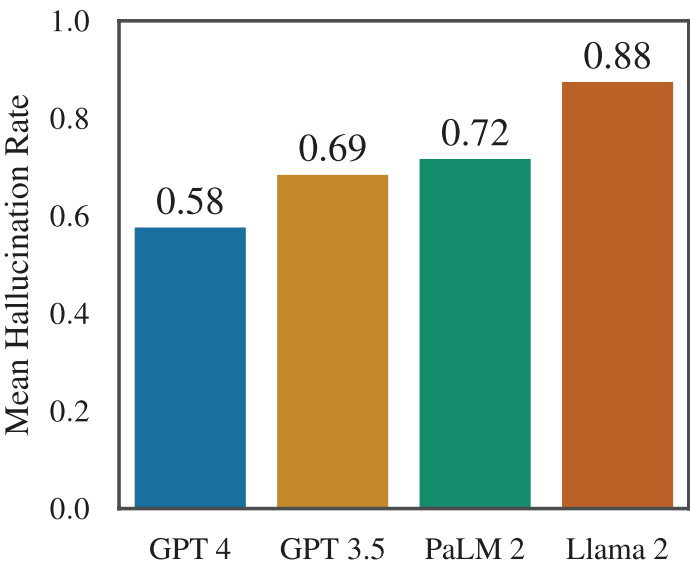
Figure 6. Hallucination rates by LLM, all reference-based tasks pooled. Hallucinations are common across all LLMs when they are asked a direct, verifiable question about a federal court case, but GPT 4 performs best overall.
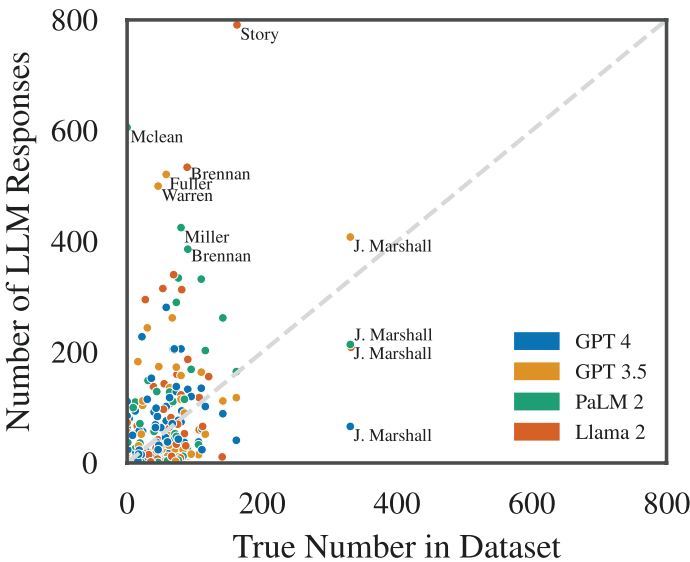
Figure 7. Number of times each justice is stated to be the author of a SCOTUS case versus the actual number of cases authored by each justice in our time period-stratified dataset. A small number of justices are disproportionately represented in LLM responses.

well-aligned with the true distribution of facts about the world, then the LLM is likely to make systematic errors when queried about those facts. Moreover, the persistence of inductive biases also increases the risk of LLMs instantiating a kind of legal monoculture (Kleinberg, Raghavan 2021). Instead of accurately restating the full variation of the law, LLMs may simply regurgitate information from a few prominent members of the response set that they have been trained on, flattening legal nuance and producing a falsely homogenous sense of the legal landscape.

5.2 Contra-factual Bias

We now turn to the first of two potential failure points that we seek to examine for LLMs performing legal tasks, beyond their sheer propensity to hallucinate: their bias toward accepting legal premises that are not anchored in reality and answering queries accordingly. We view this behavior as a particular kind of model sycophancy (the tendency of an LLM to agree with a user's preferences or beliefs, even when the LLM would reject the belief as wrong without the user's prompting; Sharma et al. 2023; Wei et al. 2023) or general cognitive error (Tversky and Kahneman 1974; Jones, Steinhardt 2022; Suri et al. 2023).

This bias poses a subtle but pernicious challenge to those aiming to use LLMs for legal re- search. When a researcher is learning about a topic, they are not only unsure about the *answer*, they are also often unsure about the *question* they are asking as well. Worse, they might not even be aware of any defects in their query; research by its nature ventures into the realm of "unknown unknowns" (Luft and Ingham 1955). This is especially true for unsophisticated *pro se* litigants, or those without much legal training to begin with. Relying on an LLM for legal research, they might inadvertently submit a question premised on non-factual legal information or folk wisdom about the law. As discussed in Section 2.3, this then forces a trade-off: if the LLM is too intent on minimizing prompt hallucinations, it runs the risk of simply accepting the user's misconception as true and producing a factual hallucination instead.

To test whether this risk is real in the legal setting, we evaluate two modified versions of our reference-based queries, but with premises that are false by construction. Specifically, we ask the LLMs to (i) provide information about an author's dissenting opinion in an appellate case in which they did not in fact dissent and (ii) furnish the year that a SCOTUS case that has never been overruled was overruled. In both cases, we consider failing to provide the requested information an acceptable answer; any uncritical answering of the prompt is treated as a hallucination.

Table 6 reports the results of this experiment and Figure 8 summarizes them by LLM. In general, LLMs seem to suffer from contra-factual bias on these legal information tasks. As in the raw hallucination tasks, contra-factual bias hallucinations are higher in lower levels of the judiciary. Substantively, they are also greatest for the question with a false overruling premise, possibly reflecting the increased complexity of the question asked.

Llama 2 performs exceptionally well, demonstrating little contra-factual hallucination. However, this success is linked to a different kind of hallucination—in many false dissent examples, for instance, Llama 2 often states that the case or justice does not exist at all. (In reality, all of our false dissent examples were created with real cases and real justices—just justices who did not author a dissent for the case.) Under our metrics for contra-factual hallucination, we choose to record these examples as successful rejections of the premise. The kind of error that Llama 2 makes here is already measured in its poor performance on other tasks, especially **Existence**.

5.3 Model Calibration

The second potential hazard that we investigate is model calibration, or the ability of LLMs to "know what they know." Ideally, a well-calibrated model would be confident in its factual responses, and not confident in its hallucinated ones (Azaria and Mitchell 2023; Kadavath et al. 2022; Tian, Mitchell, Zhou, et al. 2023; Xiong et al. 2023; Yin et al. 2023). If this property held for legal queries, users would be able to adjust their expectations accordingly and could theoretically learn to trust the LLM when it is confident, and learn to be more skeptical when it is not (Zhang, Liao, and Bellamy 2020). Even more importantly, if an LLM knew when it was likely to be hallucinating, the hallucination problem could be in principle solvable through some form of reinforcement learning from human feedback (RLHF) or fine-tuning, with unconfident answers simply being suppressed (Tian, Mitchell, Yao, et al. 2023).

To study our LLMs' calibration on legal queries, we estimate the expected calibration error (ECE) for each of our tasks. We describe our estimation strategy in full in the Online Appendix, but, intuitively,

Table 6 Hallucination rates across levels of the federal judiciary (contra-factual tasks)

Task	Prompt	SCOTUS (1794–2015; n = 1,000)				USCOA (1895–2019; n = 1,000)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
False dissent premise	Zero-shot	0.691	0.338	0.990	0.000	0.842	0.408	0.983	0.021
		(0.015)	(0.015)	(0.003)	(0.000)	(0.012)	(0.016)	(0.004)	(0.005)
False overruling premise	Zero-shot	0.531	0.821	1.000	0.027				
		(0.016)	(0.012)	(0.000)	(0.005)				

Note: Table reports estimated hallucination rates. Standard errors are shown in parentheses.

it entails extracting a confidence score for each LLM answer that we obtain and comparing it to the empirical hallucination rate that we observe. Table 7 reports the results of this analysis at the task level, and Figure 9 pools our findings at the LLM level by plotting those two metrics—confidences and empirical non-hallucination frequencies—against each other, binned into 10 equally-sized bins (represented by the dots). In a perfectly calibrated model, the confidences and empirical frequencies would be perfectly correlated along the y = x diagonal.

Overall, we note that PaLM 2 (pooled ECE = 0.057), GPT 3.5 (pooled ECE = 0.099), and GPT 4 (pooled ECE = 0.190) are significantly better calibrated than Llama 2 (pooled ECE = 0.421). Interestingly, although GPT 4 is our best performing model in terms of raw hallucination rates (Figure 6), it is actually less calibrated than PaLM 2 and GPT 3.5, which are otherwise inferior. This suggests that even the newest and most advanced LLMs may not always be superior in all desirable senses—although GPT 4 is currently the LLM least prone to hallucination, our results imply that when it *does* hallucinate, it does so in a way that is more likely to mislead users than GPT 3.5 or PaLM 2.

Diving into the task-level results (Table 7), we see that across all LLMs, calibration is poorer on our more complex tasks, like **Doctrinal agreement**, and on tasks directed toward lower levels of the judicial hierarchy. ECE is also higher on our partially open-ended tasks such as **Court** and **Author**. In these tasks, the LLM has a large but finite universe of responses, and the high ECE for these tasks reflects the LLMs' tendencies to over-report on the most prominent or widely known members of the response set.

In all cases, the calibration error is in the positive direction: our LLMs systematically *overestimate* their confidence relative to their actual rate of hallucination.[12] This finding, too, suggests that users should exercise caution when interpreting LLMs' responses to legal queries, especially those of Llama 2. Not only may they receive a hallucinated response, but they may receive one that the LLM is overconfident in and liable to repeat again.

6 Discussion

We began this article with a question that has surged in salience over the last twelve months: Will AI systems like ChatGPT soon reshape the practice of law and democratize access to justice? Although there is much enthusiasm for LLMs' potential to revolutionize these domains, we highlight the problem
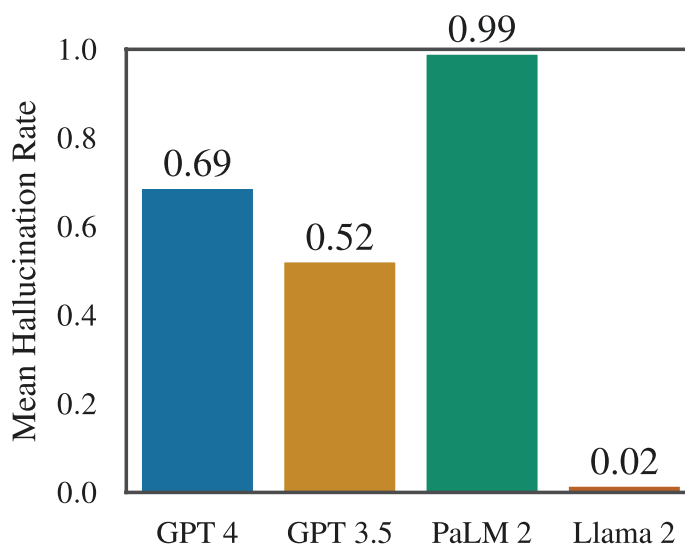
Figure 8. Hallucination rates by LLM, all contra-factual tasks pooled. Llama 2 is very unlikely to hallucinate on these tasks because it almost always rejects the premise in the question. However, this tendency also leads it to perform more poorly on tasks with correct premises (cf. Figure 6).

---

[12] In the Online Appendix, we explore whether this bias can be corrected with an *ex post* scaling adjustment, but conclude that challenges remain.

Table 7 Expected calibration error (ECE) across levels of the federal judiciary

Task	Prompt	SCOTUS (1794–2015; n = 100)				USCOA (1895–2019; n = 100)				USDC (1932–2019; n = 100)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
Existence	Zero-shot	0.026	0.008	0.119	0.262	0.124	0.007	0.182	0.117	0.074	0.004	0.183	0.178
		(0.005)	(0.001)	(0.004)	(0.006)	(0.005)	(0.001)	(0.003)	(0.005)	(0.004)	(0.000)	(0.003)	(0.006)
	Few-shot	0.149	0.038	0.041	0.998	0.073	0.064	0.014	0.998	0.024	0.053	0.038	0.096
		(0.005)	(0.002)	(0.002)	(0.000)	(0.005)	(0.002)	(0.001)	(0.000)	(0.003)	(0.002)	(0.001)	(0.000)
Court	Zero-shot	0.000	0.000	0.000	0.003	0.362	0.322	0.184	0.387	0.133	0.163	0.132	0.163
		(0.000)	(0.000)	(0.000)	(0.000)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)
	Few-shot	0	0.001	0	0.009	0.345	0.355	0.161	0.452	0.145	0.099	0.094	0.099
		(0.000)	(0.000)	(0.000)	(0.001)	(0.006)	(0.006)	(0.006)	(0.008)	(0.005)	(0.004)	(0.005)	(0.005)
Citation	Zero-shot	0.113	0.069	0.026	0.068	0.143	0.073	0.004	0.036	0.104	0.043	0.002	0.022
		(0.005)	(0.004)	(0.003)	(0.003)	(0.004)	(0.005)	(0.001)	(0.001)	(0.003)	(0.003)	(0.000)	(0.001)
	Few-shot	0.191	0.029	0.059	0.063	0.138	0.036	0.01	0.032	0.099	0.018	0.001	0.051
		(0.005)	(0.003)	(0.004)	(0.003)	(0.005)	(0.003)	(0.001)	(0.001)	(0.004)	(0.002)	(0.000)	(0.002)
Author	Zero-shot	0.443	0.208	0.130	0.419	0.349	0.094	0.131	0.583	0.302	0.228	0.063	0.305
		(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.003)	(0.002)	(0.004)	(0.005)	(0.005)	(0.002)	(0.004)
	Few-shot	0.454	0.347	0.142	0.454	0.357	0.142	0.118	0.656	0.32	0.096	0.045	0.481
		(0.006)	(0.006)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.005)
Disposition	Zero-shot	0.203	0.431	0.291	0.199	0.17	0.557	0.382	0.148				
		(0.007)	(0.007)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)				
	Few-shot	0.239	0.314	0.165	0.283	0.165	0.205	0.227	0.439				
		(0.006)	(0.008)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)				
Overruling year	Zero-shot	0.308	0.246	0.116	0.510								
		(0.025)	(0.022)	(0.019)	(0.019)								
	Few-shot	0.377	0.68	0.154	0.754								
		(0.025)	(0.022)	(0.019)	(0.018)								

Table 7. Continued

Task	Prompt	SCOTUS (1794–2015; n = 100)				USCOA (1895–2019; n = 100)				USDC (1932–2019; n = 100)			
		GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2	GPT 4	GPT 3.5	PaLM 2	Llama 2
Doctrinal agreement	Zero-shot	0.369	0.527	0.165	0.564								
		(0.006)	(0.007)	(0.006)	(0.006)								
	Few-shot	0.319	0.409	0.152	0.548								
		(0.007)	(0.008)	(0.007)	(0.007)								

[a] 1810–2022 (n = 279)
[b] 1796–2005 (n = 5,000).
Note: Table reports expected calibration error between empirical hallucination rates and estimated conditional probabilities. Conditional probabilities are estimated by sampling 10 responses from the model at temperature 1 and assessing their agreement with the model's greedy response. Bootstrapped standard errors are shown in parentheses.

of legal hallucinations, which remains a serious obstacle to the adoption of these models. Performing the first systematic empirical test of popular perceptions (Roberts 2023; Romoser 2023; Weiser 2023), we show that factual legal hallucinations are widespread in the LLMs that we study—OpenAI's ChatGPT 4, OpenAI's ChatGPT 3.5, Google's PaLM 2, and Meta's Llama 2—on the bulk of the legal knowledge tasks that we profile (Section 5.1).

We also push beyond conventional wisdom by documenting the correlates of these hallucinations and by surfacing two additional behaviors that threaten LLMs' utility for legal applications: (i) their susceptibility to contra-factual bias, i.e., their inability to handle queries containing an erroneous or mistaken starting point (Section 5.2), and (ii) their certainty in their responses, i.e., their inability to always "know what they know" (Section 5.3). Unfortunately, we find that LLMs frequently provide seemingly genuine answers to legal questions whose premises are false by construction, and that under their default configurations they are imperfect predictors of their own tendency to confidently hallucinate legal falsehoods.

These findings complicate the existing literature that suggests that LLMs are performing increasingly well on a number of legal benchmarking tasks (Ash et al. 2024; Blair-Stanek, Holzenberger, and Van Durme 2023; Choi et al. 2022; Fei et al. 2023; Guha et al. 2023; Nay et al. 2023; Katz et al. 2023; Trozze, Davies, and Kleinberg 2023). Our study is related to this prior research, but is oriented in a slightly different vein. Instead of examining LLMs' ability to engage in le-gal *reasoning*, we assess LLMs' capacity to internalize legal *knowledge*. Ultimately, LLMs will need to excel in both of these respects if they are going to be effectively integrated into the legal profession. So long as they suffer from gaps in their background legal knowledge—as our results suggest—they will be unable to function as reliable sources of legal counsel and advice, no matter how strong their in-context reasoning abilities become.

Our results therefore temper optimism for the ability of off-the-shelf, publicly available LLMs to accelerate access to justice (Perlman 2023; Tan, Westermann, and Benyekhlef 2023; Tito 2017). Indeed, our findings suggest that the risks of using these generic foundation models are especially high for litigants who are:

1. Filing in courts lower in the judicial hierarchy or those located in less prominent jurisdictions,
2. Seeking more complex forms of legal information,
3. Formulating questions with mistaken premises, or
4. Unsure of how much to trust the LLMs' responses.

In short, we find that the risks are *highest* for those who would benefit from LLMs *most*—under-resourced or *pro se* litigants. Some of these risks—namely, (3) and (4)—might be mitigated with improved user
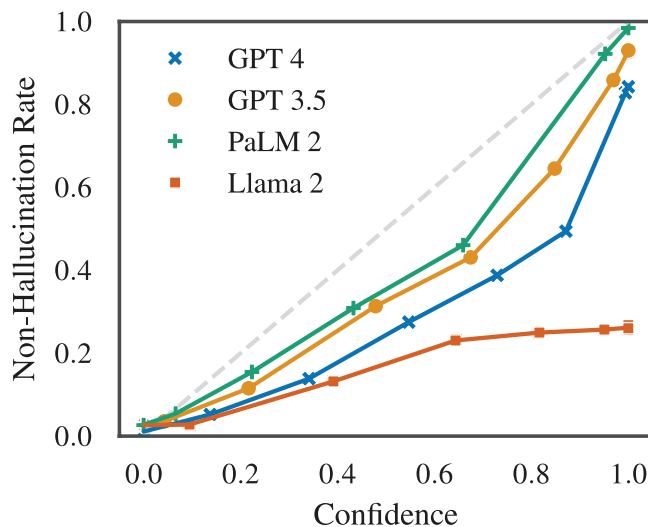
Figure 9. Calibration curves by LLM, all reference-based tasks pooled. PaLM 2 is best calibrated on legal queries, followed by GPT 3.5, GPT 4, and lastly Llama 2, which is significantly worse than the first three models.

education, but others—(1) and (2)—are more intractable. LLMs would ideally do best at localized legal information (rather than SCOTUS-level information), be able to correct users when they ask misguided questions (rather than accepting their premises at face value), and be able to moderate their responses with the appropriate level of confidence (rather than hallucinating with conviction). Consequently, we echo concerns that the proliferation of LLMs may ultimately exacerbate, rather than eradicate, existing inequalities in access to legal services (Draper and Gillibrand 2023; Simshaw 2022). At the same time, increased reliance on LLMs also has the potential to produce a kind of legal monoculture (Creel and Hellman 2022; Kleinberg and Raghavan 2021), with users being fed information from only a limited subset of judicial sources that elide many of the deeper nuances of the law. This new monoculture, in turn, is likely to reify the geographic, temporal, and judge-level biases that we diagnose above, as the foundation-like property of these models permits those biases to propagate into any downstream tools built on top of the original LLM (Bommasani et al. 2022).

Some recent research suggests that hallucinations can be diminished through the adoption of techniques like retrieval-augmented generation (RAG) (Shuster et al. 2021; Cui et al. 2023; Peng et al. 2023; Savelka et al. 2023), advanced prompting (such as chain-of-thought prompting or chain-of-verification) (Si et al. 2023; Lei et al. 2023; Mündler et al. 2023; Ji, Yu, et al. 2023; Dhuliawala et al. 2023; Suzgun and Kalai 2024), specialized fine-tuning (Tian, Mitchell, Yao, et al. 2023; Razumovskaia et al. 2023; Zhang et al. 2023), factuality-focused decoding methods (Shi et al. 2023; Mallen et al. 2023; Li et al. 2024; Chuang et al. 2024), or external database checks (Chern et al. 2023; Peng et al. 2023; Qin et al. 2023; Gou et al. 2024). These methods have shown promising results in significantly reducing hallucinated content and enhancing the accuracy, reliability, and faithfulness of model outputs. However, we caution that these approaches are not without limitations.

For example, the effectiveness of RAG-based methods heavily relies on the quality of their retrieval mechanisms (Wu et al. 2024). Moreover, accurately parsing and understanding the content of input queries poses a challenge, especially when queries are inherently ambiguous or irrelevant to the domain of focus (Tonmoy et al. 2024). Additionally, the task of retrieving relevant and precise information from extensive corpora can be computationally demanding and resource- intensive, necessitating continuous updating and modification of knowledge databases to keep pace with the latest information (Chen et al. 2023; Siriwardhana et al. 2023; Ram et al. 2023; Cheng et al. 2024). There may also be situations where the knowledge database might contain conflicting or contradictory information, making it unclear which pieces of relevant information to extract (Wang et al. 2023; Yu et al. 2023; Gao et al. 2023). For instance, when a legal case is overruled, or when there is a circuit split on an issue, the retrieval module must have some mechanism to distinguish outdated or jurisdictionally irrelevant sources from apposite and binding law.

Furthermore, methods for detecting hallucinations and evaluating their mitigation are themselves not foolproof. Evaluation datasets and metrics may not always accurately reflect real-world performance reliability (Ji et al. 2023; Lucas et al. 2023; Zhang et al. 2023). Biases could be embedded within the evaluation dataset, or the automated metric employed to quantify hallucination may lack comprehensiveness or task-specificity (Kang, Blevins, and Zettlemoyer 2024). Therefore, it is far from clear whether these technical improvements will be able to truly solve the hallucination problem.

Finally, we also emphasize that the challenges presented by legal hallucinations are not only empirical, but also normative. Although data-rich and moneyed players certainly stand at an advantage when it comes to building hallucination-free legal LLMs for their own private use, it is not clear that even infinite resources can entirely solve the conceptual problems we diagnose. As we discuss in Section 2.3, model fidelity to the training corpus, model fidelity to the user's prompt, and model fidelity to the facts of the world—i.e., the law—are normative commitments that stand in tension with each other, despite all being independently desirable technical properties of an LLM. Ultimately, since hallucinations of *some* kind are generally inevitable at the token level (Kalai and Vempala 2023; Xu, Jain, and Kankanhalli 2024), developers of legal LLMs will need to make choices about which type(s) of hallucinations to minimize, and they should make these choices transparent to their downstream users. Only then can individual litigants decide for themselves whether the legal information they seek to obtain from LLMs is trustworthy or not.

To aid in future research in this area, we release a test dataset of our queries and answers on the HuggingFace platform, which scholars can use to continue to evaluate LLMs as they advance in legal

sophistication.[13] In the meantime, more experienced legal practitioners may find some value in consulting LLMs for certain tasks, but even these users should remain vigilant in their use, taking care to verify the accuracy of their prompts and the quality of their chosen LLM's responses. Similarly, legal scholars and educators seeking to use LLMs as automated research assistants (Livermore, Herron, and Rockmore 2024) or student aids (Choi and Schwarcz 2024) must be cautious to not inadvertently inject these LLMs' subtle knowledge biases into their own applications. Put differently, our findings underscore the importance of human-centered AI. Responsible integration of LLMs into legal tasks must *augment* lawyers, researchers, and litigants and not, as Chief Justice Roberts has put it, risk "dehumanizing the law" (Roberts 2023, 5).

Supplementary Material

Supplementary material is available at *JLA* online.

References

Agrawal, Ayush, *et al.* 2023. Do Language Models Know When They're Hallucinating References? *Findings of the Association for Computational Linguistics: EACL 2024, 912-928.* St. Julian's, Malta: Association for Computational Linguistics. https://aclanthology.org/2024.findings-eacl.62

Ambrogi, Bob. 2023. As Allen & Overy Deploys GPT-based Legal App Harvey Firmwide, Founders Say Other Firms Will Soon Follow. *LawSites*. https://www.lawnext.com/2023/02/as-allen-overy-deploys-gpt-based-legal-app-harvey-firmwide-founders-say-other-firms-will-soon-follow.html.

Anil, Rohan, *et al.* 2023. PaLM 2 Technical Report. doi:10.48550/arXiv.2305. 10403

Ash, Elliott, *et al.* 2024. Translating Legalese: Enhancing Public Understanding of Court Opinions with Legal Summarizers In Proceedings of the Symposium on Computer Science and Law, 136–57. CSLAW '24. NY, USA: Association for Computing Machinery.

Azaria, Amos, & Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. doi:10.48550/arXiv.2304.13734

Bar-Gill, Oren, Cass R. Sunstein & Inbal Talgam-Cohen. 2023. Algorithmic Harm in Consumer Markets. 15 *J. Legal Anal.* 1–47.

Barocas, Solon & Andrew D. Selbst. 2016. Big Data's Disparate Impact. 104 *California Law Rev.* 671–732.

Ben-Shahar, Omri. 2023. Privacy Protection, At What Cost? Exploring the Regulatory Resistance to Data Technology in Auto Insurance. 15 *J. Legal Anal.* 129–57.

Black, Ryan C & James F. Spriggs II. 2013. The Citation and Depreciation of U.S. Supreme Court Precedent. 10 *J. Empirical Legal Stud.* 325–58.

Blair-Stanek, Andrew, Nils Holzenberger, and Benjamin Van Durme. 2023. "Can GPT-3 Perform Statutory Reasoning?" In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. Braga, Portugal: Association for Computing Machinery.

Bommasani, Rishi, *et al.* 2022. On the Opportunities and Risks of Foundation Models. doi:10.48550/arXiv.2108.07258

Brown, Tom B., *et al.* 2020. Language Models Are Few-Shot Learners. doi: 10.48550/arXiv.2005.14165

Cao, Meng, Yue Dong, & Jackie Cheung. 2022. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In Smaranda Muresan, Preslav Nakov, & Aline Villavicencio, eds., Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3340–54. Dublin, Ireland: Association for Computational Linguistics.

Cao, Ziqiang, *et al.* 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. 32 *Proceedings of the AAAI Conference on Artificial Intelligence* 4784–4791.

Caselaw Access Project. 2023. Caselaw Access Project. https://case.law.

Carlini, Nicholas, *et al.* 2022. Quantifying Memorization Across Neural Language Models. doi:10. 48550/arXiv.2202.07646

Chann, Seherman. 2023. Non-Determinism in GPT-4 Is Caused by Sparse MoE. https://152334H.github.io/blog/non-determinism-in-gpt-4/.

---

[13] https://huggingface.co/datasets/reglab/legal_hallucinations

Chen, Jiawei, *et al*. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. 38 AAAI Conference on Artificial Intelligence 17754–62. doi: 10.48550/arXiv.2309.01431

Cheng, Xin, *et al*. 2024. Lift Yourself Up: Retrieval-augmented Text Generation with Self Memory. 37 *Conference on Neural Information Processing Systems* 1–20.

Chern, I, *et al*. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. doi:10.48550/arXiv.2307.13528

Chien, Colleen V, *et al*. 2024. How Generative AI Can Help Address the Access to Justice Gap Through the Courts. *Loyola Los Angeles Law Rev.*, forthcoming.

Choi, Jonathan H, *et al*. 2022. "ChatGPT Goes to Law School. 71 *J. Legal Educ.* 387–400.

Choi, Jonathan H & Daniel Schwarcz. 2024. AI Assistance in Legal Analysis: An Empirical Study. *J. Legal Educ.* doi: 10.2139/ssrn.4539836. forthcoming

Chuang, Yung-Sung, *et al*. 2024. "DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models." In The Twelfth International Conference on Learning Representations.

Congress.gov. 2023. Table of Supreme Court Decisions Overruled by Subsequent Decisions. https://constitution.congress.gov/resources/decisions-overruled/.

Creel, Kathleen & Deborah Hellman. 2022. The Algorithmic Leviathan: Arbitrariness, Fair- ness, and Opportunity in Algorithmic Decision-Making Systems. 52 *Canad. J. Philos.* 26–43.

Cui, Jiaxi, *et al*. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. doi:10.48550/arXiv.2306.16092

Cushman, Jack, Matthew Dahl & Michael Lissner. 2021. Eyecite: A Tool for Parsing Legal Citations. 6 *Open Source Software* 3617.

Dell'Acqua, Fabrizio, *et al*. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. doi:10.2139/ssrn.4573321

Deroy, Aniket, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How Ready Are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization? doi:10.48550/arXiv.2306.01248

Dhuliawala, Shehzaad, *et al*. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint arXiv:2309.11495* doi: 10.48550/arXiv.2309.11495

Draper, Chris, and Nicky Gillibrand. 2023. The Potential for Jurisdictional Challenges to AI or LLM Training Datasets. In *Proceedings of the ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice*. Braga, Portugal: CEUR Workshop Proceedings.

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

Eloundou, Tyna, *et al*. 2023. GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. doi:10.48550/arXiv.2303.10130

Engel, Christoph & Nina Grgić-Hlacă. 2021. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. 13 *J. Legal Anal.* 284–340.

Engstrom, David Freeman, Daniel E. Ho. 2020. Algorithmic Accountability in the Administrative State. 37 *Yale J. Regul.* 800–54.

Engstrom, David Freeman, *et al*. 2020. Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. Technical report. Administrative Conference of the United States.

Fei, Zhiwei, *et al*. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. doi:10.48550/arXiv.2309.16289

Feijo, Diego de Vargas, & Viviane P. Moreira. 2023. Improving Abstractive Summarization of Legal Rulings through Textual Entailment. 31 *Artif. Intell. Law* 91–113.

Feng, Shangbin, *et al*. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. doi:10.48550/arXiv.2402.00367

Fowler, James H, *et al*. 2007. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. 15 *Polit. Anal.* 324–46.

Gao, Yunfan, *et al*. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. doi:10.48550/arXiv.2312.10997

Gillis, Talia B & Jann L. Spiess. 2019. Big Data and Discrimination. 86 *Univ. Chicago Law Rev.* 459–88.

Gou, Zhibin, *et al*. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In The Twelfth International Conference on Learning Representations.

Guha, Neel, *et al*. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Henderson, Peter, Tatsunori Hashimoto & Mark Lemley. 2023. Where's the Liability in Harmful AI Speech?. 3 *J Free Speech Law* 589–650.

Henderson, Peter, *et al*. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. doi:10.48550/arXiv.2207.00220.

Ji, Ziwei, *et al*. 2023. Survey of Hallucination in Natural Language Generation. 55 *ACM Comput. Surv.* 248:1–248:38.

Ji, Ziwei, *et al*. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In Houda Bouamor, Juan Pino, & Kalika Bali, eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–43. Singapore: Association for Computational Linguistics.

Jones, Erik, and Jacob Steinhardt. 2022. Capturing Failures of Large Language Models via Human Cognitive Biases. 36 Conference on Neural Information Processing Systems 1-15. doi:10.48550/arXiv.2202.1229.

Kadavath, Saurav, *et al*. 2022. Language Models (Mostly) Know What They Know. doi:10.48550/arXiv.2207.05221

Kalai, Adam Tauman, & Santosh S. Vempala. 2023. Calibrated Language Models Must Hallucinate. doi:10.48550/arXiv.2311.14648

Kang, Haoqiang, Terra Blevins, & Luke Zettlemoyer. 2024. Comparing Hallucination Detection Metrics for Multilingual Generation. doi:10.48550/arXiv.2402.10496

Katz, Daniel Martin, *et al*. 2023. GPT-4 Passes the Bar Exam. doi:10.2139/ssrn.4389233

King, Jennifer, *et al*. 2023. The Privacy- Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, 492–505. Chicago, IL, USA: Association for Computing Machinery.

Kleinberg, Jon, *et al*. 2018. Discrimination in the Age of Algorithms. 10 *J. Legal Anal.* 113–74.

Kleinberg, Jon & Manish Raghavan. 2021. Algorithmic Monoculture and Social Welfare. 118 *Proc. Natl. Acad. Sci. U.S.A* 1–7.

Krishna, Kalpesh, Aurko Roy, & Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4940–57. Association for Computational Linguistics.

Kuersten, Ashlyn K., & Susan B. Haire. 2011. Update to the Appeals Courts Database (1997–2002). https://artsandsciences.sc.edu/poli/juri/appct.htm

Lee, Minhyeok. 2023. A Mathematical Investigation of Hallucination and Creativity in GPT Models. 11 *Mathematics* 2320.

Lee, Nayeon, *et al*. 2023. Factuality Enhanced Language Models for Open-Ended Text Generation. doi:10.48550/arXiv.2206.04624

Lei, Deren, *et al*. 2023. Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations. doi:10.48550/arXiv.2310.03951

Lemley, Mark A & Bryan Casey. 2019. Remedies for Robots. 86 *Univ. Chicago L. Rev.* 1311–96.

Li, Junyi, *et al*. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. doi: 10.48550/arXiv.2305.11747

Li, Kenneth, *et al*. 2024. Inference-Time Intervention: Eliciting Truthful Answers From a Language Model.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. doi:10.48550/arXiv.2109.07958

Livermore, Michael A, Felix Herron & Daniel Rockmore. 2024. Language Model Interpretability and Empirical Legal Studies. *J. Institut. Theoreti. Econ.*, forthcoming

Lucas, Jason, *et al*. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In Houda Bouamor, Juan Pino, and Kalika Bali eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14279–305. Singapore: Association for Computational Linguistics.

Luft, Joseph, and Harrington Ingham. 1955. The Johari Window as a Graphic Model of Interpersonal Awareness. In Proceedings of the Western Training Laboratory in Group Development. University of California, Los Angeles, Extension Office.

Mallen, Alex, *et al*. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics *(Volume* 1: *Long Papers),* 9802–22.

Manakul, Potsawee, Adian Liusie, & Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. doi:10.48550/arXiv.2303.08896

Maynez, Joshua, *et al*. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906–19. Association for Computational Linguistics.

Mayson, Sandra G. 2019. Bias In, Bias Out. 128 *Yale Law J.* 2122–473.

Min, Sewon, *et al*. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. doi:10.48550/arXiv. 2305.14251

Mündler, Niels, *et al*. 2023. Self-Contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. doi: 10.48550/arXiv.2305.15852

Nay, John J., *et al*. 2023. Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence. doi:10.48550/arXiv.2306.07075

OpenAI. 2023a. GPT-4 Technical Report. doi:10.48550/arXiv.2303.08774

———. 2023b. *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Peng, Baolin, *et al*. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. doi:10.48550/arXiv.2302.12813.

Perlman, Andrew. 2023. The Implications of ChatGPT for Legal Services and Society. *The Practice*, March/April. https://clp.law.harvard.edu/knowledge-hub/magazine/issues/generative-ai-in-the-legal-profession/the-implications-of-chatgpt-for-legal-services-and-society/.

Qin, Yujia, *et al*. 2023. Tool Learning With Foundation Models. doi:10.48550/arXiv.2304.08354

Ram, Ori, *et al*. 2023. In-Context Retrieval-Augmented Language Models. 11 *Trans. Assoc. Computat. Linguist.* 1316–31.

Razumovskaia, Evgeniia. 2023. Dial BeInfo for Faithfulness: Improving Factuality of Information-Seeking Dialogue via Behavioural Fine-Tuning. doi:10.48550/arXiv.2311.09800

Roberts, John G. 2023. 2023 Year-End Report on the Federal Judiciary. https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf

Rodgers, Ian, John Armour & Mari Sako. 2023. How Technology Is (or Is Not) Transforming Law Firms. 19 *Ann. R. Law Social Sci.* 299–317.

Romoser, James. 2023. No, Ruth Bader Ginsburg Did Not Dissent in Obergefell — and Other Things ChatGPT Gets Wrong about the Supreme Court. *SCOTUSblog*. https://www.scotusblog.com/2023/01/no-ruth-bader-ginsburg-did-not-dissent-in-obergefell-and-other-things-chatgpt-gets-wrong-about-the-supreme-court/.

Savelka, Jaromir, *et al*. 2023. Explaining Legal Concepts with Augmented Large Language Models (GPT-4). doi:10.48550/arXiv.2306.09525

Sharma, Mrinank, *et al*. 2023. Towards Understanding Sycophancy in Language Models. doi:10.48550/arXiv.2310.13548

Shi, Weijia, *et al*. 2023. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. doi:10.48550/arXiv.2305.14739

Shuster, Kurt, *et al*. 2021. "Retrieval Aug- mentation Reduces Hallucination in Conversation." In Findings of the Association for Computational Linguistics: EMNLP 2021, 3784–803. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Si, Chenglei, *et al*. 2023. "Prompting GPT-3 To Be Reliable." In *The Eleventh Inter-national Conference on Learning Representations*.

Simshaw, Drew. 2022. Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services. 24 *Yale J. Law Technol.* 150–226.

Siriwardhana, Shamane, *et al*. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. 11 *Trans. Assoc. Comput. Linguist.* 1–17.

Solow-Niederman, Alicia. 2020. Administering Artificial Intelligence. 93 *Southern California Law Rev.* 633–96.

Songer, Donald R. 2008. *The United States Courts of Appeals Database, 1925–1996.*

Spaeth, Harold J., *et al*. 2022. *2022 Supreme Court Database, Version 2022 Release 01*. http://supremecourtdatabase.org/

Suri, Gaurav, *et al*. 2023. *Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5*. doi: 10.48550/arXiv.2305.04400

Suzgun, Mirac, & Adam Tauman Kalai. 2024. *Meta-prompting: Enhancing Language Models with Task-Agnostic Scaffolding*. doi:10.48550/arXiv.2401.12954

Tan, Jinzhe, Hannes Westermann, & 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.

Tian, Katherine, *et al*. 2023. *Fine-Tuning Language Models for Factuality*. doi:10.48550/arXiv.2311.08401

Tian, Katherine, *et al*. 2023. *Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback*. doi:10.48550/arXiv.2305.14975

Tito, Joel. 2017. *How AI Can Improve Access to Justice*. https://www.centreforpublicimpact.org/insights/joel-tito-ai-justice

Tonmoy, SM, *et al*. 2024. *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. doi:10.48550/arXiv.2401.01313

Touvron, Hugo, *et al*. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. doi:10.48550/arXiv.2307.09288

Trozze, Arianna, Toby Davies, & Bennett Kleinberg. 2023. *Large Language Models in Cryptocurrency Securities Cases: Can ChatGPT Replace Lawyers?* doi:10.48550/ arXiv.2308.06032

Tversky, Amos & Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. 185 *Science* 1124–31.

van Deemter, Kees. 2024. The Pitfalls of Defining Hallucination. *Comput. Linguist.* 1–10, forthcoming

Volokh, Eugene. 2023. Large Libel Models? Liability for AI Output. 3 *J. Free Speech Law* 489–558.

Wang, Cunxiang, *et al*. 2023. *Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity*. doi:10.48550/ arXiv.2310.07521

Wei, Jerry, *et al*. 2023. *Simple Synthetic Data Reduces Sycophancy in Large Language Models*. doi:10.48550/arXiv.2308.03958

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

Wittgenstein, Ludwig. 1998 [1921]. *Tractatus Logico-Philosophicus*. Translated by C. K. Ogden. Dover.

Wu, Kevin, *et al*. 2024. *How Well Do LLMs Cite Relevant Medical References? An Evaluation Framework and Analyses*. doi:10.48550/arXiv.2402.02008

Xiong, Miao, *et al*. 2023. *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs*. doi:10.48550/arXiv.2306.13063

Xu, Weijia, *et al*. 2023. *Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection*. doi:10.48550/arXiv.2301.07779

Xu, Ziwei, Sanjay Jain, & Mohan Kankanhalli. 2024. *Hallucination Is Inevitable: An Innate Limitation of Large Language Models*. doi:10.48550/arXiv.2401.11817

Yin, Zhangyue, *et al*. 2023. *Do Large Language Models Know What They Don't Know?* doi:10.48550/arXiv.2305.18153

Yu, Wenhao, *et al*. 2023. *Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models*. doi:10.48550/arXiv.2311.09210

Zhang, Hanning, *et al*. 2023. *R-Tuning: Teaching Large Language Models to Refuse Unknown Questions*. doi:10.48550/arXiv.2311.09677

Zhang, Yue, *et al*. 2023. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. doi:10.48550/arXiv.2309.01219.

Zhang, Yunfeng, Q. Vera Liao, & Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305. FAT* '20. NY, USA: Association for Computing Machinery.