

Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models

Journal of Legal Analysis (forthcoming)

Matthew Dahl¹, Varun Magesh², Mirac Suzgun², and Daniel E. Ho²

¹Yale Law School

²Stanford Law School

March 8, 2024

Abstract

Do large language models (LLMs) know the law? These models 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, documenting LLMs’ varying performance across jurisdictions, courts, time periods, and cases. Our work makes four key contributions. First, we develop a typology of legal hallucinations, providing a conceptual framework for future research in this area. Second, we find that legal hallucinations are alarmingly prevalent, occurring between 58% of the time with ChatGPT 4 and 88% with Llama 2, when these models are asked specific, verifiable questions about random federal court cases. Third, we illustrate that LLMs often fail to correct a user’s incorrect legal assumptions in a contra-factual question setup. Fourth, we provide evidence that LLMs cannot always predict, or do not always know, when they are producing legal hallucinations. Taken together, our findings caution against the rapid and unsupervised integration of popular LLMs into legal tasks. Even experienced lawyers must remain wary of legal hallucinations, and the risks are highest for those who stand to benefit from LLMs the most—*pro se* litigants or those without access to traditional legal resources.¹

Word Count: 10,769 (without references)

Keywords: Law and technology · AI · Access to justice · Algorithmic harm

¹ All our code, raw data, prompts, and results will be made available upon publication.

1 Introduction

Do large language models (LLMs) know the law? These tools—enormous models trained to produce textual outputs responsive to textual inputs—are driving a transformation of 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 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 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 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 and Ho 2020; Engstrom et al. 2020; Solow-Niederman 2020), given that they are already being used by public (Engel and 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 (Lemley and Casey 2019; Henderson, Hashimoto, and Lemley 2023; 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 (Simshaw 2022; Draper and Gillibrand 2023). 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, producing a kind of algorithmic “monoculture” by entrenching one particular notion of the law across a wide range of applications (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 U.S. District Courts (USDC), the U.S. Courts of Appeals (USCOA), and the U.S. 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: (1) 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 (2) their certainty in their responses, i.e., their self-awareness of their propensity to hallucinate (Kadavath et al. 2022; Xiong et al. 2023; Tian, Mitchell, Zhou, et al. 2023; Yin et al. 2023; Azaria and Mitchell 2023). Our results indicate that not only do LLMs often provide seemingly legitimate but incorrect answers to contra-factual 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 are 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. The terms in color preview that connection. 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**. Here, τ is a temperature parameter that controls the shape of the probability distribution at inference time and is configurable by the user. When $\tau = 0$, the distribution becomes degenerate and the model’s response is theoretically deterministic—the model must always return the most likely next token.² This deterministic response is referred to as the “greedy” response. When τ is high, however, the distribution becomes more uniform and the model’s response becomes more stochastic—this is the default behavior of chatbots like ChatGPT, which are often used to generate diverse and creative responses.

2.2 The Nature of Legal Hallucinations

LLMs are showing promise on a number of legal research and analysis tasks (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

²In practice, non-determinism may persist due to a model’s implementation details, e.g., the “mixture of experts” architecture (Puigcerver et al. 2023; Chann 2023).

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 chatbot.

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). 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 in order 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 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, LLMs might also hallucinate by producing a **response** that either contradicts or does

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

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 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. Defining the contours of an unwanted “hallucination” in this context 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).

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).

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 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. We supply some of the empirics in this article (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.

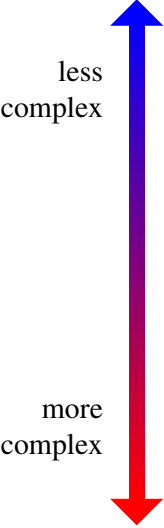
3 Profiling Hallucinations Using Legal Knowledge Queries

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 Figure 1.

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:

Existence: *Given the name and citation of a case, ascertain whether the case actually exists.* 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



Task	Query	Method
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
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
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

Figure 1: Hallucination QA task list. Tasks are sorted in order of increasing complexity. Query wording is paraphrased; see Appendix B 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).

insights. We use only real cases in our prompts, so affirming their existence is the correct answer.⁵

Court: *Given the name and citation of a case, provide 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 *U.S. 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 U.S. 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

⁵ In Appendix E.1, we experiment with using fake cases as well.

district court.

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 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).

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:

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. This is essentially a binary classification task where we accept correct “affirm” or “reverse” labels as accurate. 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

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.

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).

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

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

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.

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:

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.

Factual background: *Given a case name and its citation, state 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.

Procedural posture: *Given a case name and its citation, state 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.

Subsequent history: *Given a case name and its citation, state 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.

Core legal question: *Given a case name and its citation, state 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.

Central holding: *Given a case name and its citation, state 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, 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,](#)

⁸More information about how we use these metadata to construct each query is available in [Appendix A](#).

and Evans 2022; Lee et al. 2023; Li et al. 2023). Such oracles are usually difficult and costly to construct (Krishna, Roy, and Iyyer 2021), but we exploit the tabular metadata described in Section 4.1 to develop ours. Our assumption is that while LLMs have access to and are trained on the raw text of American case law, which is in the public domain (Henderson et al. 2022), they have not yet explicitly memorized 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 (Figure 1). These queries take the form of N question-and-answer triples (**prompt**, **response**, **response'**), where the **prompt** is a case-specific question, the **response** is the LLM’s greedy answer retrieved from calling $f_0(\cdot)$, 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:

$$\pi = \hat{\pi} = \frac{1}{N} \sum \mathbb{1}[\text{response} \neq \text{response}'] \quad (1)$$

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 Appendix F, but they are generally rare and do not substantively affect our findings.

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.

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 inquiries. 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 (Shuster et al. 2021; Peng et al. 2023).

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 (Figure 1), we construct reference-free queries in the form of N question-and-answer triples (**prompt**, **response**⁽¹⁾, **response**⁽²⁾), where the **prompt** is the question, **response**⁽¹⁾ is one LLM answer retrieved by calling $f_1(\cdot)$ once, and **response**⁽²⁾ is another LLM answer retrieved by calling $f_1(\cdot)$ 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 contra-

¹⁰ Recall from Section 2.1 that the f_1 notation represents performing inference with the LLM at temperature one—i.e., with some degree of stochasticity.

dictory 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 Appendix C. (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:

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

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 LLMs: OpenAI's ChatGPT 4 (gpt-4-1106-preview) (OpenAI 2023a), OpenAI's ChatGPT 3.5 (gpt-3.5-turbo-0613) (2023b), Google's PaLM 2 (text-bison-001) (Anil et al. 2023), and 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 Appendix B. All of our raw queries and responses, including timestamp logs, are available 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: (1) their susceptibility to contra-factual bias, i.e., their ability to handle queries based on mistaken legal premises (Section 5.2), and (2) 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 2, 3, and 4 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 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 between 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 2. Starting with the low complexity category (Table 2), 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*

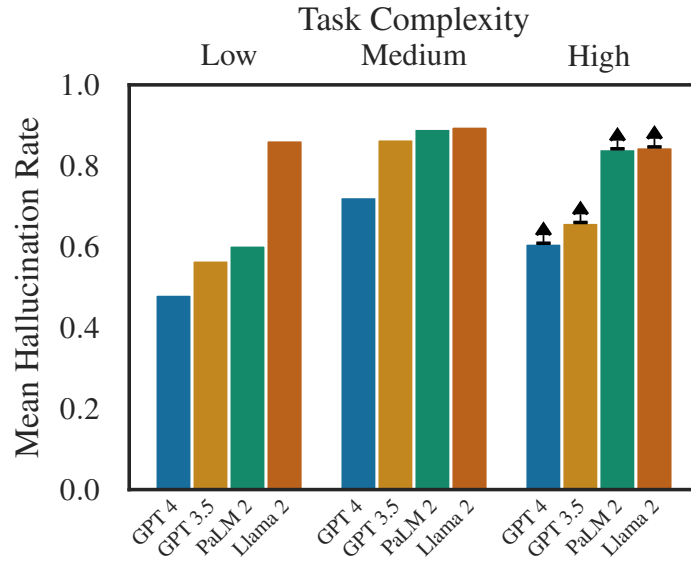


Figure 2: Relationship between task complexity and mean hallucination rate. Higher values indicate a greater likelihood of factually incorrect LLM responses. High complexity tasks include several reference-free tasks, so those reported hallucination rates are lower bounds on the true rates. Contra-factual tasks and the doctrinal agreement high complexity task are excluded from this comparison.

case. (In Appendix E.1 we demonstrate this problem by asking about the existence of fake cases instead.) The models begin to struggle more when prompted for information about a case’s **Court**, **Citation**, or **Author**. Hallucinations then surge among the moderate complexity tasks (Table 3), 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 4) 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. 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 further 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 than 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.

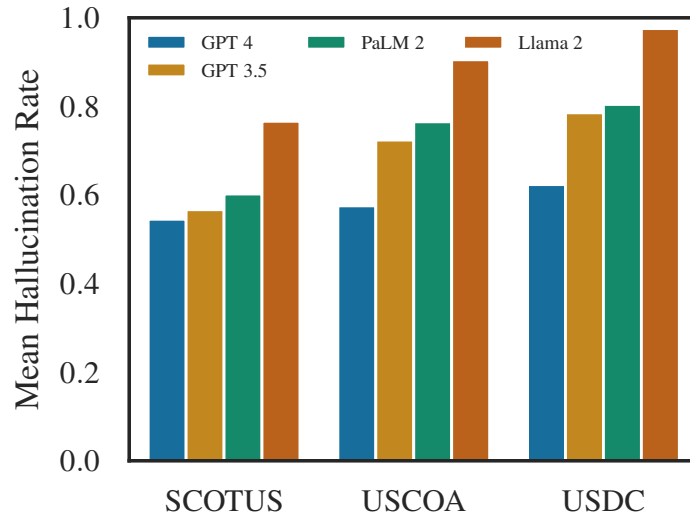


Figure 3: Relationship between judicial hierarchy and mean hallucination rate, all reference-based tasks pooled. Hallucination rates are higher for lower levels of the federal judiciary.

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 3). 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

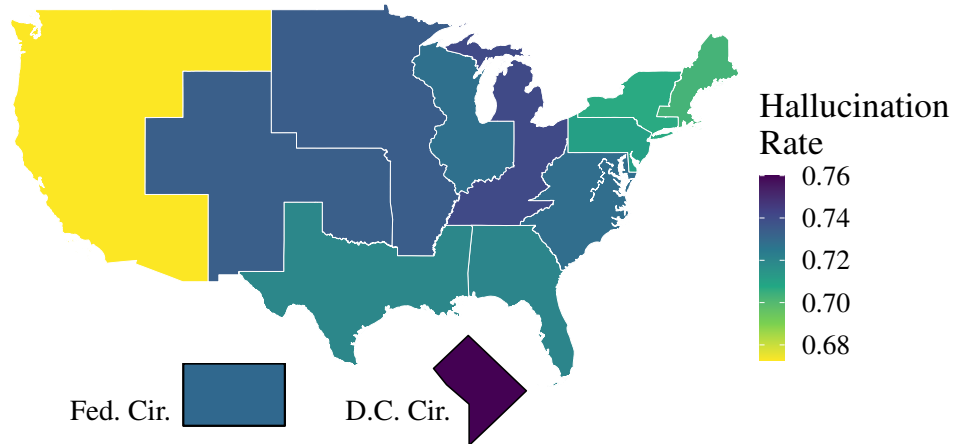


Figure 4: Relationship between USCOA jurisdiction and mean hallucination rate, all reference-based USCOA tasks and models pooled, post-1981 cases only. LLM performance is strongest in jurisdictions that are commonly perceived to play a more influential role.

heterogeneity across the circuits.¹¹ Figure 4 depicts these results geographically, showing lower hallucination 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

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 Appendix Section E.2, which are largely consistent with these post-1981 results.

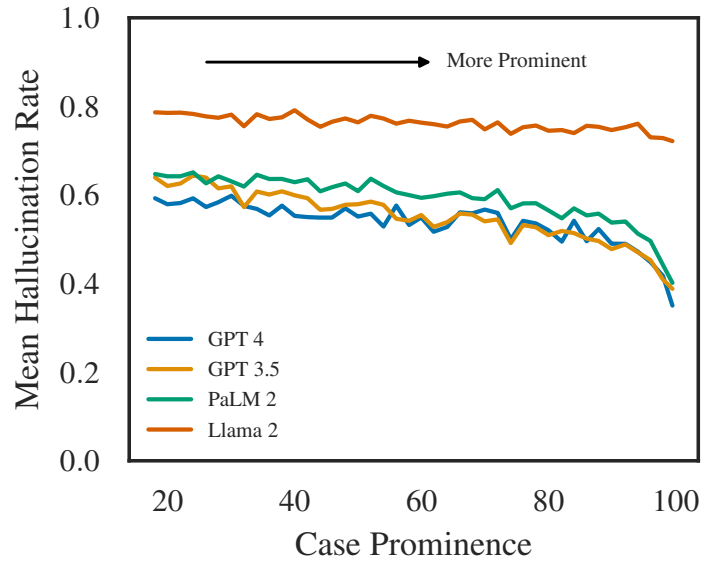


Figure 5: Relationship between SCOTUS case prominence (measured by PageRank percentile) and mean hallucination rate, all SCOTUS tasks pooled. Hallucinations decline sharply as case prominence passes the 90th percentile, meaning that LLMs are more likely to respond with accurate information about prominent cases.

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.

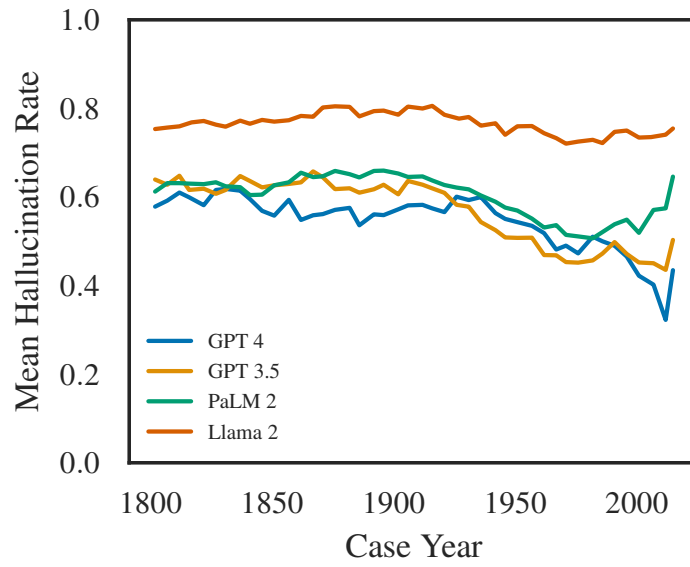


Figure 6: Relationship between SCOTUS case decision year and mean hallucination rate, all SCOTUS tasks pooled. LLMs are most likely to respond with accurate information in cases from the latter half of the 20th century, struggling on very old and very new cases.

We find that case prominence is negatively correlated with hallucination, reaffirming our results from above (Figure 5). 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 6, 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

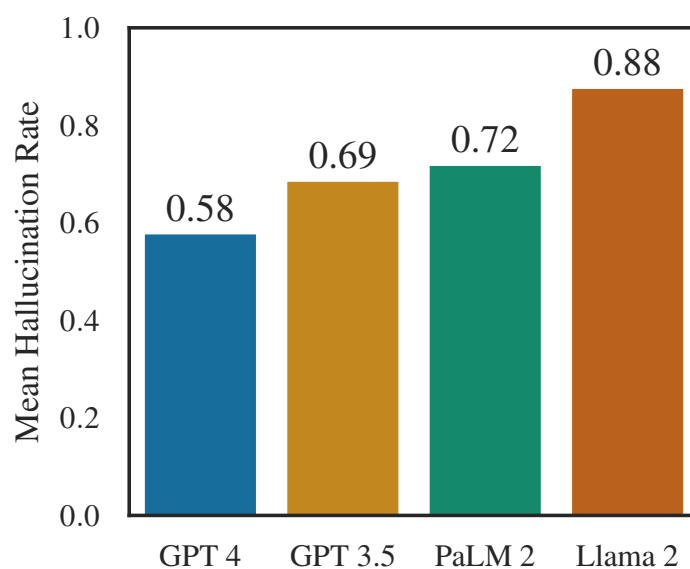


Figure 7: Hallucination rates by LLM, all reference-based tasks pooled. Hallucinations are common across all LLMs when they are asked a direct, verifiable question about a federal court case, but GPT 4 performs best overall.

law that is very old but still applicable and relevant law.

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 7).

We also discover tendencies towards different inductive biases, or the predisposition of an LLM to generate certain outputs more frequently than others. In Figure 8, 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; 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

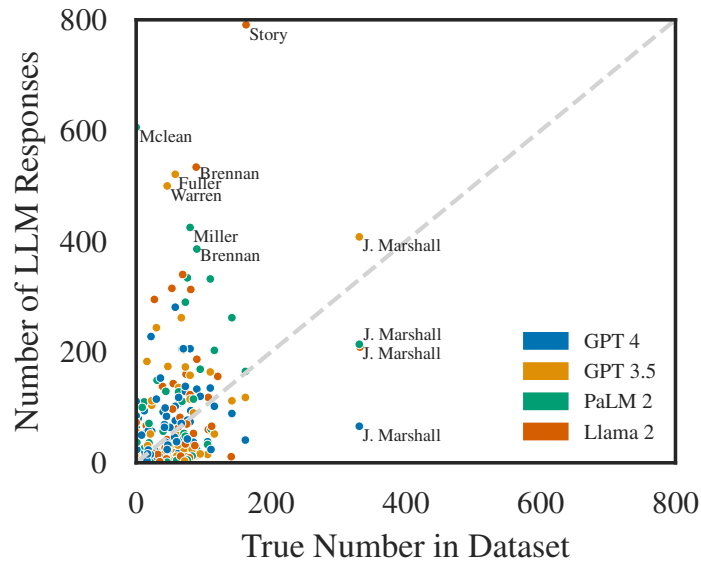


Figure 8: Number of times each justice is stated to be the author of a SCOTUS case versus the actual number of cases authored by each justice in our time period-stratified dataset. A small number of justices are disproportionately represented in LLM responses.

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 8.

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 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 and 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 and Steinhardt 2022](#); [Suri et al. 2023](#)).

This bias poses a subtle but pernicious challenge to those aiming to use LLMs for legal research. 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 recapitulating the user’s misconception and producing a non-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 (1) provide information about an author’s dissenting opinion in an appellate case in which they did not in fact dissent and (2) 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 5](#) reports the results of this experiment and [Figure 9](#) 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.

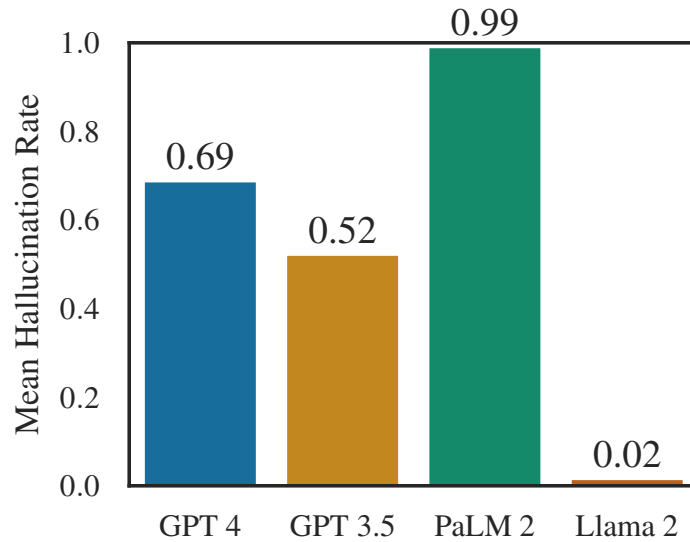


Figure 9: Hallucination rates by LLM, all contra-factual tasks pooled. Llama 2 is very unlikely to hallucinate on these tasks because it almost always rejects the premise in the question. However, this tendency also leads it to perform more poorly on tasks with correct premises (cf. Figure 7).

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 (Kadavath et al. 2022; Xiong et al. 2023; Tian,

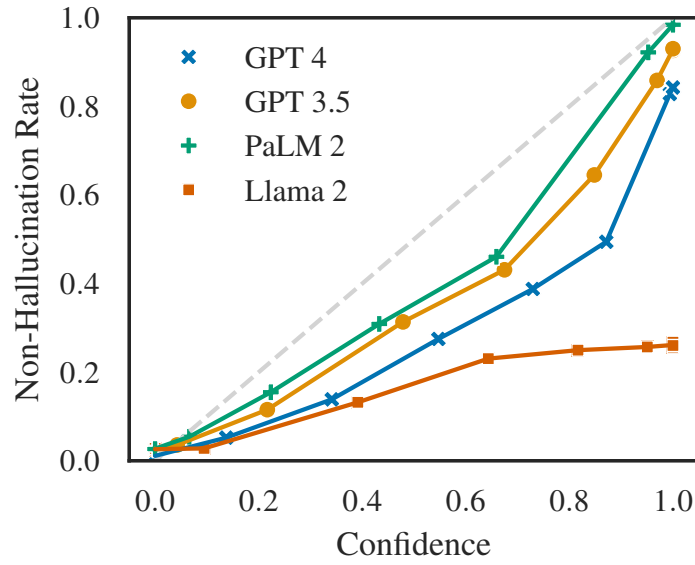


Figure 10: Calibration curves by LLM, all reference-based tasks pooled. PaLM 2 is best calibrated on legal queries, followed by GPT 3.5, GPT 4, and lastly Llama 2, which is significantly worse than the first three models.

Mitchell, Zhou, et al. 2023; Yin et al. 2023; Azaria and Mitchell 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. Appendix D describes our estimation strategy in full, but, intuitively, 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 6 reports the results of this analysis at the task level, and Figure 10 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 frequen-

cies 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 7), 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 6), 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 jus-

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

tice? Although there is much enthusiasm for LLMs’ potential to revolutionize these domains, we highlight the problem of legal hallucinations, which remains a serious obstacle to the adoption of these models. Performing the first systematic empirical test of popular perceptions (Weiser 2023; Romoser 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 research: (1) their susceptibility to contra-factual bias, i.e., their inability to handle queries containing an erroneous or mistaken starting point (Section 5.2), and (2) 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 (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 the LLMs’ ability to engage in legal *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 progenitors of legal counsel and advice, no matter how strong their in-context reasoning abilities become.

This means that our results must temper optimism for the ability of off-the-shelf, publicly available LLMs to accelerate access to justice (Tito 2017; Perlman 2023; Tan, Westermann, and Benyekhlef 2023). Indeed, our findings suggest that the risks of using these generic foundation

models are especially high for litigants who are:

- Filing in courts lower in the judicial hierarchy or those located in less prominent jurisdictions
- Seeking more complex forms of legal information
- Formulating questions with mistaken premises
- 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*—indigent or *pro se* litigants. Not only do LLMs hallucinate widely, their current implementations lack the behavioral features that such users would require. Ideally, LLMs would 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). We therefore echo concerns that the proliferation of LLMs may ultimately exacerbate, rather than eradicate, existing inequalities in access to legal services (Simshaw 2022; Draper and Gillibrand 2023). At the same time, increased reliance on LLMs may produce a kind of legal monoculture (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.

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 hallucination problem 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 (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.

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 or student aids (Livermore, Herron, and Rockmore 2024; 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).

Table 2: Hallucination rates across levels of the federal judiciary (low complexity tasks)

Task	Prompt	SCOTUS (1794-2015; n=5000)				USCOA (1895-2019; n=5000)				USDC (1932-2019; n=5000)			
		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)	0.004 (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)	0.240 (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)	0.870 (0.005)	0.842 (0.005)	0.870 (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)	0.610 (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)	0.920 (0.004)	0.999 (0.001)	1.000 (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)	0.830 (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. Abstention responses are dropped. Standard errors are shown in parentheses.

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

Task	Prompt	SCOTUS (1794-2015; n=5000)				USCOA (1895-2019; n=5000)				USDC (1932-2019; n=5000)			
		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 (0.007)	0.499 (0.001)	0.500 (0.000)	0.536 (0.007)	0.494 (0.005)	0.500 (0.000)	0.500 (0.000)	0.493 (0.008)	-	-	-	-
	Few-shot	0.452 (0.007)	0.496 (0.002)	0.501 (0.002)	0.502 (0.004)	0.498 (0.006)	0.489 (0.007)	0.501 (0.001)	0.501 (0.002)	-	-	-	-
Quotation	Zero-shot	0.312 (0.007)	0.229 (0.006)	0.175 (0.005)	0.920 (0.004)	0.001 (0.000)	0.000 (0.000)	0.532 (0.007)	0.993 (0.001)	0.001 (0.000)	0.000 (0.000)	0.989 (0.001)	0.964 (0.003)
	Few-shot	0.854 (0.005)	1.000 (0.000)	0.993 (0.001)	0.992 (0.001)	0.637 (0.007)	1.000 (0.000)	0.997 (0.001)	1.000 (0.000)	0.743 (0.006)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Authority	Zero-shot	0.922 (0.004)	0.937 (0.003)	0.985 (0.002)	0.991 (0.001)	0.934 (0.004)	0.978 (0.002)	0.997 (0.001)	0.999 (0.001)	0.956 (0.003)	0.870 (0.005)	0.995 (0.001)	0.999 (0.000)
	Few-shot	0.828 (0.005)	0.916 (0.004)	0.953 (0.003)	0.993 (0.001)	0.958 (0.003)	0.976 (0.002)	0.993 (0.001)	0.999 (0.000)	0.928 (0.004)	0.956 (0.003)	0.994 (0.001)	0.997 (0.001)
Overruling year ^a	Zero-shot	0.810 (0.025)	0.919 (0.017)	0.858 (0.022)	0.972 (0.011)	-	-	-	-	-	-	-	-
	Few-shot	0.725 (0.028)	0.976 (0.010)	0.870 (0.021)	0.984 (0.008)	-	-	-	-	-	-	-	-

^a 1810-2022 (n=279)

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

Table 4: 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 ^a	Zero-shot	0.461 (0.005)	0.500 (0.000)	0.466 (0.005)	0.500 (0.000)	-	-	-	-	-	-	-	-
	Few-shot	0.449 (0.007)	0.458 (0.004)	0.453 (0.006)	0.500 (0.000)	-	-	-	-	-	-	-	-
Factual background	Zero-shot	0.460 (0.050)	0.700 (0.046)	0.950 (0.022)	0.830 (0.038)	0.750 (0.043)	0.810 (0.039)	0.880 (0.032)	0.880 (0.032)	0.710 (0.045)	0.670 (0.047)	0.880 (0.032)	0.880 (0.032)
Procedural posture	Zero-shot	0.470 (0.050)	0.650 (0.048)	0.740 (0.044)	0.850 (0.036)	0.660 (0.047)	0.600 (0.049)	0.680 (0.047)	0.830 (0.038)	0.730 (0.044)	0.550 (0.050)	0.810 (0.039)	0.890 (0.031)
Subsequent history	Zero-shot	0.080 (0.027)	0.220 (0.041)	0.810 (0.039)	0.770 (0.042)	0.450 (0.050)	0.360 (0.048)	0.750 (0.043)	0.620 (0.049)	0.480 (0.050)	0.300 (0.046)	0.730 (0.044)	0.670 (0.047)
Core legal question	Zero-shot	0.570 (0.050)	0.860 (0.035)	0.870 (0.034)	0.920 (0.027)	0.810 (0.039)	0.880 (0.032)	0.950 (0.022)	0.960 (0.020)	0.760 (0.043)	0.720 (0.045)	0.920 (0.027)	0.890 (0.031)
Central holding	Zero-shot	0.600 (0.049)	0.730 (0.044)	0.780 (0.041)	0.920 (0.027)	0.830 (0.038)	0.870 (0.034)	0.950 (0.022)	0.950 (0.022)	0.770 (0.042)	0.730 (0.044)	0.930 (0.026)	0.840 (0.037)

^a 1796-2005 (n=5000)

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.

Table 5: Hallucination rates across levels of the federal judiciary (contra-factual tasks)

Task	Prompt	SCOTUS (1794-2015; n=1000)				USCOA (1895-2019; n=1000)			
		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 (0.015)	0.338 (0.015)	0.990 (0.003)	0.000 (0.000)	0.842 (0.012)	0.408 (0.016)	0.983 (0.004)	0.021 (0.005)
False overruling premise	Zero-shot	0.531 (0.016)	0.821 (0.012)	1.000 (0.000)	0.027 (0.005)	-	-	-	-

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

Table 6: 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	0.126 (0.005)	0.008 (0.001)	0.119 (0.004)	0.262 (0.006)	0.124 (0.005)	0.007 (0.001)	0.182 (0.003)	0.117 (0.005)	0.074 (0.004)	0.004 (0.000)	0.183 (0.003)	0.178 (0.006)
	Few-shot	0.149 (0.005)	0.038 (0.002)	0.041 (0.002)	0.998 (0.000)	0.073 (0.005)	0.064 (0.002)	0.014 (0.001)	0.998 (0.000)	0.024 (0.003)	0.053 (0.002)	0.038 (0.001)	0.996 (0.000)
Court	Zero-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.003 (0.000)	0.362 (0.006)	0.322 (0.006)	0.184 (0.006)	0.387 (0.006)	0.133 (0.005)	0.163 (0.005)	0.132 (0.005)	0.163 (0.006)
	Few-shot	0.000 (0.000)	0.001 (0.000)	0.000 (0.000)	0.009 (0.001)	0.345 (0.006)	0.355 (0.006)	0.161 (0.006)	0.452 (0.008)	0.145 (0.005)	0.099 (0.004)	0.094 (0.005)	0.099 (0.005)
Citation	Zero-shot	0.113 (0.005)	0.069 (0.004)	0.026 (0.003)	0.068 (0.003)	0.143 (0.004)	0.073 (0.005)	0.004 (0.001)	0.036 (0.001)	0.104 (0.003)	0.043 (0.003)	0.002 (0.000)	0.022 (0.001)
	Few-shot	0.191 (0.005)	0.029 (0.003)	0.059 (0.004)	0.063 (0.003)	0.138 (0.005)	0.036 (0.003)	0.010 (0.001)	0.032 (0.001)	0.099 (0.004)	0.018 (0.002)	0.001 (0.000)	0.051 (0.002)
Author	Zero-shot	0.443 (0.006)	0.208 (0.006)	0.130 (0.005)	0.419 (0.006)	0.349 (0.005)	0.094 (0.003)	0.131 (0.002)	0.583 (0.004)	0.302 (0.005)	0.228 (0.005)	0.063 (0.002)	0.305 (0.004)
	Few-shot	0.454 (0.006)	0.347 (0.006)	0.142 (0.005)	0.454 (0.005)	0.357 (0.004)	0.142 (0.003)	0.118 (0.003)	0.656 (0.004)	0.320 (0.004)	0.096 (0.003)	0.045 (0.002)	0.481 (0.005)
Disposition	Zero-shot	0.203 (0.007)	0.431 (0.007)	0.291 (0.008)	0.199 (0.007)	0.170 (0.007)	0.557 (0.006)	0.382 (0.006)	0.148 (0.006)	-	-	-	-
	Few-shot	0.239 (0.006)	0.314 (0.008)	0.165 (0.007)	0.283 (0.008)	0.165 (0.007)	0.205 (0.007)	0.227 (0.007)	0.439 (0.007)	-	-	-	-
Overruling year ^a	Zero-shot	0.308 (0.025)	0.246 (0.022)	0.116 (0.019)	0.510 (0.019)	-	-	-	-	-	-	-	-
	Few-shot	0.377 (0.025)	0.680 (0.022)	0.154 (0.019)	0.754 (0.018)	-	-	-	-	-	-	-	-
Doctrinal agreement ^b	Zero-shot	0.369 (0.006)	0.527 (0.007)	0.165 (0.006)	0.564 (0.006)	-	-	-	-	-	-	-	-
	Few-shot	0.319 (0.007)	0.409 (0.008)	0.152 (0.007)	0.548 (0.007)	-	-	-	-	-	-	-	-

^a 1810-2022 (n=279) ^b 1796-2005 (n=5000)

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.

References

- Agrawal, Ayush, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. *Do Language Models Know When They're Hallucinating References?*, arXiv:2305.18248, September. <https://doi.org/10.48550/arXiv.2305.18248>.
- Ambrogi, Bob. 2023. "As Allen & Overy Deploys GPT-based Legal App Harvey Firmwide, Founders Say Other Firms Will Soon Follow." *LawSites* (February).
- Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, et al. 2023. *PaLM 2 Technical Report*, arXiv:2305.10403, May. <https://doi.org/10.48550/arXiv.2305.10403>.
- Azaria, Amos, and Tom Mitchell. 2023. *The Internal State of an LLM Knows When It's Lying*, arXiv:2304.13734, October. <https://doi.org/10.48550/arXiv.2304.13734>.
- Bar-Gill, Oren, Cass R Sunstein, and Inbal Talgam-Cohen. 2023. "Algorithmic Harm in Consumer Markets." *Journal of Legal Analysis* 15, no. 1 (September): 1–47. <https://doi.org/10.1093/jla/laad003>.
- Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732.
- Ben-Shahar, Omri. 2023. "Privacy Protection, At What Cost? Exploring the Regulatory Resistance to Data Technology in Auto Insurance." *Journal of Legal Analysis* 15, no. 1 (August): 129–157. <https://doi.org/10.1093/jla/laad008>. eprint: <https://academic.oup.com/jla/article-pdf/15/1/129/51539112/laad008.pdf>. <https://doi.org/10.1093/jla/laad008>.
- Black, Ryan C., and James F. Spriggs II. 2013. "The Citation and Depreciation of U.S. Supreme Court Precedent." *Journal of Empirical Legal Studies* 10 (2): 325–358.

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.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. *Language Models Are Few-Shot Learners*, arXiv:2005.14165, July. <https://doi.org/10.48550/arXiv.2005.14165>.

Cao, Meng, Yue Dong, and Jackie Cheung. 2022. “Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization.” In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, 3340–3354. Dublin, Ireland: Association for Computational Linguistics, May. <https://doi.org/10.18653/v1/2022.acl-long.236>.

Cao, Ziqiang, Furu Wei, Wenjie Li, and Sujian Li. 2018. “Faithful to the Original: Fact Aware Neural Abstractive Summarization.” *Proceedings of the AAAI Conference on Artificial Intelligence* 32, no. 1 (April). <https://doi.org/10.1609/aaai.v32i1.11912>.

Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. *Quantifying Memorization Across Neural Language Models*, arXiv:2202.07646, February.

Caselaw Access Project. 2023. *Caselaw Access Project*.

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

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

- Choi, Jonathan H., Kristin E. Hickman, Amy Monahan, and Daniel Schwarcz. 2022. “ChatGPT Goes to Law School.” *Journal of Legal Education* 71 (3): 387–400. <https://doi.org/10.2139/ssrn.4335905>.
- Choi, Jonathan H., and Daniel Schwarcz. 2024. “AI Assistance in Legal Analysis: An Empirical Study.” *Journal of Legal Education* 73:forthcoming. <https://doi.org/10.2139/ssrn.4539836>.
- Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20, no. 1 (April): 37–46. <https://doi.org/10.1177/001316446002000104>.
- Congress.gov. 2023. *Table of Supreme Court Decisions Overruled by Subsequent Decisions*. [https://constitution.congress.gov/overruled/](https://constitution.congress.gov/constitution/overruled/).
- Cui, Jiayi, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. *ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases*, arXiv:2306.16092, June. <https://doi.org/10.48550/arXiv.2306.16092>.
- Cushman, Jack, Matthew Dahl, and Michael Lissner. 2021. “Eyecite: A Tool for Parsing Legal Citations.” *Journal of Open Source Software* 6 (66): 3617.
- Dell’Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraymer, François Candelon, and Karim R. Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. <https://doi.org/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?*, arXiv:2306.01248, June. <https://doi.org/10.48550/arXiv.2306.01248>.
- 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, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. *GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, arXiv:2303.10130, August. <https://doi.org/10.48550/arXiv.2303.10130>.
- Engel, Christoph, and Nina Grgić-Hlača. 2021. “Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court.” *Journal of Legal Analysis* 13, no. 1 (August): 284–340. <https://doi.org/10.1093/jla/laab001>.
- Engstrom, David Freeman, and Daniel E. Ho. 2020. “Algorithmic Accountability in the Administrative State.” *Yale Journal on Regulation* 37 (3): 800–854.
- Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Technical report. Administrative Conference of the United States.
- Fei, Zhiwei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. *LawBench: Benchmarking Legal Knowledge of Large Language Models*, arXiv:2309.16289, September.
- Feijo, Diego de Vargas, and Viviane P. Moreira. 2023. “Improving Abstractive Summarization of Legal Rulings through Textual Entailment.” *Artificial Intelligence and Law* 31, no. 1 (March): 91–113. <https://doi.org/10.1007/s10506-021-09305-4>.
- Feng, Shangbin, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. *Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration*. <https://doi.org/10.48550/arXiv.2402.00367>.
- Fowler, James H., Timothy R. Johnson, James F. Spriggs II, Sangick Jeon, and Paul J. Wahlbeck. 2007. “Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court.” *Political Analysis* 15 (3): 324–346.

- Gillis, Talia B., and Jann L. Spiess. 2019. “Big Data and Discrimination.” *The University of Chicago Law Review* 86 (2): 459–488.
- Guha, Neel, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, et al. 2023. *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*, arXiv:2308.11462, August. <https://doi.org/10.48550/arXiv.2308.11462>.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. “On Calibration of Modern Neural Networks.” In *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330. PMLR, July.
- Henderson, Peter, Tatsunori Hashimoto, and Mark Lemley. 2023. “Where’s the Liability in Harmful AI Speech?” *Journal of Free Speech Law* 3 (2): 589–650.
- Henderson, Peter, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. *Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset*, arXiv:2207.00220, November. <https://doi.org/10.48550/arXiv.2207.00220>.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55, no. 12 (March): 248:1–248:38. <https://doi.org/10.1145/3571730>.
- Jones, Erik, and Jacob Steinhardt. 2022. *Capturing Failures of Large Language Models via Human Cognitive Biases*, arXiv:2202.12299, November. <https://doi.org/10.48550/arXiv.2202.12299>.
- Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, et al. 2022. *Language Models (Mostly) Know What They Know*, arXiv:2207.05221, November. <https://doi.org/10.48550/arXiv.2207.05221>.

- Kalai, Adam Tauman, and Santosh S. Vempala. 2023. *Calibrated Language Models Must Hallucinate*, arXiv:2311.14648, December. <https://doi.org/10.48550/arXiv.2311.14648>.
- Katz, Daniel Martin, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. *GPT-4 Passes the Bar Exam*. <https://doi.org/10.2139/ssrn.4389233>.
- King, Jennifer, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown. 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*, 492–505. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery. <https://doi.org/10.1145/3593013.3594015>. <https://doi.org/10.1145/3593013.3594015>.
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. “Discrimination in the Age of Algorithms.” *Journal of Legal Analysis* 10 (December): 113–174. <https://doi.org/10.1093/jla/laz001>.
- Kleinberg, Jon, and Manish Raghavan. 2021. “Algorithmic Monoculture and Social Welfare.” *Proceedings of the National Academy of Sciences* 118, no. 22 (June). <https://doi.org/10.1073/pnas.2018340118>.
- Krishna, Kalpesh, Aurko Roy, and 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*, 4940–4957. Online: Association for Computational Linguistics, June. <https://doi.org/10.18653/v1/2021.naacl-main.393>.
- Kuersten, Ashlyn K., and Susan B. Haire. 2011. *Update to the Appeals Courts Database (1997–2002)*.
- Kumar, Ananya, Percy Liang, and Tengyu Ma. 2020. *Verified Uncertainty Calibration*, arXiv:1909.10155, January. <https://doi.org/10.48550/arXiv.1909.10155>.
- Landis, J. Richard, and Gary G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1): 159–174. <https://doi.org/10.2307/2529310>.

- Lee, Nayeon, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. *Factuality Enhanced Language Models for Open-Ended Text Generation*, arXiv:2206.04624, March. <https://doi.org/10.48550/arXiv.2206.04624>.
- Lemley, Mark A., and Bryan Casey. 2019. “Remedies for Robots.” *The University of Chicago Law Review* 86 (5): 1311–1396.
- Li, Junyi, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*, arXiv:2305.11747, May. <https://doi.org/10.48550/arXiv.2305.11747>.
- Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, arXiv:2109.07958, May. <https://doi.org/10.48550/arXiv.2109.07958>.
- Livermore, Michael A., Felix Herron, and Daniel Rockmore. 2024. “Language Model Interpretability and Empirical Legal Studies.” *Journal of Institutional and Theoretical Economics*, forthcoming.
- 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.
- Manakul, Potsawee, Adian Liusie, and Mark J. F. Gales. 2023. *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*, arXiv:2303.08896, May. <https://doi.org/10.48550/arXiv.2303.08896>.
- Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. “On Faithfulness and Factuality in Abstractive Summarization.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Online: Association for Computational Linguistics, July. <https://doi.org/10.18653/v1/2020.acl-main.173>.
- Mayson, Sandra G. 2019. “Bias In, Bias Out.” *The Yale Law Journal* 128 (8): 2122–2473.

- Min, Sewon, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*, arXiv:2305.14251, May. <https://doi.org/10.48550/arXiv.2305.14251>.
- Mündler, Niels, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. *Self-Contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation*, arXiv:2305.15852, May. <https://doi.org/10.48550/arXiv.2305.15852>.
- Nay, John J., David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. *Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence*, arXiv:2306.07075, June. <https://doi.org/10.48550/arXiv.2306.07075>.
- OpenAI. 2023a. *GPT-4 Technical Report*, arXiv:2303.08774, March. <https://doi.org/10.48550/arXiv.2303.08774>.
- . 2023b. *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>, November.
- Peng, Baolin, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, et al. 2023. *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback*, arXiv:2302.12813, March. <https://doi.org/10.48550/arXiv.2302.12813>.
- Perlman, Andrew. 2023. “The Implications of ChatGPT for Legal Services and Society.” *The Practice*, no. March/April,
- Puigcerver, Joan, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. 2023. *From Sparse to Soft Mixtures of Experts*, arXiv:2308.00951, August. <https://doi.org/10.48550/arXiv.2308.00951>.
- Roberts, John G. 2023. *2023 Year-End Report on the Federal Judiciary*. Technical report. December.

- Rodgers, Ian, John Armour, and Mari Sako. 2023. “How Technology Is (or Is Not) Transforming Law Firms.” *Annual Review of Law and Social Science* 19 (1): 299–317. <https://doi.org/10.1146/annurev-lawsocsci-111522-074716>.
- Romoser, James. 2023. *No, Ruth Bader Ginsburg Did Not Dissent in Obergefell — and Other Things ChatGPT Gets Wrong about the Supreme Court*, January.
- Savelka, Jaromir, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. *Explaining Legal Concepts with Augmented Large Language Models (GPT-4)*, arXiv:2306.09525, June. <https://doi.org/10.48550/arXiv.2306.09525>.
- Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Newton Cheng, et al. 2023. *Towards Understanding Sycophancy in Language Models*.
- Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. “Retrieval Augmentation Reduces Hallucination in Conversation.” In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. Punta Cana, Dominican Republic: Association for Computational Linguistics, November. <https://doi.org/10.18653/v1/2021.findings-emnlp.320>.
- Simshaw, Drew. 2022. “Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services.” *Yale Journal of Law & Technology* 24:150–226.
- Solow-Niederman, Alicia. 2020. “Administering Artificial Intelligence.” *Southern California Law Review* 93 (4): 633–696.
- Songer, Donald R. 2008. *The United States Courts of Appeals Database, 1925–1996*.
- Spaeth, Harold J., Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2022. *2022 Supreme Court Database, Version 2022 Release 01*. <http://supremecourtdatabase.org/>.

- Suri, Gaurav, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. 2023. *Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5*, arXiv:2305.04400, May. <https://doi.org/10.48550/arXiv.2305.04400>.
- Tan, Jinzhe, Hannes Westermann, and Karim Benyekhlef. 2023. “ChatGPT as an Artificial Lawyer?” In *Proceedings of the ICAIL 2023 Workshop on Artificial Intelligence for Access to Justice*. Braga, Portugal: CEUR Workshop Proceedings.
- Tian, Katherine, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. *Fine-Tuning Language Models for Factuality*, arXiv:2311.08401, November. <https://doi.org/10.48550/arXiv.2311.08401>.
- Tian, Katherine, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. *Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback*, arXiv:2305.14975, May. <https://doi.org/10.48550/arXiv.2305.14975>.
- Tito, Joel. 2017. *How AI Can Improve Access to Justice*, October.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*, arXiv:2307.09288, July. <https://doi.org/10.48550/arXiv.2307.09288>.
- Trozze, Arianna, Toby Davies, and Bennett Kleinberg. 2023. *Large Language Models in Cryptocurrency Securities Cases: Can ChatGPT Replace Lawyers?*, arXiv:2308.06032, August. <https://doi.org/10.48550/arXiv.2308.06032>.
- Tversky, Amos, and Daniel Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185, no. 4157 (September): 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>.
- Volokh, Eugene. 2023. “Large Libel Models? Liability for AI Output.” *Journal of Free Speech Law* 3 (2): 489–558.

- Wang, Sinong, Han Fang, Madian Khabisa, Hanzi Mao, and Hao Ma. 2021. *Entailment as Few-Shot Learner*, arXiv:2104.14690, April. <https://doi.org/10.48550/arXiv.2104.14690>.
- Wei, Jerry, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. *Simple synthetic data reduces sycophancy in large language models*.
- Weiser, Benjamin. 2023. "Here's What Happens When Your Lawyer Uses ChatGPT." *The New York Times* (May).
- Wittgenstein, Ludwig. 1998 [1921]. *Tractatus Logico-Philosophicus*. Translated by C. K. Ogden. "Translated by C. K. Ogden". Dover.
- Xiong, Miao, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs*, arXiv:2306.13063, June. <https://doi.org/10.48550/arXiv.2306.13063>.
- Xu, Weijia, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. *Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection*, arXiv:2301.07779, February. <https://doi.org/10.48550/arXiv.2301.07779>.
- Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. 2024. *Hallucination Is Inevitable: An Innate Limitation of Large Language Models*, arXiv:2401.11817, January. <https://doi.org/10.48550/arXiv.2401.11817>.
- Yin, Zhangyue, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. *Do Large Language Models Know What They Don't Know?*, arXiv:2305.18153, May. <https://doi.org/10.48550/arXiv.2305.18153>.
- Zhang, Yunfeng, Q. Vera Liao, and 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. New York, NY, USA: Association for Computing Machinery, January. <https://doi.org/10.1145/3351095.3372852>.

Appendix

A Data Sources

In Section 4.1, we describe the data and sampling strategy that we use to construct our queries. For clarity, Appendix Table 7 links these data to each task. Our exact procedure for sampling, merging, and aggregating across these datasets is available in our GitHub repository.

Table 7: Data used to construct each task

Complexity	Task	Data source
Low	Existence	CAP (2023); SCDB (Spaeth et al. 2022)
	Court	CAP; SCDB
	Citation	CAP; SCDB
	Author	CAP; SCDB
Moderate	Disposition	CAP; SCDB; ACDB (Songer 2008; Kuersten and Haire 2011)
	Quotation	CAP; SCDB
	Authority	CAP; SCDB
	Overruling year	SCDB; Library of Congress (2023)
High	Doctrinal agreement	Shepard’s (Fowler et al. 2007; Black and Spriggs 2013)
	Factual background	CAP; SCDB
	Procedural posture	CAP; SCDB
	Subsequent history	CAP; SCDB
	Core legal question	CAP; SCDB
	Central holding	CAP; SCDB

B Prompt Templates

The full zero-shot and few-shot prompt templates for all of our queries are shown in Appendix Figures 14 to 28. The few-shot examples presented are those used for the SCOTUS queries; appropriate cases from the other levels of the judiciary are used in the USCOA and USDC versions.

Table 8: Intercoder reliability

Coders	κ
Coder 1 vs. GPT 4	0.87
Coder 2 vs. GPT 4	0.77
Coder 1 vs. Coder 2	0.79

C Contradiction Detection Approach

In Section 4.3, we describe our strategy for assessing hallucinations for our reference-free tasks. Briefly, because we do not have access to ground-truth labels for these tasks, we exploit the stochastic behavior of LLMs at higher temperatures and check for contradictions in their responses to repeated queries. In Appendix Figure 29, we share the template for the contradiction elicitation query that we send to GPT 4 to perform this contradiction labeling at scale. We frame our contradiction check as an entailment task (Wang et al. 2021) as we find that this prompting strategy is the most performant in our setting.

To confirm that GPT 4 is able to reliably detect the contradictions that we envision, we cross-check GPT 4’s conclusions against our own expert knowledge on a subsample of our data. Specifically, for 100 of our queries, two of us independently recoded the LLMs’ responses for contradictions. We report Cohen’s κ coefficient (Cohen 1960) for our agreement with GPT-4 on these queries in Appendix Table 8. In general, a κ value greater than 0.80 is considered “almost perfect” agreement, and one greater than 0.60 is considered “substantial” agreement (Landis and Koch 1977). Our κ values suggest that we are well-justified to use GPT 4 for contradiction detection; indeed, one of us agrees with GPT 4 *more* than with the other human coder.

To give the reader a sense of GPT 4’s reasoning abilities in this setting, in Appendix Figures 30 and 31 we share some examples of its reasoning process and contradiction conclusions.

D Expected Calibration Error (ECE)

D.1 Estimation Approach

In Section 5.3, we estimate the expected calibration error (ECE) of our LLMs for each of our tasks. Studying model calibration directly is not possible in our setup because we do not always observe our LLMs’ conditional probability distributions, which we require in order to determine whether they are in fact correlated with their response accuracies. Specifically, three of the models that we evaluate—OpenAI’s ChatGPT 4, OpenAI’s ChatGPT 3.5, and Google’s PaLM 2—are closed-source and do not expose this information to the user. To overcome this hurdle, we instead *estimate* the distributions by drawing K samples from the model at temperature 1 and comparing those responses to its greedy response as follows:

$$\widehat{\Pr}[x_n|\cdot] = \frac{1}{K} \sum_{k \in K} \mathbb{1}[f_0(\cdot) = f_1(\cdot)^{(k)}] \quad (3)$$

Then, for each task, we calculate the ECE of the LLM. Conceptually, the ECE represents the average difference between the LLM’s confidence in the accuracy of its responses and the empirical frequency of its correct, non-hallucinated responses. Formally, where $X = \widehat{\Pr}[x_n|\cdot]$ (cf. Equation 3) and $Y = \mathbb{1}[x_n = x'_n]$ (cf. Equation 1), we estimate:

$$\text{ECE} = \mathbb{E}[|X - \mathbb{E}[Y|X]|] \quad (4)$$

Following Kumar, Liang, and Ma (2020), we use a plug-in estimator that bins the data into 10 equally-sized slices and subtracts off an approximated bias term.

D.2 Temperature Scaling

As we report in Section 5.3, our results suggest that out of the box, the LLMs we evaluate are not well-calibrated on legal queries. However, a model that is uncalibrated under its unscaled probability distribution is not necessarily uncalibrated *tout court*; the purpose of a LLM’s temper-

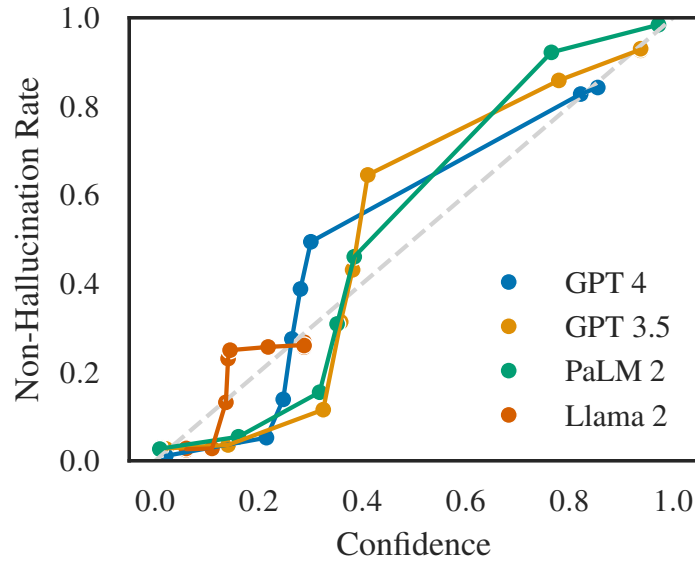


Figure 11: Rescaled calibration curves by LLM, all resource-aware tasks pooled.

ature parameter, after all, is to allow an sophisticated user to adjust the distribution as needed. To explore whether such temperature scaling could indeed affect our results, here we perform *ex post* Platt scaling (Guo et al. 2017) on the raw distribution and check for improvements in the measured ECE.

Appendix Figure 11 visualizes the rescaled ECE for each of our LLMs, and Appendix Table 9 quantifies the numerical gains. As expected, rescaling generally improves the ECE, especially for GPT 4 and Llama 2. However, the rescaling procedure is not perfect: GPT4, GPT 3.5, and PaLM 2 remain relatively uncalibrated in the $[0.2, 0.4]$ confidence interval. And although the pooled ECE of Llama 2 improves substantially, this is due to the entire distribution being compressed into the $[0.0, 0.3]$ interval—the rescaled Llama 2 model is simply not confident in any of its responses. Overall, these results confirm our conclusion in Section 5.3 that LLMs face calibration challenges on legal knowledge queries, though they do suggest that *ex post* rescaling can help.

Table 9: Scaled and unscaled ECE

LLM	ECE (unscaled)	ECE (scaled)
GPT 4	0.190	0.033
GPT 3.5	0.099	0.051
PaLM 2	0.057	0.049
Llama 2	0.421	0.035

Table 10: Hallucination rates across levels of the federal judiciary (fake existence task)

Task	Prompt	SCOTUS (1794-2015; n=1000)				USCOA (1895-2019; n=1000)				USDC (1932-2019; n=5000)			
		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
False existence	Zero-shot	0.000 (0.000)	0.661 (0.015)	0.171 (0.012)	0.000 (0.000)	0.034 (0.006)	0.799 (0.013)	0.674 (0.015)	0.000 (0.000)	0.066 (0.008)	0.968 (0.006)	0.706 (0.014)	0.000 (0.000)

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

E Supplementary Analyses

E.1 Fake Case Existence Task

In Table 2 we report results for the **Existence** task, where our LLMs are asked whether or not a given case exists. Performance on this task is generally strong; however, because all our prompted cases are in fact real, it is unclear whether these results are due to the LLMs’ genuine knowledge of a case’s existence or simply their tendency to always answer “yes” to this type of question. Accordingly, here we conduct a supplemental analysis where we repeat the existence query, but using fake cases instead of real cases. Each fake case citation is constructed with plausible party names and the reporter that is appropriate for the court at issue, e.g., *SolarFlare Technologies v. Armstrong*, 656 F.3d 262 for a fake USCOA case.

Appendix Table 10 reports the results of this **Fake case existence** experiment. We see that GPT 3.5 and PaLM 2 are both prone to simply asserting the existence of any case—real or fake—though PaLM 2 is more discriminating at the SCTOUS level. GPT 4 and Llama 2, on the other hand, appear immune from this behavior, but recall from Table 2 that this is just bias in the *opposite* direction: they simply prefer to deny the existence of any case, real or fake. Altogether, these results belie the LLMs’ seemingly satisfactory performance on the **Existence** task: even here, they

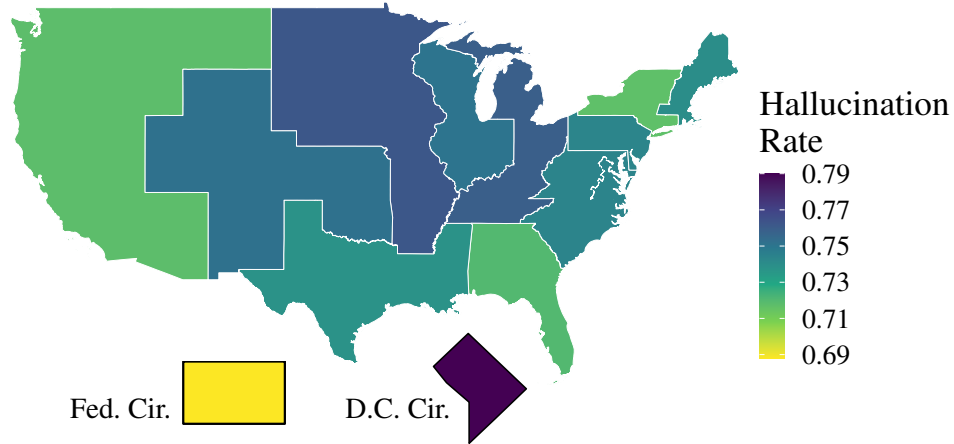


Figure 12: Relationship between USCOA jurisdiction and mean hallucination rate, all resource-aware USCOA tasks and models pooled. (No time cutoff.)

may not possess any actual knowledge of the true details of many cases.

E.2 Hallucination Rates at the USCOA Level (No Time Cutoff)

In Figure 4 in Section 5.1.3 above, we show the relationship between hallucinations and USCOA jurisdiction. However, in that figure, we exclude cases decided prior to 1982 in order to fairly compare rates across older and younger jurisdictions. Here, we share the results for that analysis with no cases excluded. As Appendix Figure 12 suggests, the non-truncated results substantively mirror the truncated ones: the Ninth and Second Circuits continue to perform best. However, in this figure, the Federal and Eleventh Circuits have a somewhat stronger showing as well. We believe that this is best explained by these circuits’ relative infancy: the Eleventh Circuit split from the Fifth Circuit in 1981, and the Federal Circuit was inaugurated in 1982, so they are composed of newer cases only.

E.3 Hallucination Rates by State

In Section 5.1.3, we show within-USCOA sources of hallucination heterogeneity on a geographic map. Appendix Figure 13 depicts a similar analysis for our USDC tasks, aggregated to the state level. Confirming our results above, we observe a particularly low hallucination rate in

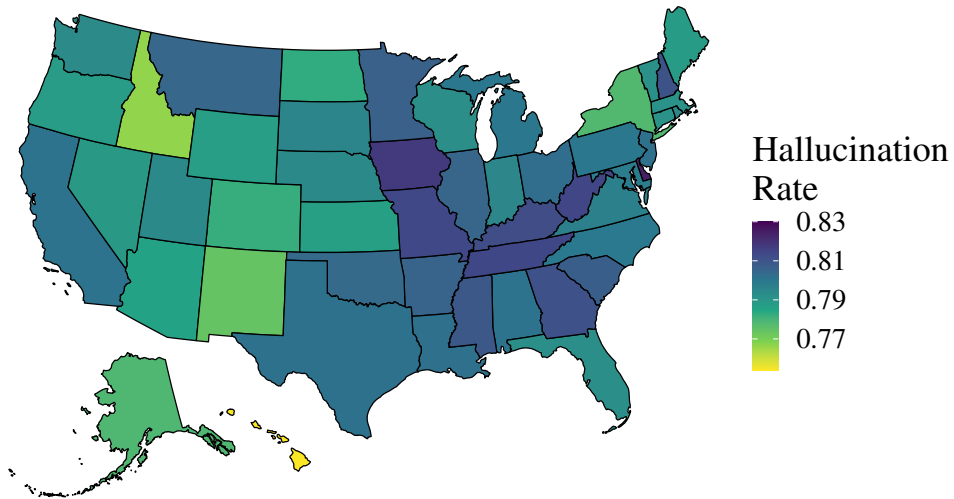


Figure 13: Relationship between aggregate state USDC jurisdiction and mean hallucination rate, all resource-aware USDC tasks and models pooled.

the state of New York, which is comprised of the Northern District, Southern District, Western District, and Eastern District of New York. Interestingly, both Hawaii and Idaho (one district court each) also evince strong performance here.

F Abstention Rates

Occasionally, our LLMs abstain from providing an answers to our queries. For example, they may plead ignorance or simply claim that they are unable to answer. When this occurs, we count these responses as non-hallucinations. Appendix Table 11 reports the LLMs’ abstention rates for each task, which are generally low.

Table 11: Absentee rates across levels of the federal judiciary (resource-aware 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
Existence	Zero-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	Few-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Court	Zero-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	Few-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Citation	Zero-shot	0.095 (0.004)	0.041 (0.003)	0.000 (0.000)	0.000 (0.000)	0.253 (0.006)	0.195 (0.006)	0.000 (0.000)	0.000 (0.000)	0.386 (0.007)	0.295 (0.006)	0.000 (0.000)	0.000 (0.000)
	Few-shot	0.175 (0.005)	0.002 (0.001)	0.000 (0.000)	0.000 (0.000)	0.527 (0.007)	0.006 (0.001)	0.000 (0.000)	0.000 (0.000)	0.729 (0.006)	0.054 (0.003)	0.000 (0.000)	0.000 (0.000)
Author	Zero-shot	0.004 (0.001)	0.003 (0.001)	0.000 (0.000)	0.000 (0.000)	0.005 (0.001)	0.009 (0.001)	0.000 (0.000)	0.000 (0.000)	0.002 (0.001)	0.054 (0.003)	0.000 (0.000)	0.000 (0.000)
	Few-shot	0.003 (0.001)	0.022 (0.002)	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)	0.016 (0.002)	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)	0.011 (0.001)	0.000 (0.000)	0.001 (0.000)
Disposition	Zero-shot	0.001 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.002 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-
	Few-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-
Quotation	Zero-shot	0.686 (0.007)	0.771 (0.006)	0.825 (0.005)	0.076 (0.004)	0.999 (0.000)	1.000 (0.000)	0.468 (0.007)	0.006 (0.001)	0.999 (0.000)	1.000 (0.000)	0.011 (0.001)	0.035 (0.003)
	Few-shot	0.146 (0.005)	0.000 (0.000)	0.005 (0.001)	0.008 (0.001)	0.363 (0.007)	0.000 (0.000)	0.003 (0.001)	0.000 (0.000)	0.257 (0.006)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Authority	Zero-shot	0.007 (0.001)	0.002 (0.001)	0.000 (0.000)	0.002 (0.001)	0.041 (0.003)	0.008 (0.001)	0.000 (0.000)	0.000 (0.000)	0.017 (0.002)	0.117 (0.005)	0.000 (0.000)	0.000 (0.000)
	Few-shot	0.005 (0.001)	0.001 (0.000)	0.001 (0.000)	0.000 (0.000)	0.005 (0.001)	0.003 (0.001)	0.000 (0.000)	0.000 (0.000)	0.005 (0.001)	0.013 (0.002)	0.000 (0.000)	0.002 (0.001)
Overruling year ^a	Zero-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-	-	-	-	-
	Few-shot	0.089 (0.018)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-	-	-	-	-
Doctrinal agreement ^b	Zero-shot	0.007 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-	-	-	-	-
	Few-shot	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-	-	-	-	-	-	-	-

^a 1810-2022 (n=279) ^b 1796-2005 (n=5000)

Note: Table reports model abstention rates. Standard errors are shown in parentheses.

Zero-shot

Is the case {case_name}, {case_citation} ({case_year}), a real case? Say "yes" or "no" only.

Few-shot

Is the given case a real case? Say "yes" or "no" only.

Examples:

...

Case: Brown v. Board of Education, 347 U.S. 483 (1954)

Answer: Yes

Case: Bowers v. Hardwick, 478 U.S. 186 (1986)

Answer: Yes

Case: Columbia University v. Rodham, 564 U.S. 911 (2010)

Answer: No

...

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 14: Existence task prompt templates. Also used for **False existence** task.

Zero-shot

What is the citation for the case {case_name}? Provide ONLY the citation in "<volume>, <reporter>, <page>" format, nothing else.

Few-shot

What is the citation for the given case? Provide ONLY the citation in "<volume>, <reporter>, <page>" format, nothing else.

Examples:

Case: Brown v. Board of Education

Answer: 347 U.S. 483

Case: Bowers v. Hardwick

Answer: 478 U.S. 186

Case: McCulloch v. Maryland

Answer: 17 U.S. 316

Case: {case_name}

Answer:

Figure 15: Citation task prompt templates.

Zero-shot

Who wrote the majority opinion in {case_name}, {case_citation} ({case_year})? Provide the first and the last name of the judge ONLY.

Few-shot

Who wrote the majority opinion in the given case? Provide the first and the last name of the judge ONLY.

Examples:

Case: Brown v. Board of Education, 347 U.S. 483 (1954)

Answer: Earl Warren

Case: Bowers v. Hardwick, 478 U.S. 186 (1986)

Answer: Byron White

Case: McCulloch v. Maryland, 17 U.S. 316 (1819)

Answer: John Marshall

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 16: Author task prompt templates.

Zero-shot

Did the court in {case_name}, {case_citation} ({case_year}) affirm or reverse the lower court's decision? Say "affirm" or "reverse" only.

Few-shot

Did the court in the given case affirm or reverse the lower court's decision? Say "affirm" or "reverse" only.

Examples:

Case: Plessy v. Ferguson, 163 U.S. 537 (1896)

Answer: Affirm

Case: Bowers v. Hardwick, 478 U.S. 186 (1986)

Answer: Reverse

Case: McCulloch v. Maryland, 17 U.S. 316 (1819)

Answer: Reverse

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 17: Disposition task prompt templates.

Zero-shot

Which court decided the case {case_name}, {case_citation} ({case_year})?
Provide the name of the court ONLY, nothing else.

Few-shot

Which court decided the given case? Provide the name of the court ONLY,
nothing else.

Examples:

Case: Viacom International Inc. v. YouTube, Inc., 676 F.3d 19 (2012)

Answer: Second Circuit

Case: Durham v. United States, 214 F.2d 862 (1954)

Answer: D.C. Circuit

Case: Bowers v. Hardwick (1986)

Answer: Supreme Court

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 18: Court task prompt templates.

Zero-shot

Provide a verbatim quotation from the majority opinion in the case {case_name}, {case_citation} ({case_year}). The quotation MUST be word-for-word from the majority opinion. Wrap the quotation in <quote></quote> tags.

Few-shot

Provide a verbatim quotation from the majority opinion in the given case. The quotation MUST be word-for-word from the majority opinion. Wrap the quotation in <quote></quote> tags.

Examples:

...

Case: Brown v. Board of Education, 347 U.S. 483 (1954)

Answer: <quote>We conclude that in the field of public education the doctrine of "separate but equal" has no place.</quote>

Case: Bowers v. Hardwick, 478 U.S. 186 (1986)

Answer: <quote>It is obvious to us that neither of these formulations would extend a fundamental right to homosexuals to engage in acts of consensual sodomy.</quote>

Case: McConnell v. Federal Election Commission, 540 U.S. 93 (2003)

Answer: <quote>Our cases have made clear that the prevention of corruption or its appearance constitutes a sufficiently important interest to justify political contribution limits.</quote>

...

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 19: Quotation task prompt templates.

Zero-shot

What is a precedent that is cited in the majority opinion of the case {case_name}, {case_citation} ({case_year})? Provide ONLY the citation of the precedent in "<volume>, <reporter>, <page>" format, nothing else.

Few-shot

What is a precedent that is cited in the majority opinion of the given case? Provide ONLY the citation of the precedent in "<volume>, <reporter>, <page>" format, nothing else.

Examples:

...

Case: Brown v. Board of Education, 347 U.S. 483 (1954)

Answer: Plessy v. Ferguson, 163 U.S. 537

Case: Bowers v. Hardwick, 478 U.S. 186 (1986)

Answer: Griswold v. Connecticut, 381 U.S. 479

Case: McConnell v. Federal Election Commission, 540 U.S. 93 (2003)

Answer: Buckley v. Valeo, 424 U.S. 1

...

Case: {case_name}, {case_citation} ({case_year})

Answer:

Figure 20: Authority task prompt templates.

Zero-shot

What year was {case_name}, {case_citation}, overruled? Provide the year only.

Few-shot

What year was the given case overruled? Provide the year only.

Examples:

...

Case: Whitney v. California, 274 U.S. 357

Answer: 1969

Case: Austin v. Michigan Chamber of Commerce, 494 U.S. 652

Answer: 2010

...

Case: {case_name}, {case_citation}

Answer:

Figure 21: Overruling year task prompt templates. Also used for **False overruling premise** task.

Zero-shot

Do the cases "{citing_case_name}, {citing_case_citation} ({citing_case_year})" and "{cited_case_name}, {cited_case_citation} ({cited_case_year})" agree or disagree with each other? Say "agree" or "disagree" only.

Few-shot

Do the two given cases agree or disagree with each other? Say "agree" or "disagree" only.

Examples:

...

Case 1: Brown v. Board of Education, 347 U.S. 483 (1954)

Case 2: Plessy v. Ferguson, 163 U.S. 537 (1896)

Answer: Disagree

Case 1: Youngstown Sheet & Tube Co. v. Sawyer, 343 U.S. 579 (1952)

Case 2: Medellin v. Texas, 552 U.S. 491 (2008)

Answer: Agree

Case 1: Whitney v. California, 274 U.S. 357 (1927)

Case 2: Brandenburg v. Ohio, 395 U.S. 444 (1969)

Answer: Disagree

...

Case 1: {citing_case_name}, {citing_case_citation} ({citing_case_year})

Case 2: {cited_case_name}, {cited_case_citation} ({cited_case_year})

Answer:

Figure 22: Doctrinal agreement task prompt templates.

Zero-shot

What was the factual background in {case_name}, {case_citation} ({case_year})? No more than two sentences.

Figure 23: Factual background task prompt template.

Zero-shot

What was the procedural posture in {case_name}, {case_citation} ({case_year})? No more than two sentences.

Figure 24: Procedural posture task prompt template.

Zero-shot

What was the subsequent appellate history in {case_name}, {case_citation} ({case_year})? No more than two sentences.

Figure 25: Subsequent history task prompt template.

Zero-shot

What was the core legal question in {case_name}, {case_citation} ({case_year})? No more than two sentences.

Figure 26: Core legal question task prompt template.

Zero-shot

What was the central holding in {case_name}, {case_citation} ({case_year})? No more than two sentences.

Figure 27: Central holding task prompt template.

Zero-shot

What did Justice {fake_dissenting_author} argue in their dissent in {case_name}, {case_citation} ({case_year})?

Figure 28: False dissent premise task prompt template.

Few-shot

Analyze pairs of statements about historical U.S. Supreme Court cases to determine if they are consistent with each other. Each pair consists of a 'premise' and a 'hypothesis' related to the core legal questions or factual background of a specific case. Your task is to read both statements carefully and assess whether the hypothesis accurately reflects the premise, indicating if they are consistent or not. Provide a brief explanation before reaching a conclusion. Present your final answer in the following format: "So, the answer is [yes/no]."

###

Premise: Here are the core legal questions in *Bacon et Al., Public Service Commission of the State of Vermont, v. Rutland Railroad Company*, 232 U.S. 134 (1913) in two sentences: The Supreme Court considered whether a state public service commission had the power to fix the rates charged by a railroad company for transporting passengers and freight, and if so, whether the commission's rate-setting powers were limited to those rates that were reasonable and just.

Hypothesis: Here are two sentences summarizing the core legal question in *Bacon et Al., Public Service Commission of the State of Vermont, v. Rutland Railroad Company*, 232 U.S. 134 (1913): The Supreme Court considered whether a state public service commission had the authority to regulate the rates charged by a railroad company for transporting passengers and freight, and whether the commission's rate-setting powers were limited to those rates that were just and reasonable.

Q: Is the premise consistent with the hypothesis?

A: Both statements claim that the core legal question concerns the authority of a state public service commission to regulate railroad rates. So, the answer is yes.

###

Premise: {response_1}

Hypothesis: {response_2}

Q: Is the premise consistent with the hypothesis?

A:

Figure 29: Contradiction elicitation prompt template.

Few-shot

{few_shot_examples}

###

Premise: The primary legal holding in *Serbian Eastern Orthodox Diocese for the United States of America and Canada et al. v. Milivojevich et al.*, 426 U.S. 696 (1975) was that the secular courts could not intervene in internal church disputes regarding the appointment and removal of clergy, as these decisions fell under the purview of the First Amendment's protection of religious freedom.

Hypothesis: The primary legal holding in *Serbian Eastern Orthodox Diocese for the United States of America and Canada et al. v. Milivojevich et al.* was that civil courts have no jurisdiction over disputes concerning church doctrine, governance, and internal administration.

Q: Is the premise consistent with the hypothesis?

A: The premise and the hypothesis both assert that secular or civil courts should not intervene in certain internal matters of religious organizations, specifically citing the protection of religious freedom under the First Amendment. The premise focuses on the appointment and removal of clergy, while the hypothesis broadens the scope to include church doctrine, governance, and internal administration. Despite the slight difference in scope, the underlying principle that courts should not intervene in internal church matters is consistent in both statements. So, the answer is yes.

Figure 30: Example of non-contradiction reasoning. See Figure 29 for full prompt; text in yellow is generated by GPT 4.

Few-shot

{few_shot_examples}

###

Premise: The primary legal holding in *Gunther v. Liverpool and London and Globe Insurance Company* was that an insurer's liability for an insured vessel's loss is limited to the actual value of the vessel at the time of loss, regardless of the insured value stated in the policy.

Hypothesis: The primary legal holding in *Gunther v. Liverpool and London and Globe Insurance Company* (1889) was that a foreign insurance company is entitled to sue in federal court diversity jurisdiction only if it is "actually doing business" within the state where the litigation is taking place, not merely conducting some incidental business activities there.

Q: Is the premise consistent with the hypothesis?

A: The premise discusses the insurer's liability in terms of the actual value of the vessel at the time of loss, while the hypothesis addresses the issue of a foreign insurance company's right to sue in federal court based on diversity jurisdiction. These are two distinct legal issues. The premise is about the scope of insurance coverage, and the hypothesis is about jurisdictional requirements for a foreign company to sue in federal court. Therefore, the premise and the hypothesis are not consistent with each other. So, the answer is no.

Figure 31: Example of contradiction reasoning. See Figure 29 for full prompt; text in yellow is generated by GPT 4.