Prompting from the bench: Large-scale pretraining is not sufficient to prepare LLMs for ordinary meaning analysis

ABHISHEK PURUSHOTHAMA*, Georgetown University, USA
JUNGHYUN MIN*, Georgetown University, USA
BRANDON WALDON, University of South Carolina, USA
NATHAN SCHNEIDER, Georgetown University, USA

In the U.S. judicial system, a widespread approach to legal interpretation entails assessing how a legal text would be understood by an 'ordinary' speaker of the language. Recent scholarship has proposed that legal practitioners leverage large language models (LLMs) to ascertain a text's ordinary meaning. But are LLMs up to the task? As textual interpretation questions arise in spheres ranging from criminal law to civil rights, we argue it is crucial that models not be taken as authoritative without rigorous evaluation. This work offers an empirical argument against LLM-assisted interpretation as recently practiced by legal scholars and federal judges, who reasoned the large amount of data that models see in training would enable models to illuminate how people ordinarily use certain words or phrases. In controlled experiments, we find failures in robustness which cast doubt on this assumption and raise serious questions about the utility of these models in practice. For the models in our evaluation, slight changes to the format of a question can lead to wildly different conclusions—a vulnerability that parties with an interest in the outcome could exploit. Comparing with a dataset where people were asked similar legal interpretation questions, we see that these models are at best moderately correlated to human judgments—not strong enough given the stakes in this domain.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Applied computing** → **Law**.

Additional Key Words and Phrases: legal interpretation, large language models, legal NLP

ACM Reference Format:
Abhishek Purushothama, Junghyun Min, Brandon Waldon, and Nathan Schneider. 2026. Prompting from the bench: Large-scale pretraining is not sufficient to prepare LLMs for ordinary meaning analysis. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3805689.3812346

1 Introduction

Legal decisions often come down to the interpretation of written text (e.g., in a statute or contract). Usually such interpretation is straightforward. At times, however, the text is appreciably imprecise or ambiguous, leading to disputes about how to apply the text to a particular set of circumstances. In the U.S. legal system, judges often

*Both authors contributed equally to this research.

Authors' Contact Information: Abhishek Purushothama, ap2089@georgetown.edu, Georgetown University, Washington, District of Columbia, USA; Junghyun Min, jm3743@georgetown.edu, Georgetown University, Washington, District of Columbia, USA; Brandon Waldon, bwaldon@mailbox.sc.edu, University of South Carolina, Columbia, South Carolina, USA; Nathan Schneider, nathan.schneider@georgetown.edu, Georgetown University, Washington, District of Columbia, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812346

John is a contractor with insurance that covers property loss, damage, or personal injury claims that arise due to his 'landscaping' work.

John is employed by a family, the Smiths, to install an in-ground trampoline in the family's backyard. A few years after John completes the project, the Smiths successfully sue John for injuries that their daughter sustained while playing on the trampoline. John files a claim with his insurance company to recover losses incurred from the lawsuit.

Considering just how "landscaping" would be understood by ordinary speakers of English, is John covered by the insurance—yes or no?

Fig. 1. A legal interpretation scenario represented as a QA task with binary questions. The example is based on the case *Snell v. United Specialty Insurance Co.* and constructed in the style of our task.

place considerable weight on the 'ordinary meaning' (meaning as would be understood by ordinary speakers of American English) of a legal term [61].[1]

One recent case (*Snell v. United Specialty Insurance Co.*) involved interpretation of the term 'landscaping' (see fig. 1). A second case (*U.S. v. Deleon*) turned on the interpretation of the phrase 'physically restrained' and whether the phrase describes indirectly restricting movement by threatening someone with a gun[2].

How can a judge ascertain the 'ordinary meaning' of a legal text? Often, the judge will deploy armchair intuition buttressed by hypotheticals and dictionaries [34]. On occasion, judges reference corpora [23, 62] and surveys [68].

Enter LLMs. The advent of LLMs has produced considerable excitement in the legal field including for legal text interpretation. Hoffman and Arbel [26] opined that LLMs are valuable resources for ascertaining the ordinary meaning of legal text. Federal judicial opinions [42, 43] have included 'direct queries' to LLMs on the ordinary meaning of 'landscaping' and 'physically restrained'[3] on the grounds that these models supposedly capture and can articulate patterns of ordinary language use.

These developments have been accompanied by a growing body of academic scholarship assessing the benefits, risks, and impact of these technologies from the standpoint of both legal theory and practice. The valence of this work ranges from critical to optimistic. The critical end of the spectrum includes recent evidence that LLMs' interpretive judgments are highly sensitive to how queries about the meaning are phrased, suggesting that LLMs are not reliable tools for legal interpretation [13, 75]. Recent scholarship on the optimistic end of the spectrum suggests, by contrast, that LLMs provide a consistent and accurate window into 'ordinary' speakers' linguistic intuitions [26, 40].

The question of how to establish the 'ordinary meaning' of an expression is a fraught one in legal theory [61, 69, 70]. However, the scale of data that LLMs are trained on, and the fluency of LLM-generated text, have sparked much interest among lawyers and judges [15, 26, 42]. Optimists conjecture that LLMs may prove useful for empirical ordinary meaning analysis, providing more objectivity and comprehensiveness than any individual could. But first, LLMs must be carefully evaluated to establish that they can deliver on this promise.

Hence, evaluations of LLM-based ordinary meaning analysis must consider:

[1]In this paper, we are officially agnostic whether 'ordinary meaning' is a useful legal analytical construct (or even a coherent linguistic concept). Our focus is on whether LLMs are useful for pursuing interpretative methodologies that are purportedly grounded in 'ordinary meaning.'

[2]We provide some additional cases related to other areas of law in appendix A.

[3]In *Snell* and *Deleon*, Judge Kevin Newsom posed the two following direct queries, respectively: "*What is the ordinary meaning of landscaping?*" and "*What is the ordinary meaning of physically restrained?*". See Waldon et al. [75] for a detailed discussion of the direct query method.

(1) whether LLMs—which appear to be able to answer natural language questions but are subject to unexpected or unpredictable **sensitivity** to inputs[4]— are also sensitive to the framing of the question for this task; and
(2) whether LLM judgments are **sound** insofar as they approximate the interpretive judgments of lay human readers [10][5].

We offer these as *necessary* conditions for real-world use, such that failure to meet either one should disqualify LLMs for ordinary meaning analysis. LLMs (or alternative systems) that meet these two criteria might still suffer from other critical drawbacks that render them unsuitable for the task. Both proponents and detractors of LLMs agree that empirical data—particularly surrounding the methodological robustness of LLM-assisted analysis and the correspondence of LLM outputs to ordinary human intuition—can help to advance this debate.

To that end, we conduct a large-scale[6] investigation into the stability of LLM-based legal interpretation methods, with a focus on LLMs' ability to produce ordinary meaning analyses consistent with human native speaker intuitions. We consider whether LLMs' training data *scale*, coupled with the ability to fluently navigate *question–answer style interactions*, translates to the capacity for robust ordinary meaning analysis. Specifically, we investigate whether large-scale pretraining and instruction tuning are sufficient to yield systems that reliably provide sound answers to queries about ordinary meaning. Our study includes GPT-4, a platform system, as a point of reference.

A growing body of legal and NLP research on the linguistic capabilities of LLMs (reviewed in section 2) informs our two hypotheses: (1) LLM judgments are highly sensitive to subtle manipulations in how ordinary meaning queries are posed to the model. (2) LLM judgments are poorly correlated to human judgment.

To test our hypotheses, we create queries based on a previously developed set of 138 scenarios that assess linguistic interpretation in a variety of hypothetical insurance contract disputes [73], and query 14 open-weight models and one closed-weight platform model, across 9 systematic question variants (section 3). We find in section 4 that LLM judgments are inconsistent across model family and size, as well as across choices in question phrasing. The outputs of some instruction-tuned models of specific sizes are correlated to human judgment only in some question variants; moreover, neither decoded tokens nor probability distributions offer a reliable source of human-like 'ordinary meaning' judgments. Though LLMs possess undeniably fluent generation capabilities, our results add to a growing chorus of skepticism about LLMs as tools for legal interpretation (section 5).

Our paper provides the most extensive experimental evaluation of LLMs for legal interpretation to date.[7] We extend previous results on LLM prompt sensitivity to fresh data in legal interpretation. We additionally present an analysis of human correlation using existing human judgment data. Our rigorous experimental and analytical formulation lays the foundation for the evaluation work necessary to establish whether LLMs can be considered trustworthy tools for legal interpretation. We make the code, results, and analysis available publicly[8].

2 Background

Legal interpretation is pervasive in the courts, reaching far beyond a single judge, court, or area of law. Often, judges place considerable weight on *ordinary meaning* in deciding the proper interpretation of the text.[9]

[4]We discuss previous works demonstrating these effects in Section 2.
[5]Consequently, using an LLM specialized to the legal domain or evaluating it with legal experts may be desirable in some circumstances, but would not be in line with 'ordinary meaning' inquiry, which seeks to probe the understanding of the public writ large.
[6]Our study significantly expands the body of evidence that informs the debate of LLM legal interpretation. Previous work has mainly provided qualitative arguments rather than full-fledged evaluation: Waldon et al. [75] test LLMs on just two interpretive scenarios, and Choi [13] utilizes just five.
[7]We have documented specific limitations of the work in Appendix H.
[8]Code is available in the repository: https://github.com/bwaldon/llms-legal-interp
[9]That is, U.S. judges are often inclined to consider how the text would be understood by nonspecialists. As U.S Supreme Court Chief Justice John Roberts recently put it: "So the most probably useful way of settling all these questions [about ordinary meaning] would be to take a poll of 100 ordinary – ordinary speakers of English and ask them what it means, right?" (Oral argument, *Facebook Inc. v. Duguid et al.* [53].)

While such analyses have previously relied on hypotheticals, dictionaries, corpora, and surveys [23, 34, 62, 68, 74], in the current era of generative AI enthusiasm, some scholars have expressed optimism that LLMs will revolutionize legal interpretation. Multiple variations of this idea have been put forward [15, 26, 40, 42, 43].

One such claim is that LLM-based methods need not be perfectly accurate to be useful; they just have to be more robust to misuse than alternatives [26]. However, LLMs have been shown to be sensitive to prompts at multiple levels. Sclar et al. [58] showed how spurious character-level features can lead to large changes in model performance, while Pezeshkpour and Hruschka [49] showed that the order of options can affect model performance in multiple choice question answering. Zhuo et al. [81] have organized such prompt sensitivity into a dataset-independent metric, but use LLM probabilities as measures of confidence. Turning to legal interpretation, Choi [13] tested the robustness of LLMs by generating 2,000 paraphrases for each of the 5 scenarios from Hoffman and Arbel [26], expanding from 20 machine generated paraphrases. Choi [13] found evidence of considerable prompt sensitivity across various models, training methods, and output processing methods. Waldon et al. [75] similarly demonstrated the ease with which an LLM judgment could be manipulated through subtle prompt editing. Taken together, this prior work establishes a need for empirical analysis of how known biases and sensitivities affect LLM legal interpretation. Below, we rigorously test LLMs in legal interpretation in a large set of scenarios, using systematically controlled prompt variants.

A second line of reasoning holds that for hard cases that turn on ordinary textual meaning, LLMs are potentially *better* than humans (and superior to existing tools) because they have been trained on so much ordinary English data (we call this the 'omniscience' argument). Judge Kevin Newsom, in a concurring opinion in *Snell* (the 'landscaping' case), gives voice to this idea, citing the diversity of ordinary usage that these models might draw upon: "models train on a mind-bogglingly enormous amount of raw data … as I understand LLM design, those data run the gamut from the highest-minded to the lowest, from Hemingway [sic] novels and Ph.D. dissertations to gossip rags and comment threads. Because they cast their nets so widely, LLMs can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life" [42]. He appears to tone down the omniscience argument in a subsequent case, saying LLMs "may well serve a valuable auxiliary role as we aim to triangulate ordinary meaning," complementing "traditional interpretive tools [such as] dictionaries" [43].

Is it plausible that a careful judge, armed with established tools of ordinary meaning analysis and a bevy of law clerks, would arrive at a better-informed result because an LLM was consulted? Newsom's rationale is predicated on certain assumptions about LLMs that have been called into question [51, 75]. In particular, his comment about the vast training data behind LLM chatbots suggests he believes systems are capable of conducting *metalinguistic reasoning*[10] about the language in their training data in order to synthesize an interpretive conclusion. But recent scholarship casts doubt on the idea that LLMs are capable of deep metalinguistic reflection [7, 8, 12, 67]. Instead, it seems likely that LLMs are good at imitating or summarizing metalinguistic text in the training data—be it dictionary definitions, textbooks, or online language forum discussions [8, 75]. Thus, to the extent that an LLM produces plausible-sounding responses to an interpretive prompt, it likely draws on what humans *say* about language [2], rather than what they *do* as language users. Still, if LLMs could accurately synthesize ordinary people's opinions about meaning, then they might be a reliable and cost-effective tool for ordinary meaning analysis, as hypothesized by Hoffman and Arbel [26].

It is crucial, then, to test whether LLMs' interpretive judgments are reliably correlated to (ordinary) human interpretations. By some measures, LLM probability models exhibit human-like sensitivity to grammatical phenomena in English sentences: these include linguistic or syntactic acceptability [19, 77, 78], and semantic plausibility [32, 33]. However, there is increasing 'behavioral' evidence of differences between LLMs and humans

[10]Metalinguistic reasoning is the ability to not only use language, but also explicitly reason about the use of language [7], distinct from conventional reasoning that describes 'the general human capacity for truth-seeking and problem solving' [50].

when it comes to learning and processing language [4, 41, 44] and answering questions [64]. To test whether LLMs and humans arrive at similar interpretive conclusions, we use a dataset of human judgments reported by Waldon et al. [73] to evaluate LLMs' interpretive judgments.

3 Legal Interpretation with LLMs

Consider the landscaping example in fig. 1. A term within the contractor's insurance contract, 'landscaping,' must be interpreted to determine whether the insurance covers the described scenario. The task explicitly demands a judgment as to how ordinary speakers would understand the contract language in context. This judgment may not cohere with one's beliefs regarding the 'correct' interpretation of the provision, or regarding how a judge will actually resolve the legal dispute at hand.[11] Our assessment metrics assume that 'ordinary interpretation' is both highly varied and to some extent subjective. Our first metric—robustness to variation—evaluates the stability of model judgments across multiple prompt formulations. Our second metric—human correlation (see section 4.3)—evaluates the extent to which model judgments cohere with the intuitions of a relevant human population. In the remainder of this section, we describe a study designed to investigate LLMs' legal interpretation capabilities as assessed against these two metrics.

Materials. Our study adapts materials originally developed by Waldon et al. [73] for a human study of legal interpretation and consists of 138 items based on real-world insurance contracts. Table 1 provides an example item. Each item names a category of insurance coverage (e.g., Vehicle Damage) and provides a definition of that category. The item then describes a policyholder's loss, which may or may not be covered by the named category. Waldon et al. [73] analyzed human responses from 1,338 U.S.-based native English speakers recruited via Prolific. The participants were shown a vignette composed of the scenario and insurance text. They were then asked to determine whether the vignette's protagonist (e.g., Ken) was covered by the insurance, with three response options: Yes, No, or Can't Decide. A sample vignette (fig. 5) can be found in Appendix B.

Query element	**Example ('Vehicle Damage')**
Insurance text	Steve's car insurance policy includes coverage for "Vehicle Damage," defined as "loss or damage to the policy holder's 1) car; or 2) car accessories (while in or on the car)"
Scenario	One day, Steve is involved in a minor accident. His GPS navigation system, which was in the car at the time, was damaged. Steve files a claim with his insurance company for the damage.
Framing for 'ordinary meaning'	Considering just how "accessory" would be understood by ordinary speakers of English,
Question	is Ken covered by the insurance—yes or no?
Cue	Final answer is:

Table 1. Elements of the interpretive queries. 'Insurance text' and 'Scenario' are sourced from the study item, and the 'Framing' and 'Question' are added to solicit interpretive judgment. The 'Cue' was found with exploratory tests to induce high rate of first token judgment.

Interpretation as binary QA. Following the design of the human experiments, we query LLM interpretation with binary questions. As illustrated in table 1, we frame polar judgment in legal interpretation as binary QA, where cues[12] constrain the output space (e.g., that it should answer directly with "yes" or "no"). The LLM judgment is

[11] Our investigation is perspectivist, as we are not assessing model behavior against a single, fixed 'ground truth' linguistic interpretation or judicial outcome [17].

[12] Explicit cues (e.g. "Is Ken covered by the insurance—yes or no? Final answer is:") are likely different from how judges may use LLMs. But for empirical evaluation, we believe our `Yes/No` variant (table 2) is able to characterize a wide range of felicitous responses to how judges actually use LLMs—likely with simple questions (e.g. "Is Ken covered by the insurance?").

operationalized as the first output token probability. Our design allows us to obtain probabilities which can be used for more robust analysis, including correlation to human judgment. It is also robust against incongruent text output, where the output string does not have a clear response. Motivated by prior work suggesting a lack of alignment between token probability and intended model judgment [76], we manually compare our operationalization of LLM judgment as first token probability to manual extraction from decoded output and show that our operationalization is more robust (Appendix D).

Additionally, we consider several types of yes and no tokens across casing as indicators of coverage judgment, and operationalize their respective sums as judgment probabilities, which we further describe in Appendix E.1. However, we do not assume a direct association between token probability and judgment confidence. This is true of our robustness analysis and human correlation analysis. For human correlation analysis, we specifically motivate a linking hypothesis detailed in section 4.3.

Formulation of prompt variants. To investigate whether LLMs' interpretive judgments are robust to minor changes in prompt design, we constructed a template of 9 question variants and applied that template to each of the 138 items. An example paradigm is presented in table 2.

Variant	**Question**
	Considering just how the word 'landscaping' would be understood by ordinary speakers of English,
`Yes/No`	is John covered by the insurance—yes or no?
`No/Yes`	is John covered by the insurance—no or yes?
`Negation`	is John not covered by the insurance—yes or no? Final answer is:
`Agreement`	do you agree with the statement, "John is covered by the insurance"—yes or no?:
`AgrWithNeg`	do you agree with the following statement: "John is not covered by the insurance"— yes or no?
`Disagreement`	do you disagree with the following statement: John is covered by the insurance"— yes or no?
`DisagrWithNeg`	do you disagree with the following statement: John is not covered by the insurance"— yes or no?
`Options`	is John covered by the insurance? Options: A. John is covered. B. John is not covered.
`OptionsFlipped`	is John covered by the insurance? Options: A. John is not covered. B. John is covered.

Table 2. Systematic variation of the question (table 1) in the interpretive queries. See further discussion in Section 4.2.

Some variants reflect phenomena that are already attested to be challenging for LLMs. García-Ferrero et al. [18] and Truong et al. [71] show, for example, that LLMs find natural language negation words challenging and lack a deep understanding of the phenomenon. Moreover, Sharma et al. [59] and Hong et al. [27] demonstrate that in some contexts, LLM outputs are modulated by prompts that overtly solicit agreement (e.g., *Do you agree that…?*). Our study builds upon these previous findings in the domain of legal interpretation.

For most variants, an affirmative yes response would correspond to the Covered judgment, but in some variants (e.g., the `Negation` variant in table 2), the Covered judgment would be expressed with a no token. For clarity, we report and discuss probabilities corresponding to Covered and NotCovered judgments, rather than discussing yes and no token probabilities. We note that the `DisagrWithNeg` question variant is complex even for humans, as both negation words and negative prefixes add to the cognitive load [16, 57, 60]. We still include the variant to evaluate whether language models are affected by such increased syntactic and semantic complexity and to assess their robustness in handling convoluted equivalence in prompt variation.

Models. Although our evaluation is focused on robustness and human correlation, our choice of models (i.e., exclusively pretrained or minimally post-trained models) isolates a theoretically relevant subclass of LLMs given the state of legal discourse around the use of LLMs for legal interpretation. In addition to pretraining

and instruction tuning, various methods have been employed to improve LLMs in specific directions, including aligning LLMs to generate responses closer to *human* preferences [37, 82], training LLMs to follow procedural (e.g., chain-of-thought) steps [14], and applying inference meta-algorithms [79]. However, pretraining is still the primary stage of LLM development, with instruction tuning providing the next distinct general purpose stage.

Hence, our model selection includes both 'base' (pretraining only) and instruction-tuned models, of varying sizes up to 70B parameters, and GPT-4 [13]. They include `Llama` [22], `OLMo` [25], `Mistral` [31], and `Gemma` [66] and, we include `GPT-2` [52] both as a model representing the smaller end and one that can support future careful investigation into how pretraining methodologies influence performance on this task.[14] The full list is in table 3, and implementation details in Appendix E. Given our evaluation data's release in 2023, data contamination is a possibility, which we discuss in Appendix C.

As discussed in section 2, ordinary meaning analysis asks how language would be understood by a lay speaker of English—not a legal expert such as a lawyer or a judge. We thus evaluate general-domain LLMs rather than models specialized for legal text and legal tasks (e.g., LegalBERT [11]).

Family	**Models**
Llama-3	`1B`, `1B-Inst` (3.2), `3B`, `3B-Inst` (3.2), `8B`, `8B-Inst` (3.1), `70B` (3.1), `70B-Inst` (3.3)
GPT	`GPT-2-medium`, `GPT-4`
OLMo-2	`7B`, `7B-Inst`
Ministral	`8B-Inst`
Gemma	`7b`, `7b-it`

Table 3. Our experiments include 15 models across 5 families: all are pretrained, some are instruction-tuned, and span range of sizes. GPT-2 is the smallest, a reference model on the low end (for considerations of modified pretraining investigations) and GPT-4 is a closed reference model on the other end (and has gone through additional stages such as alignment).

4 Results and Analysis

From the models, we collect both categorical (Covered or NotCovered) and distributional (probabilistic) judgments over tokens that represent the Covered judgment (yes, Yes, YES for `Yes/No` variant), tokens that represent the NotCovered judgment (no, No, NO for `Yes/No` variant), and the residual other tokens. We use these judgments to analyze how robust the models are with respect to variation across family, size, and question. Additionally, we examine correlation (or lack thereof) between human and LLM judgment.

4.1 Analysis of judgments for `Yes/No` prompts

We begin by focusing on what is arguably the most basic question variant: `Yes/No`, illustrated in table 1. The categorical and distributional judgments for all models with the `Yes/No` question variant are reported in table 4.

Some models behave like stopped clocks. We observe that many models repeat the same categorical judgment across all scenarios, or are highly biased towards one judgment: 6 of the 15 models tested provide the same response to `Yes/No` questions representing more than 127 (92%) of the 138 tested scenarios. Three models provided the same judgment for all 138 scenarios. In these cases, it is doubtful that the model judgment reflects substantive engagement with the provided scenario. These models, rather, are only as useful as the proverbial stopped clock that correctly tells the time twice a day.

[13] This is a closed weight model, whose judgment was obtained via API. See Appendix E.2 for details.

[14] In future work, we also plan to investigate the extent to which the ordinary meaning judgments of such models reflect generalizations of language use as observed in the models' pre-training data (or, e.g., reflect metalinguistic *mentions* in that data).

	Categorical Counts		Distributional Spread			Categorical Counts		Distributional Spread	
Model	Covered	NotCovered	Min	Max	**Model**	Covered	NotCovered	Min	Max
Llama-70B	28	110	0.21	0.48	OLMo-2-7B	70	68	0.19	0.56
+Inst	59	79	0.02	0.85	+Inst	53	85	0.00	0.99
Llama-8B	80	58	0.14	0.37	Ministral-8B-Inst	75	63	0.21	0.59
+Inst	0	138	0.11	0.74	gemma-7b	39	99	0.19	0.49
Llama-3B	127	11	0.09	0.52	+it	131	7	0.00	1.00
+Inst	53	85	0.16	0.69	GPT-4	77	61	0.00	1.00
Llama-1B	138	0	0.06	0.29	GPT-2-medium	5	133	0.13	0.31
+Inst	138	0	0.15	0.59	Human Majority	84	54	–	–

Table 4. Count and probability range for each model's Covered and NotCovered judgments in response to Yes/No questions. Both distribution and the effect of instruction tuning vary significantly across models. Human majority is also provided for reference.

Different distributions for each model. We observe that each model allocates its distributional judgments in different ranges. For example, GPT-4 allocates probabilities in a wide range $[0.0, 1.0]$, while Ministral-8B-Inst allocates them in a much narrower range $[0.19, 0.58]$. However, this variation across models do not represent the varying confidence of the models' interpretive judgments (Section 3). Rather than basing analysis on the absolute values of the probabilities, we look at the distributions separately for each model to determine whether a given model provides a useful signal for interpretation.

Instruction tuning is associated with a wider range of judgment probabilities but also introduces unpredictable bias. Across the board, instruction-tuned models utilize a wider range of judgment probabilities than their base counterparts, with instruction-tuned OLMo, gemma, and GPT-4 models utilizing the entire space in $[0, 1]$ as shown in table 4. However, other changes introduced by instruction tuning are less consistent—the magnitude and direction of changes in both categorical and distributional judgment varies by model. For example, instruction tuning on Llama-8B leads to a predominance of NotCovered judgments, while instruction tuning of Llama-70B leads to predominance of Covered judgment, when prompted using the Yes/No question variant.

4.2 Robustness to question variation

For a model to be considered reliable at answering interpretive questions, it should be robust to minor variation in how the question is phrased. As described in section 3, we measure model responses while varying the phrasing of the question in the prompt and leaving the content unchanged, and analyze how the variation affects the models' categorical and distributional judgments. We report three major findings in this section of our study.

Given the nine question variants, for each item and model, one of the categorical judgments (Covered, NotCovered) will be the majority judgment and the other the minority judgment. The strength of the majority can vary—from 5 of 9 variants (an indication of brittleness) to 9 of 9 variants (unanimity, indicating robustness). A frequency table for majority judgments collated by items for each model is shown in table 5.

Ubiquitous lack of consistency across question variants. We observe lack of consistency across the 9 question variants in each LLM we study. As illustrated in table 5, in 2,061 of 2,070 model item-model combinations (138 scenarios for each of the 15 models), both judgments can be found across the question variants with each model; only in 9 item-model combinations is the judgment fully consistent across all question variants.

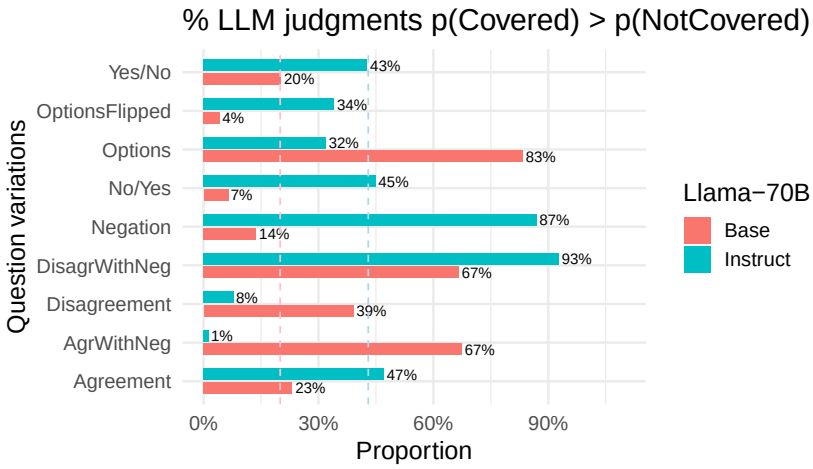

Fig. 2. `Llama-70B` model responses across question variants, each of which results in a large shift away from values in either directions given the `Yes/No` variant, indicated with the dotted lines.

Model judgments are sensitive even to variations such as reversing the order of the provided answer choices or introducing a negation word, as seen in fig. 2 for `Llama-70B` models, where the swap from `Agreement` to `AgrWithNegation` prompt type produces a 44% absolute shift in the base model, and a 46% absolute shift in the instruction-tuned model, with respect to the rate of categorical COVERED judgments. This ubiquitous lack of consistency exposes a generalization failure on the part of the models. Worse, it invites users invested in a particular outcome to engage in "prompt shopping," varying the prompt until the desired response is produced [51, 75].

Some questions are more likely to elicit minority judgments. While unanimity is difficult to come by, some question variants are still more likely to induce the 'minority judgment' which disagrees with the majority of judgments produced by the same model given the same scenario. As shown in table 6, we find that the `Disagreement` variant yields the minority judgment most frequently, while the `Yes/No` variant is least likely to yield the minority judgment. Of nine question variants we investigate, `Disagreement`, `AgrWithNegation`, `DisagrWithNegation` and `Options` are the most frequent elicitors of minority judgment, accounting for 67% of minority responses. However, even in the absence of these four variants, unanimity occurs only in 33% of items.

Some question variants have a stronger distributional effect than others. We quantify the impact of each question variant (relative to the default `Yes/No` variant) by measuring the distance between 1) the judgment distribution given that variant and 2) the judgment distribution given the default `Yes/No` variant. We measure distribution distance with the Jensen-Shannon distance [JSD; 46], which is based on KL divergence [35] but is symmetric and provides a distance measurement between 0.0 (identical distributions) and 1.0 (maximally different). The question variant that yields the most distant distribution for each model is listed in table 13 (appendix D), where average distance is obtained by calculating JSD at the item level then aggregating by question variant for each model. The `Disagreement with negation` variant most frequently (5 of the 15 models) leads to the biggest change in the distributional judgments in the most models, including up to 0.56 for GPT-4. The lack of robustness supports our first hypothesis, and should be considered a limitation of LLMs for use in legal interpretation, since such inconsistency shows models are brittle in the face of superficial prompt variation.

Model	5	6	7	8	9	Model	5	6	7	8	9
Llama-70B	33	68	36	1	0	Ministral-8B-Inst	24	65	30	18	1
+Inst	25	33	78	2	0	OLMo-2-7B	46	65	20	7	0
Llama-8B	40	52	46	0	0	+Inst	51	57	30	0	0
+Inst	6	39	59	31	3	Gemma-7b	44	58	29	7	0
Llama-3B	95	40	3	0	0	+it	9	31	79	19	0
+Inst	75	48	15	0	0	GPT-4	4	9	57	63	5
Llama-1B	12	57	69	0	0	GPT-2-medium	50	83	5	0	0
+Inst	129	9	0	0	0						

Table 5. Number of items by number of question variants that yielded the majority judgment for the model. For example, there were 33 items for which `Llama-70B` produced one judgment for 5 variants, and the other judgment for 4 variants. Each judgment is a binary choice between Covered and NotCovered. 9 variants producing the same judgment indicates unanimity, which occurs in 3 items for `Llama-70B-Inst` and in 5 for `GPT-4`.

Variant	Count	%Minor.	Variant	Count	%Minor.
Disagr.	1256	21	Options F.	493	8
Agr. w/ Neg.	1045	17	Negation	489	8
Disagr. w/ Neg.	918	15	N/Y	275	5
Options	809	14	Y/N	188	3
Agr.	501	8	**Total**	**4975**	

Table 6. The number of minority judgments for each question variant, and the percentage proportion in minority judgments. The counts are sorted vertically in descending order. An equal proportion would lead to a 0.09 proportion for each variant.

4.3 Correlation to human judgment

Finally, we compare LLMs' responses and distributional judgments to human judgment data from Waldon et al. [73] to test our hypothesis that model responses are poorly correlated to human judgments. In the Waldon et al. [73] study, each human participant was asked for a response of yes, no, or can't decide (yes and no correspond to Covered and NotCovered judgments, respectively).[15] The dataset contains judgments for all 138 items with total of judgments from 1346 participants, with an average of 30 judgments per scenario.

If LLMs are good models of population-level interpretive consensus among lay human speakers, we expect there to be model-internal quantity (or a derived value) to bear a monotonic relationship to the degree of human consensus. In our correlation analysis, the dependent variable is the proportion of human Covered responses on the same item. We use LLM judgment (a measure derived from LLM token probability) as the independent variable. We explicitly avoid conflating LLM probability with the confidence of the prediction being correct; token probability distributions represent generation probabilities, not estimates of confidence [29], and current LLMs are not sufficiently calibrated to assume a correlation between probability and confidence [30].

Our linking hypothesis posits that LLM judgment can be linearly correlated to human consensus. We link the endpoints of the LLM-derived measure $p(\text{Covered}) = \{0, 1\}$ to unanimous human judgments and consider four operationalizations that satisfy this boundary condition. We find R^2 values to be stable across the four operationalizations (reported in the appendix table 14 for completeness), suggesting that our linking hypothesis

[15] More details on the response collection is provided for reference in appendix B).

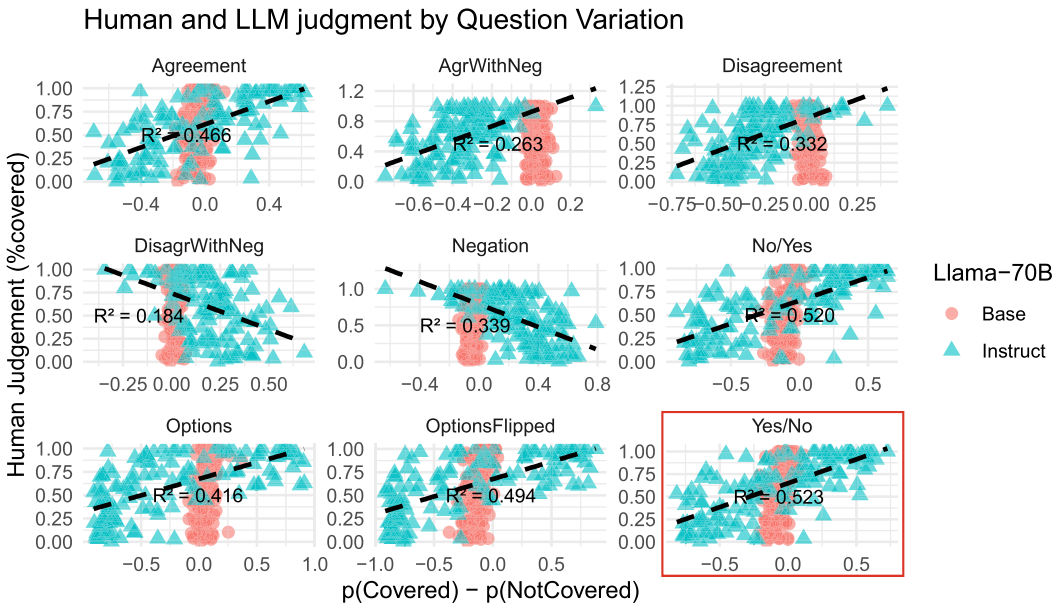

Fig. 3. `Llama-70B` model probabilities versus human consensus across question variants. Dotted lines and the corresponding R^2 are best best-fit lines between human and Instruct LLM. The `Yes/No` question variant, highlighted in red, represents the highest $R^2 = 0.523$ value for the model, a moderate correlation.

is robust to minor variations in formulation. [16] The operationalization of LLM judgment used in the rest of the paper is the difference in model probabilities $\Delta = p(\textsc{Covered}) - p(\textsc{NotCovered})$ due to its benefit in representing consensus in either direction. With this operationalization, our linear linking hypothesis accounts for the probability judgments for both Covered and NotCovered with the decision boundary at 0. A perfect correlation would be fit to the equation $\%\textsc{Covered} = 0.5\Delta + 0.5$.

We use this linking function to perform a R^2 analysis [80] with LLM judgment as the independent variable in determining the proportion of human majority judgment. Our analysis is not a benchmarking objective, rather an analytical view at how much signal of *human* judgment they provide.

Only some LLMs are moderately correlated to human judgment some of the time. Our results show that only larger instruction-tuned models' judgments exhibit some correlation to human judgment, and only for some specific question variants. This is despite evidence for scaling–models do show better fit to human data and stronger performance in predicting human consensus as the number of parameters (and likely the amount of training data) increases. In particular, only instruction-tuned models with 70 billion parameters or more report an R^2 value greater than 0.5, as shown in table 7 and fig. 7. Correlation between `Llama-70B-Inst` response and human judgment across question variants is illustrated in fig. 3, where one variant yields a negative correlation ($m < 0$) with a R^2 value of 0.18. The significant variation in R^2 values across question variants suggests that the models' responses are not, in general, representative of human-like judgment, but rather highly influenced by

[16] We acknowledge that there may be other monotonic linking functions that could generally lead to higher R^2, but given the goal and scope of this work, we stick to a linear link as a reasonable default.

prompt design and question form. This highlights a second limitation to LLMs in legal interpretation: at best, the models we examine achieve moderate correlation to human judgment in select model-question variant pairs.

Furthermore, even when there is correlation between token probabilities and human judgment, the output tokens can be overwhelmingly biased. `Llama-70B-Inst`'s responses to the `AgrWithNeg` variant shown in fig. 3 illustrates that despite a (weak) correlation of $R^2 = 0.26$ when examining gradations of probabilities, the actual responses are overwhelmingly NotCovered, where $\Delta < 0$, further raising caution to LLMs' use for legal interpretation.

		Corr. Analysis		Oracle				Corr. Analysis		Oracle	
Model	**Var.**	***m***	**R^2**	**Acc.**	**Thr.**	**Model**	**Var.**	***m***	**R^2**	**Acc.**	**Thr.**
Llama-70B	Opt. F.	2.0	0.25	0.69	−0.17	Ministral-8B-Inst	Yes/No	0.97	0.27	0.69	−0.11
+Inst	Yes/No	0.65	**0.52**	**0.79**	−0.28	OLMo-2-7B	No/Yes	1.3	0.07	0.64	−0.10
Llama-8B	Agr.	1.1	0.04	0.58	−0.02	+Inst	Yes/No	0.21	0.18	0.64	−0.63
+Inst	Options	1.1	0.20	0.67	−0.36	gemma-7b	Yes/No	1.4	0.10	0.59	−0.10
Llama-3B	Disagr.	−0.67	0.03	0.42	0.11	+it	Neg.	0.21	0.06	0.60	−0.16
+Inst	Yes/No	0.46	0.06	0.59	−0.25	GPT-4	Yes/No	0.26	**0.60**	**0.83**	−0.25
Llama-1B*	No/Yes	−0.91	0.01	0.38	0.13	GPT-2-medium*	Options	0.43	0.01	0.62	−0.19
+Inst*	Neg.	−0.37	0.01	0.38	0.34	Majority class baseline				0.61	

Table 7. For each model, we report the question variant with the highest correlation to human judgments along with the R^2. We also report the accuracy with an oracle threshold (knowing the labels) of a binary classifier predicting the human majority judgment from the LLM probability difference Δ. We offer the always-Covered accuracy of the majority class as a reference baseline. The correlations except three indicated with a * have a p-value < 0.05. These are also the three weakest correlations. Unlike p-values reported for performance with random sampling effects, the p-values provided is a significance measure of the *best-fit* line m and b, rather than the coefficient of determination R^2. Refer to fig. 7 for a detailed visualization of each model-prompt pair. An oracle model uses LLM judgment to linearly predict binary human consensus.

Even the best-correlated LLMs are unreliable predictors of human judgment. As another measure of correlation, we consider an oracle threshold classifier that predicts the binary human judgment as a function of the LLM's probability-difference Δ. The threshold is set based on the human data to match where the best-fit line crosses 0.5 in human judgment. Given a probability difference value exceeding the threshold, the classifier predicts the Covered judgment; otherwise, it predicts NotCovered. Intuitively, this tells us how accurate the LLM would be at giving probabilities consistent with the human majority judgment, if only we knew how to perfectly interpret the model-specific probability scale. Binary classification accuracies appear in the penultimate column of table 7.

The model-variant pair best correlated to human judgment, `GPT-4` responses to `Y/N` ($R^2 = 0.60$), yields a linear model that accurately predicts human consensus 83% of the time. Even if this may initially appear satisfactory, we contend that even 1 in 6 binary classification error rate in an interpretative tool as considered in Hoffman and Arbel [26], Newsom [42, 43] defeats its purpose. For comparison, a majority-class classifier always voting Covered would be correct 61% of the time.

We describe an example item where `GPT-4` judgment drastically differs from human judgment in fig. 4, where `GPT-4` overwhelmingly prefers the Covered judgment to the NotCovered judgment with $p(\text{Covered}) - p(\text{NotCovered}) = 0.85$. For comparison, only 17% of human judgments predicted Covered, preferring the NotCovered judgment.

Response incongruence: answering the wrong question. On some question variants, model responses are sometimes negatively correlated to human judgment when the question includes negation, which suggests that LLMs' attested difficulties with negation [18, 72] persist in legal interpretation. However, we find this behavior to be

Belinda's home insurance policy includes coverage for "Escape of Water" damage, defined as "damage to buildings and contents caused by a sudden, unforeseen flow of water from an internal source, such as a burst pipe or faulty appliance."

> Belinda lives in a two-story house with a bathtub on the second floor. One day, Belinda, a sculpture artist, fills the tub with paint in order to dip her sculptures in paint, but the tub suddenly cracks and begins to leak, causing the paint to seep through her floor and damage her walls. Belinda files a claim with her insurance company for the damage.

Considering just how "internal source" would be understood by ordinary speakers of English, is Belinda covered by the insurance—yes or no? Final answer:

Fig. 4. Prompt to which GPT-4 overwhelmingly prefers the Covered judgment to the NotCovered judgment with $\Delta = 0.85$, compared to human judgment %Covered = 0.17.

unpredictable across model type, question variant, and scenario. We observe negative correlation $m < 0$ between human judgment and difference in LLM judgment probability in `DisagrWithNeg` and `Negation` prompt types in fig. 3, in `DisagrWithNeg` prompt type in fig. 6. Another version of response incongruence is due to failure in instruction-following. From a manual review of all 1,242 responses from `Llama-70B`, we find that a total of 14 responses to `Options` and `OptionsFlipped` may be incongruent, conveying judgment in explicit response rather than selecting a provided choice. See Appendix D for more detail.

On the most optimistic interpretation of the results—that the largest models with certain prompt types achieve nontrivial correlation, which might be improved with further engineering to the point of reliability—we underscore two points of caution: (a) we have not ruled out the possibility of data contamination (Appendix C), and (b) our experimental data is specific to insurance contract scenarios. Thus, there is a long road ahead for any efforts to develop and verify LLMs for practical use.

5 Related Work

Legal NLP, legal language, and LLM robustness are popular areas of research. Nonetheless, there is limited scholarly work on the utility of LLMs for novel interpretation questions in the law, and most of it is based on conceptual arguments and case studies rather than controlled experimentation.

In this section, we engage with scholarship on LLM-based legal interpretation, as well as some broadly related studies in legal NLP.

Legal interpretation. Our study asks whether LLMs can reliably answer novel questions about ordinary meaning. Several recent papers have expressed skepticism. Lee and Egbert [38] contrast LLMs with corpus linguistics, arguing that LLM interpretations are less replicable, transparent, and generalizable. Waldon et al. [75] argue that the way LLMs are designed makes them vulnerable to misinterpretation and misuse. In a similar vein, Grimmelmann et al. [24] argue that LLMs are unproven for legal interpretation due to a *reliability* gap (LLMs are not always consistent and reproducible), and an *epistemic* gap (interpretive conclusions in text produced by the model are not necessarily accurate measurements of ordinary meaning as understood by humans). Our robustness and correlation evaluations are ways of quantifying the respective gaps.

In prior literature, claims about LLM reliability have been supported with ad hoc case studies—with the exception of Choi [13], who conducted a set of experiments to test LLMs' sensitivity to prompt variation. While our goal and conclusions are similar to those of Choi [13], there are important methodological differences: (i) Choi

utilized 5 contract scenarios from Hoffman and Arbel [26] while we use 138 insurance policy scenarios [from 73]; and (ii) we perform controlled, systematic variation of the questions for our studies rather than utilizing LLMs to generate prompt variants. Both our results and Choi's underscore the need for caution when contemplating applying LLMs to legal interpretation.

Other studies have compared LLMs' interpretive judgments to those of lay human respondents. Waldon et al. [73] find that InstructGPT [47] over-predicts human consensus in their interpretive scenarios but that few-shot prompting mitigates this behavior. With the same items and human judgments, we conduct a different set of experiments focused on analyzing a range of newer models. In a follow-up to our study, Petersen et al. [48] collect new human judgments on materials based on Waldon et al.'s stimuli and investigate susceptibility to prompt variation in a newer model (GPT-5 with and without reasoning). Petersen et al. find that while GPT-5 with 'reasoning' is relatively robust to prompt variation, correlation to human judgment is still lacking.

Martínez [40], using a different dataset derived from previously adjudicated U.S. cases that involved ordinary (or plain) meaning analysis[17], compares interpretive judgments of different human groups (judges, lawyers, and laypeople) and LLMs with `Yes/No` questions. Most notably in the context of our study, Martínez [40, p. 49] reports that `GPT-4`'s response is consistent with the majority of laypeople for 78% of the items. (Filtering to the subset of items with a clear majority among laypeople, the model is consistent with that majority 83% of the time.) Martínez [40, p. 63] concludes that, when compared against the *mismatch* of judgments between judges, lawyers, and layman surveyed, this shows "LLMs can, under controlled conditions, reliably track the consensus . . . at at least as high a rate as the court". Quantitatively, our results with different data and experimental methodology are broadly compatible: note the 83% oracle accuracy for `GPT-4` in table 7. Nevertheless, we take a different perspective on the implications. We argue that, even if prompt instability were not a problem, judges should think twice before taking at face value a system that gives the wrong answer a fifth of the time (without a well-calibrated indication of confidence).

Scope of other studies on LLM instability in legal interpretation. While this paper is not the first to report that LLMs are unstable when it comes to legal interpretation, prior studies differ with respect to their empirical scope.

Waldon et al. [75] perform a red-teaming exercise grounded in the 'directed queries' employed in Newsom [42] and Newsom [43]. They then perform subtle variations to the queries specifically to elicit counter (metalinguistic) judgments. This demonstrates instability in legal interpretation with an adversarial approach. Their investigation utilizes two samples and specific manipulations for each to elicit counter judgments.

Choi [13] is grounded in the five scenarios used in Hoffman and Arbel [26]. Choi utilizes Claude 3.5 [3] to generate paraphrases of a scenario, including the question and the term under interpretation. Choi generates 2000 paraphrased scenarios for each of the five scenarios, and utilize a relative probability measure (between the two options) to look at instability.

By contrast, our investigation utilizes 138 human-constructed scenarios based on 46 real-world policy scenarios from Waldon et al. [73]. Additionally, our variations are designed hold the literal content of the question presented as constant as possible while Choi's LLM-generated variations at times involve more substantial changes in wording. Moreover, our prompt variants are systematic, without adversarial manipulations, and do not use model-generated data. (Petersen et al. [48] echo our design in this respect but employ fewer variants.)

Legal NLP. In the broader space of Legal NLP [5], two recent studies are especially relevant to this work. Luo et al. [39] instruct LLMs to consider a target legal term (e.g. 'landscaping') and generate, based on legal documents from previous cases, an interpretive explanation (along with conditions governing the applicability of the interpretation). This task concerns the extraction of a previously articulated interpretation, rather than novel

[17]We refer readers to Basile [6] for further discussion of these concepts.

conclusions about meaning. They report system performance comparable to that of human experts for the task, but they do not evaluate models for robustness across prompts.

Blair-Stanek and Van Durme [9] investigate LLM instability in the context of a legal judgment prediction task, where the models are prompted with a synopsis of the case and asked to predict which party should prevail. This task demands full accounting of all the issues in a particular case, rather than a focused interpretive question. There is no indication that the cases were selected with emphasis on language interpretation. Investigating closed commercial models with a 'reasoning' step, they find significant instability; i.e., the answer is not even consistent given the same model and prompt.

6 Conclusion

Given the excitement about and increasing critical legal scholarship on the use of LLMs for legal interpretation, we conducted a systematic study of this capability with a focus on LLM judgments regarding the 'ordinary meaning' of legal language, formulated as binary-choice QA. We examined LLMs' judgments for both robustness to prompt variation and correlation to human judgments, finding: (i) Some LLMs behave like stopped clocks, with a strong tendency to provide the same judgment for most input, regardless of the scenario. (ii) Models show a ubiquitous lack of consistency across question variants. (iii) Correlation with human judgment is at best moderate, and is strongest in larger, instruction-tuned models.

Our findings inform a growing body of scholarship on the suitability of LLMs for legal interpretation. The experiments establish an informed evaluation method along two axes: robustness and human correlation. In addition to raising practical concerns regarding LLM legal interpretation, our work addresses some of the theoretical considerations at the heart of the debate. LLM development continues to innovate across data, architecture, training, and inference, keeping pace with an ever-growing range of uses. However, for at least some proponents, the excitement around LLMs for legal interpretation is not due to the most recent innovations in LLM system design but rather to the assumption that large-scale data (seen in pretraining) endows the model with sophisticated language ability and the capacity to reflect on general patterns of English usage. Our experiments focused on models with limited post-training (except for GPT-4) cast doubt on this assumption. Both base and instruction-tuned models struggle to reach satisfying levels of robustness and accuracy when it comes to ordinary meaning analysis. Our results thus suggest that large-scale pretraining is not enough to produce models that are useful for the task at hand.

Our results furthermore show that LLMs are strongly affected by training choices (e.g. instruction tuning significantly shifts judgment probability distribution) and size (e.g. degree of correlation increases with model size). Even an extensively engineered model (GPT-4) that has undergone pretraining at commercial scale can fail to be robust and show limited correlation to human judgments.

What does this mean for the use of LLM legal interpretation on the bench? Our results and findings strongly caution against Judge Newsom's approach of directly asking an LLM to settle a question of 'ordinary' linguistic meaning. Although ChatGPT and other widely available systems boast fluent and easy-to-query chatbots, we cannot assume that their underlying LLMs give sound answers to such questions. This is not to say, however, that judges could not use LLMs productively in other ways for language interpretation, especially in concert with other empirical or analytical tools (e.g., Waldon et al. [75] offer a proposal for 'dialectical AI' that does not rely on assuming the system has good judgment). More nuanced approaches should be the focus of future work.

Our experimental coverage of LLMs is, of course, not exhaustive; perhaps a newer model or approach (e.g., involving a bespoke system or user-facing design) will mitigate some of these limitations. Any such suitability for this task should be demonstrated in rigorous experiments beyond those laid out in this paper.

But the evidence thus far spectacularly fails to meet the burden of proof.

Acknowledgments

This research was supported in part by NSF award IIS-2144881. The experimental portion of this work was made possible by the Georgetown University High Performance Computing Cluster, Calcul Québec, and the Digital Research Alliance of Canada. We thank Kevin Tobia, Ethan Wilcox, and Amir Zeldes, Wisdom Obinna, Mohammed Ahmed, members of the NERT lab, anonymous reviewers, and attendees of SOLID (Symposium on Legal Interpretation and Data) 2026 for helpful discussions and feedback that informed this work.

Generative AI usage statement

We utilized ChatGPT [45], GitHub Copilot, and AI Assistant in Pycharm during the implementation of our experiments.

References

[1] Samuel Alito. 2024. Majority and Dissenting Opinions in *CAMPOS-CHAVES v. GARLAND*. https://www.supremecourt.gov/opinions/23pdf/22-674_bq7d.pdf

[2] Fatemah Yousef Almeman, Steven Schockaert, and Luis Espinosa Anke. 2024. WordNet under Scrutiny: Dictionary Examples in the Era of Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 17683–17695. https://aclanthology.org/2024.lrec-main.1538/

[3] Anthropic. 2020. Claude 3.5 Sonnet Model Card Addendum. https://api.semanticscholar.org/CorpusID:270667923

[4] Tatsuya Aoyama and Ethan Wilcox. 2025. Language Models Grow Less Humanlike beyond Phase Transition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 24938–24958. https://doi.org/10.18653/v1/2025.acl-long.1214

[5] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *ACM Comput. Surv.* 58, 6 (Dec. 2025), 163:1–163:37. https://doi.org/10.1145/3777009

[6] Marco Basile. 2024. Ordinary Meaning and Plain Meaning. *Va. L. Rev.* 110 (2024), 135.

[7] Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2025. Large Linguistic Models: Investigating LLMs' Metalinguistic Abilities. *IEEE Transactions on Artificial Intelligence* 6, 12 (2025), 3453–3467. https://doi.org/10.1109/TAI.2025.3575745

[8] Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. ELQA: A Corpus of Metalinguistic Questions and Answers about English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2031–2047. https://doi.org/10.18653/v1/2023.acl-long.113

[9] Andrew Blair-Stanek and Benjamin Van Durme. 2026. LLMs Provide Unstable Answers to Legal Questions. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (ICAIL '25)*. Association for Computing Machinery, New York, NY, USA, 425–429. https://doi.org/10.1145/3769126.3769245

[10] Piotr Bystranowski, Ivar Hannikainen, Guilherme Almeida, and Kevin Tobia. 2025. *Statutory Interpretation's Empirical Turn: Empirical Contributions to Cases, Doctrine, and Theory.*

[11] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[12] Jiali Cheng and Hadi Amiri. 2025. Linguistic Blind Spots of Large Language Models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Tatsuki Kuribayashi, Giulia Rambelli, Ece Takmaz, Philipp Wicke, Jixing Li, and Byung-Doh Oh (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, USA, 1–17. https://doi.org/10.18653/v1/2025.cmcl-1.3

[13] Jonathan H. Choi. 2025. Off-the-Shelf Large Language Models Are Unreliable Judges. https://doi.org/10.2139/ssrn.5188865 social science research network:5188865 August 2025 version.

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53. http://jmlr.org/papers/v25/23-0870.html

[15] Christoph Engel and Richard H. McAdams. 2024. Asking GPT for the Ordinary Meaning of Statutory Terms. *University of Illinois Journal of Law, Technology & Policy* 2024, 2 (July 2024), 235–296.

[16] Sara Farshchi, Annika Andersson, Joost van de Weijer, and Carita Paradis. 2021. Processing sentences with sentential and prefixal negation: an event-related potential study. *Language, Cognition and Neuroscience* 36, 1 (2021), 84–98. https://doi.org/10.1080/23273798.2020.1781214

[17] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist Approaches to Natural Language Processing: A Survey. *Language Resources and Evaluation* 59, 2 (June 2025), 1719–1746. https://doi.org/10.1007/s10579-024-09766-4

[18] Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8596–8615. https://doi.org/10.18653/v1/2023.emnlp-main.531

[19] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, Online, 70–76. https://doi.org/10.18653/v1/2020.acl-demos.10

[20] Neil Gorsuch. 2020. Majority and Dissenting Opinions in *BOSTOCK v. CLAYTON COUNTY, GEORGIA*. https://www.supremecourt.gov/opinions/19pdf/17-1618_hfci.pdf

[21] Neil Gorsuch. 2025. Majority and Dissenting Opinions in *BONDI v. VANDERSTOK*. https://www.supremecourt.gov/opinions/24pdf/23-852_o7jp.pdf

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi

Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[23] Stefan Gries and Brian Slocum. 2017. Ordinary Meaning and Corpus Linguistics. *BYU Law Review* 2017, 6 (Aug. 2017), 1417–1471. https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/7

[24] James Grimmelmann, Benjamin Sobel, and David Stein. 2025. Generative Misinterpretation. *Harvard Journal on Legislation* 63 (2025), 229–308. https://journals.law.harvard.edu/jol/2026/01/24/generative-misinterpretation/

[25] Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15789–15809. https://doi.org/10.18653/v1/2024.acl-long.841

[26] David A Hoffman and Yonathan Arbel. 2024. Generative interpretation. *New York University Law Review* Volume 99 (2024), 451. Issue 2. https://nyulawreview.org/issues/volume-99-number-2/generative-interpretation/

[27] Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 2239–2259. https://doi.org/10.18653/v1/2025.findings-emnlp.121

[28] Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5040–5060. https://doi.org/10.18653/v1/2023.emnlp-main.306

[29] Jerry Huang, Peng Lu, and Qiuhao Zeng. 2025. Calibrated Language Models and How to Find Them with Label Smoothing. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, Vancouver, Canada, 25593–25612. https://proceedings.mlr.press/v267/huang25w.html

[30] Daniel Jurafsky and James H. Martin. 2026. Speech and Language Processing (3rd ed. draft), Chapter 11: Information Retrieval and Retrieval-Augmented Generation. https://web.stanford.edu/~jurafsky/slp3/11.pdf. Retrieved March 25, 2026.

[31] Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. 2021. Mistral–a journey towards reproducible language model training. https://crfm.stanford.edu/2021/08/26/mistral.html

[32] Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (Eds.). Association for Computational Linguistics, Miami, Florida, US, 263–277. https://doi.org/10.18653/v1/2024.blackboxnlp-1.18

[33] Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science* 47, 11 (2023), e13386. https://onlinelibrary.wiley.com/doi/10.1111/cogs.13386

[34] Anita S Krishnakumar. 2024. Textualism in Practice. *Duke Law Journal* 74, 3 (2024), 573–679. https://scholarship.law.duke.edu/dlj/vol74/iss3/1/

[35] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (March 1951), 79–86. https://doi.org/10.1214/aoms/1177729694

[36] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*. Association for Computing Machinery, New York, NY, USA, 611–626. https://doi.org/10.1145/3600006.3613165

[37] Nathan Lambert. 2026. *Reinforcement Learning from Human Feedback*. Online, Chapter 2: Key Related Works, 18–20. https://rlhfbook.com

[38] Thomas R Lee and Jesse Egbert. 2024. Artificial Meaning? *Florida Law Review* 77 (2024), 24–26. https://scholarship.law.ufl.edu/flr/vol77/iss6/10

[39] Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating Legal Interpretation with LLMs: Retrieval, Generation, and Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 4015–4047. https://doi.org/10.18653/v1/2025.acl-long.204

[40] Eric Martínez. 2025. Traditional and Computational Canons. https://doi.org/10.2139/ssrn.5155444 social science research network:5155444

[41] R. Thomas McCoy and Thomas L. Griffiths. 2025. Modeling Rapid Language Learning by Distilling Bayesian Priors into Artificial Neural Networks. *Nature Communications* 16, 1 (May 2025), 4676. https://doi.org/10.1038/s41467-025-59957-y

[42] Kevin Newsom. 2024. Concurring opinion in *Snell v. United Specialty Insurance Co.* , 56 pages. https://media.ca11.uscourts.gov/opinions/pub/files/202212581.pdf

[43] Kevin Newsom. 2024. Concurring opinion in *United States v. Deleon.* , 56 pages. https://media.ca11.uscourts.gov/opinions/pub/files/202310478.pdf

[44] Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2644–2663. https://doi.org/10.18653/v1/2024.eacl-long.162

[45] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee,

Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs.CL]

[46] Ferdinand Österreicher and Igor Vajda. 2003. A New Class of Metric Divergences on Probability Spaces and Its Applicability in Statistics. *Annals of the Institute of Statistical Mathematics* 55, 3 (Sept. 2003), 639–653. https://doi.org/10.1007/BF02517812

[47] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. 35 (2022), 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[48] Dawson Petersen, Abhishek Purushothama, and Nathan Schneider. 2026. Sense and Sensitivity: "Reasoning" Models are More Robust, but can Diverge from Human Consensus in a Legal Interpretation Task. In *Proc. of CoNLL*. San Diego, California.

[49] Pouya Pezeshkpour and Estevam Hruschka. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2006–2017. https://doi.org/10.18653/v1/2024.findings-naacl.130

[50] Michael Proudfoot and Alan Robert Lacey. 2009. *The Routledge Dictionary of Philosophy*. Routledge. https://www.routledge.com/The-Routledge-Dictionary-of-Philosophy/Lacey-Proudfoot/p/book/9780415356442

[51] Dasha Pruss and Jessie Allen. 2025. Against AI jurisprudence: Large Language Models and the False Promises of Empirical Judging. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 3 (Oct. 2025), 2055–2066. https://doi.org/10.1609/aies.v8i3.36695

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[53] Chief Justice John G. Roberts. 2024. Oral arguments for 19-511 *Facebook, Inc. v. Duguid*. https://www.supremecourt.gov/oral_arguments/argument_transcripts/2020/19-511_l537.pdf

[54] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15295–15311. https://doi.org/10.18653/v1/2024.acl-long.816

[55] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722

[56] Antonin Scalia. 2008. Majority and Dissenting Opinions in *DISTRICT OF COLUMBIA v. HELLER*. https://supreme.justia.com/cases/federal/us/554/570/

[57] Niels O. Schiller, Lars van Lenteren, Jurriaan Witteman, Kim Ouwehand, Guido P. H. Band, and Arie Verhagen. 2017. Solving the problem of double negation is not impossible: electrophysiological evidence for the cohesive function of sentential negation. *Language, Cognition and Neuroscience* 32, 2 (2017), 147–157. https://doi.org/10.1080/23273798.2016.1236977

[58] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. Vienna, Austria, 29 pages. https://openreview.net/forum?id=RIu5lyNXjT

[59] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*. Vienna, Austria, 35. https://openreview.net/forum?id=tvhaxkMKAn

[60] Mark A. Sherman. 1976. Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior* 15, 2 (1976), 143–157. https://doi.org/10.1016/0022-5371(76)90015-3

[61] Brian G. Slocum. 2015. *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation*. University of Chicago Press, Chicago, IL. https://doi.org/10.7208/chicago/9780226304991.001.0001

[62] Lawrence M Solan and Tammy Gales. 2017. Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review* 2017, 6 (2017), 1311–1357. https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/5

[63] Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language Models Fail to Introspect About Their Knowledge of Language. In *Second Conference on Language Modeling*. Montreal, Canada, 23. https://openreview.net/forum?id=AivRDOFi5H

[64] Neha Srikanth, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. No Questions are Stupid, but some are Poorly Posed: Understanding Poorly-Posed Information-Seeking Questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 3182–3199. https://doi.org/10.18653/v1/2025.acl-long.160

[65] Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2023. Task Ambiguity in Humans and Language Models. In *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda. https://openreview.net/forum?id=QrnDe_9ZFd8

[66] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] https://arxiv.org/abs/2403.08295

[67] Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. 2024. I am a Strange Dataset: Metalinguistic Tests for Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8888–8907. https://doi.org/10.18653/v1/2024.acl-long.482

[68] Kevin Tobia. 2024. New Methods in Statutory Interpretation: Surveys, Corpus Linguistics, ChatGPT. https://doi.org/10.2139/ssrn.4932610 social science research network:4932610

[69] Kevin Tobia, Brian G Slocum, and Victoria Nourse. 2022. Ordinary Meaning and Ordinary People. *University of Pennsylvania Law Review* 171 (2022), 365. https://repository.law.upenn.edu/Documents/Detail/ordinary-meaning-and-ordinary-people/156683

[70] Kevin P Tobia. 2020. Testing ordinary meaning. *Harvard Law Review* 134 (2020), 726. https://harvardlawreview.org/wp-content/uploads/2020/11/134-Harv.-L.-Rev.-726.pdf

[71] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, Alexis Palmer and Jose Camacho-collados (Eds.). Association for Computational Linguistics, Toronto, Canada, 101–114. https://doi.org/10.18653/v1/2023.starsem-1.10

[72] Neeraj Varshney, Satyam Raj, Venkatesh Mishra, Agneet Chatterjee, Amir Saeidi, Ritika Sarkar, and Chitta Baral. 2025. Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, Trista Cao, Anubrata Das, Tharindu Kumarage, Yixin Wan, Satyapriya Krishna, Ninareh Mehrabi, Jwala Dhamala, Anil Ramakrishna, Aram Galystan, Anoop Kumar, Rahul Gupta, and Kai-Wei Chang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 580–598. https://doi.org/10.18653/v1/2025.trustnlp-main.37

[73] Brandon Waldon, Madigan Brodsky, Megan Ma, and Judith Degen. 2023. Predicting Consensus in Legal Document Interpretation. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45, 45 (2023). https://escholarship.org/uc/item/8rq5012j

[74] Brandon Waldon, Cleo Condoravdi, James Pustejovsky, Nathan Schneider, and Kevin Tobia. 2025. Reading law with linguistics: the statutory interpretation of artifact nouns. *Harvard Journal on Legislation* 62, 2 (June 2025), 415–467. https://journals.law.harvard.edu/

jol/2025/06/01/tobia-linguistics/
[75] Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. 2025. Large language models for legal interpretation? Don't take their word for it. *Georgetown Law Journal* 114, 1 (Nov. 2025), 115–183. https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2026/02/Waldon_Schneider_Wilcox_Zeldes_Tobia_Large-Language-Models-for-Legal-Interpretation-Dont-Take-Their-Word-for-It.pdf
[76] Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024. Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. In *First Conference on Language Modeling*. Philadelphia, 23 pages. https://openreview.net/forum?id=qHdSA85GyZ
[77] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 8 (2020), 377–392. https://doi.org/10.1162/tacl_a_00321
[78] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641. https://doi.org/10.1162/tacl_a_00290
[79] Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=eskQMcIbMS Survey Certification.
[80] Sewall Wright. 1921. Correlation and causation. *Journal of agricultural research* 20, 7 (1921), 557.
[81] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1950–1976. https://doi.org/10.18653/v1/2024.findings-emnlp.108
[82] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL] https://arxiv.org/abs/1909.08593

A Prevelance of ordinary and plain meaning in US Judicial System

To highlight the prevalence and importance of ordinary (and plain) meaning analysis, we provide quick reference to some well known and important cases across the legal spectrum below.

- **Criminal law:** The term 'physically restrained' in a sentencing provision was interpreted in *U.S. v. Deleon* [43], with one of the judges exploring the use of AI (as noted above).
- **Civil rights law:** The phrase 'because of sex' in a federal statute was interpreted to protect against discrimination on the basis of sexual orientation and gender identity in the U.S. Supreme Court case *Bostock v. Clayton County* [20].
- **Regulatory law:** The U.S. Supreme Court upheld an agency's interpretation of the term 'firearm' as encompassing weapons parts kits in *Bondi v. VanDerStok* [21] (see also [74]).
- **Constitutional law:** *District of Columbia v Heller* [56] concerning 2nd amendment rights,
- **Immigration law:** *Campos-Chaves v. Garland* [1] concerned the circumstances under which a judge could rescind an order to deport a noncitizen. Parties in the case disagreed about the interpretation of a condition that was formulated as a negative disjunction (*not A or B*).

B Additional details on experimental materials

Waldon et al. [73] collected human judgments for interpretive questions about insurance contract vignettes. Here we detail the structure and topics of these vignettes. Figure 5 illustrates the kind of text seen by human subjects in their study:

Each vignette consists of a hypothetical contract provision with a term and definition, coupled with a hypothetical scenario meant to test the interpretation of that definition (table 1). The provisions are drawn from a range of insurance types (table 8). The term at issue, or *locus of uncertainty*, depends on the scenario; the full list of these terms appears in table 9. We note that these vignettes were artificially constructed, but are meant to imitate real-world insurance contract scenarios.

Steve's car insurance policy includes coverage for "Vehicle Damage," defined as "loss or damage to the policy holder's 1) car; or 2) car accessories (while in or on the car)."

> One day, Steve is involved in a minor accident. His GPS navigation system, which was in the car at the time, was damaged. Steve files a claim with his insurance company for the damage.

(1) Do you think that the claim is covered under Vehicle Damage as it appears in the policy? [Yes / No / Can't Decide]
(2) You are one of 100 people who have volunteered to answer these questions. How many of the 100 do you think will agree with your answer to question (1)?
(3) How confident are you in your answer to question (1)? [(Not at all / Slightly / Moderately / Very / Totally) confident]

Fig. 5. An example vignette from the questionnaire provided to the participants by Waldon et al. [73]. The vignette corresponds to one of the 138 items. Since our study focuses on interpretative judgment, question 1 is of interest to us, and responses make up the human judgments used in correlation analysis in Section 4.3.

Insurance Types		
Emergency Damages	Hot Work	Public Liability Property Damages
Escape of Oil	House Removal	Storm Damage
Escape of Water	Identity Theft	Trace and Access
Fire	Loss and Accidental Damage	Vehicle Damage
Flooding	Loss or Damage to a Goods Carrying Vehicle	Vehicle Fire
Garden Plants	Malicious Acts or Vandalism	Vehicle Glass
General Damages	Personal Accident	Vehicle Theft
Ground Heave	Personal Accidents	Wind Damage
Hail Damage		

Table 8. The unique insurance types represented in Waldon et al.'s experimental materials.

Human Data. In Waldon et al. [73], the subjects were asked three questions (see fig. 5), first asking for judgment, second asking for confidence, third asking for a estimation of consensus. For this work, the first and second question would be of interest. However, the second question collects confidence on a discrete likert scale, and hence we do not use it to scale the discrete judgment provided by the subjects.

C Data contamination

Since our source data is from 2023, it is possible that it was part of the training for one or more models in our selected suite of LLMs, especially the larger models which show the best correlation. Table 10 catalogs the known cutoff dates for for the models. Any model with the exception of GPT-2 may have been partially or wholly exposed to the dataset.

Loci of Uncertainty		
accessory	flammable or combustible materials	permanent or total loss
accidental	flow of water	political disturbance
audio equipment	glass	professional movers
broken glass	ground heave	rapid build-up
causative "from"	hard surface	reasonable steps
cause	heating installation	regular working conditions
connected with business	internal source	requires / uses / produces
connected with occupation	key theft	sudden/unforeseen
custody or control	leaking	temporarily removed
damage	malicious people or vandals	third party
deliberate	naturally occurring fire	tracking device
family or employee	necessary and reasonable	traveling in
fire damage	outside the building	wear and tear
first responder	perceived emergency	

Table 9. Loci of uncertainty in contract provisions. (Note that these are descriptions in the dataset; the actual text prompts in Waldon et al. [73] did not contain metalinguistic terms like *causative*.)

Model	**Reported Cutoff Date**	**Data available before cutoff**?
GPT-4	Dec 1, 2023	Yes
GPT-2	Unknown (Released November 2019)	No
Llama-3	December 2023	Yes
OLMo-2	December 2023	Yes
Ministral-8B	Unknown (Released October 2024)	Maybe
Gemma-3	August 2024	Yes

Table 10. Reported cutoff date for training data or if unknown, model release date for each model. It is possible that our source evaluation data was included in the training dataset for 14 out of 15 models we use for our study.

D Alignment between first token probability and decoded output

In section 3 we described our operationalization of LLM judgment using first output token probability. However, this is a simplifying assumption: it is possible that (the probability of) the first token alone misrepresents the full LLM response. We manually reviewed all outputs of `Llama-70B` to better understand the effects of our choice to represent LLM judgment by first token probability. table 11 organizes our analysis. This decision is overall beneficial, allowing us to extract judgment from ambiguous or incoherent responses while extracting an equivalent human-extractable answer in 1,217 out of 1,242 (97.3%) responses. Our setup is erroneous in 10 responses (0.8%), yet enables judgment extraction in 23 non-human-extractable responses due to incoherence or ambiguity (1.9%).

A manual sweep also uncovered other examples of incongruence, specifically in the `OptionsFlipped` and `Options` question variants (challenges with extracting multiple choice answers are relatively well studied; Röttger et al. [54], Wang et al. [76]). In `Llama-70B` responses, a total of 14 could be considered incongruent, providing answers in plain text rather than selecting one of the `A.` or `B.` options. In 8 of them, first token probability aligns with manually extractable judgment, e.g. responding with "Denise is covered. ..." with $p(\textsc{Covered}) >$

Manual extraction yields	Count	Example
Correct extraction	1,217	
No extractable answer	18	yes or no? Peter has insurance that covers "Loss or Damage to a Goods Carrying Vehicle," which includes "key replacement in the case of theft." As part of his bakery business, Peter owns a van which he uses to make local deliveries. One day, Peter realizes he has lost the key to the van.
Ambiguous answer	5	yes or no? I think the answer is no, because the damage was not sudden. It took several hours for the damage to occur. The damage was not unforeseen, because Carol knew about the leak and did not fix it immediately. I think the answer is yes, because the damage was sudden. It happened in
Incorrect extraction	10	yes or no. The answer is "no." The insurance policy covers damage to belongings that occurs while being stored by professional removal contractors. Bill is not a professional removal contractor, so the damage to Harper's belongings is not covered by the insurance policy

Table 11. Counts and examples of correct, no, ambiguous, and incorrect extractions in responses generated by `Llama-70B Base`. Examples are all responses to the `Yes/No` prompt.

p(NotCovered). In the remaining 6, first token probability does not align, e.g. responding with "Gavin is not covered" but also with p(Covered) > p(NotCovered).

Model	3	4	5	Model	3	4	5
Llama-70B	22	21	95	Ministral-8B-Inst	16	67	55
+Inst	24	93	21	OLMo-2-7B	71	26	41
Llama-8B	33	43	62	+Inst	36	77	25
+Inst	13	98	27	Gemma-7b	45	50	43
Llama-3B	11	113	14	+it	9	47	82
+Inst	57	48	33	GPT-2-medium	67	71	0
Llama-1B	4	25	109	GPT-4	5	41	92
+Inst	13	125	0	**All**	**426**	**945**	**699**

Table 12. Number of items by number of question variants that yielded the majority judgment for the model. For example, there were 5 items for which GPT-4 produced one judgment for 3 variants, and the opposite judgment for 2 variants. Each judgment is a binary choice between Covered and NotCovered. This is a replication of table 5 but without the four most minority response-inducing variants `Disagreement`, `AgreementWithNegation`, `DisagrWithNegation` and `Options`.

E Implementation and Compute

E.1 Load-and-infer pipeline.

We use `vllm` [36] to implement our inference pipeline and use the model implementations available on https://huggingface.co/models.

All our inference was completed on Tesla L4 GPUs with 24GB of memory, with the exception of `Llama-70B`, which were run on 4xA100, each with 40GB of memory. Our inference configuration will be available as part of our public code.

Model	Variant	Mean	Std	Model	Variant	Mean	Std
Llama-70B	Options	0.09	0.04	OLMo-2-7B	Disagr. w/ Neg.	0.37	0.05
+Inst	Negation	0.34	0.17	+Inst	Agr. w/ Neg.	0.48	0.23
Llama-8B	Options F.	0.15	0.04	Ministral-8B-Inst	Options	0.16	0.04
+Inst	Disagr. w/ Neg.	0.18	0.06	Gemma-7B	Disagr. w/ Neg.	0.08	0.03
Llama-3B	Agr. w/ Neg.	0.19	0.06	+it	Options	0.78	0.04
+Inst	Options F.	0.22	0.07	GPT-2-medium	Disagr. w/ Neg.	0.17	0.02
Llama-1B	Options	0.28	0.05	GPT-4	Disagr. w/ Neg.	0.56	0.30
+Inst	Options F.	0.32	0.04				

Table 13. The question variant for each model with the largest Jensen-Shannon distance from the `Yes/No` question. Higher distance indicates greater difference between probability distributions.

Judgment as a sum of probability. In section 3, we consider LLM judgment as a sum of token probabilities that correspond to each judgment. That is, when answering yes to the question would indicate the Covered judgment, we consider $p(\text{Covered}) = p(\text{Yes}) + p(\text{yes}) + p(\text{YES})$.

Random seed and temperature. Because a significant portion of our study works with token probabilities, we set temperature to 0, and hence there is no randomness in our inference pipeline.

E.2 OpenAI GPT-4 Inference with APIs

We used OpenAI API platform for inference with GPT-4. We use `temperature=0` to get the highest determinism. We use the GPT-4 model with the model identifier `gpt-4-0613`.

F Additional details on LLM response across model, question variants

In section 4, we discuss LLMs' sensitivity to surface form and the lack of strong, consistent correlation to human judgment. We provide additional detail on our analyses: correlation between human and `GPT-4` judgment across all prompt types in fig. 6 and correlation between human and LLM judgment for each model's prompt type with the highest R^2 value in fig. 7. The additional figures echo our findings. Figure 6 illustrates sensitivity to trivial variation in input and response incongruence even in the model with the highest correlation to human judgment, `GPT-4`. Figure 7 provides a detailed visualization of each model-prompt pair described in table 7.

G Linking hypothesis

In section 4.3, we assume that our respective operationalizations of human and LLM judgment, namely the proportion of covered judgments %Covered in human responses and the probability difference between LLM covered and not covered judgments Δ, have a linear relationship. While the assumption is difficult to justify, we provide our attempt.

By definition, token probability $p(w \mid C) = f(C)$ in autoregressive language modeling [52] represents the proportion of cases where the next token w occurs given a model f and number of environments with context C. In our implementation, we consider this as a computational analogue of querying a human population and calculating the proportion of which that respond with one judgment, or, the proportion of human judgments in human responses. This is the basis of our assumption that judgment probability as sum of token probabilities $p(\text{Covered}) = p(\text{Yes}) + p(\text{yes}) + p(\text{YES})$ and proportion of covered judgments have a linear relationship.

However, due to residual token probabilities that linger in LLM probability distributions, we are unable to represent LLM judgment with a single token probability, as a low $p(\text{Covered})$ does not indicate a high

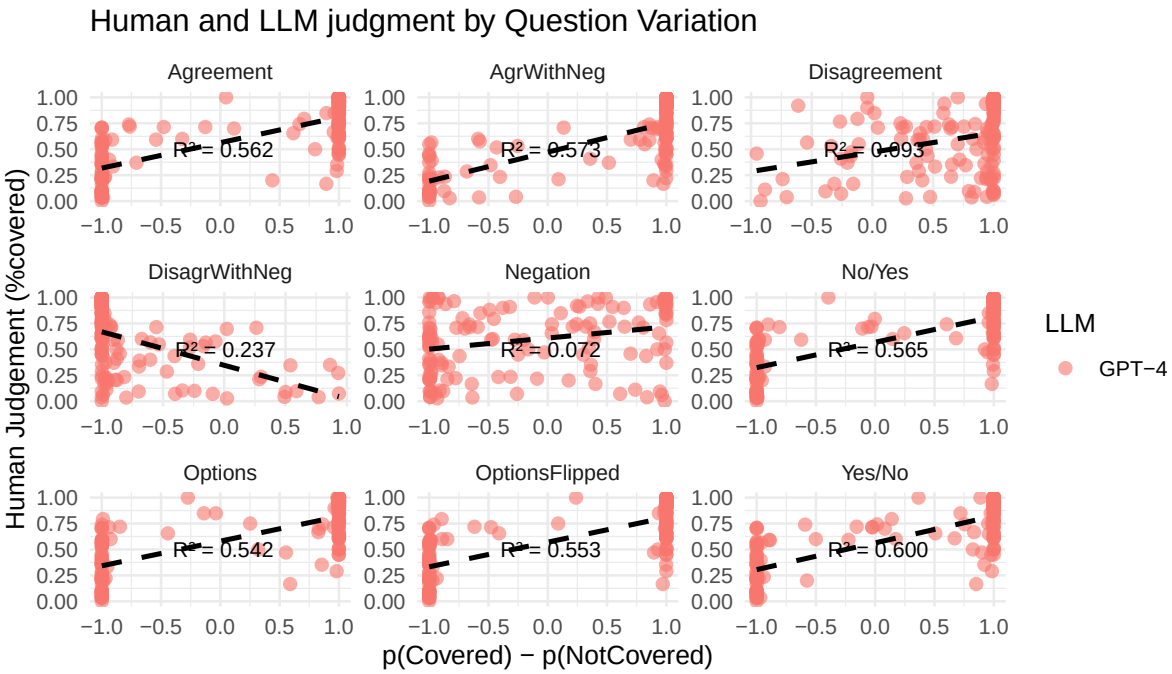

Fig. 6. GPT-4 judgment probabilities versus human consensus across question variants. Dotted lines are best best-fit lines between human and instruction-tuned LLM.

p(NotCovered). We thus take the difference of the two probabilities, $\Delta = p(\text{Covered}) - p(\text{NotCovered})$ to represent LLM judgment that has a linear relationship with human majority judgment. This allows for clear analysis, as the expectation value of a judgment $E(\Delta) = 0$, and $\Delta > 0$ yields the Covered judgment and $\Delta < 0$ the NotCovered judgment.

Based on our linking hypothesis, consider the ideal case where our two variables are perfectly correlated to each other with $R^2 = 1$. In such case where $\Delta \in [-1, 1]$ and $p(\text{Covered}) \in [0, 1]$, we predict that the proportion of human covered judgments is 0 when probability difference is -1 since $p(\text{Covered}) = 0$, $p(\text{NotCovered}) = 1$. It follows that proportion of human covered judgments is 1 when probability difference is 1; and 0.5 when probability difference 0. The best fit line would then have $m = 0.5$, $b = 0.5$. Here, all of the variance across human judgment is explained by Δ.

Other metrics may be used to perform the same experiments, e.g. with normalization with the sum of relevant judgment tokens, disregarding residual token probabilities p(other).

We outline R^2 measurements across 4 different formulas to represent LLM judgment in table 14. The four formulations differ, but as can be seen in the table, this does not affect the correlation values.
Let us define $\Delta = p(\text{Covered}) - p(\text{NotCovered})$ and $\Sigma = p(\text{Covered}) + p(\text{NotCovered})$. Then, R^2_{diff} is with judgment as difference Δ, R^2_{covered} is with judgment as just the covered judgment $p(\text{Covered})$, R^2_{rel} is with judgment as relative quantity $p(\text{Covered})/\Sigma$, and R^2_{norm} is with judgment as normalized difference Δ/Σ.

We observe that the change in operationalization results in no significant change in R^2 values. In fig. 7 you can see the lack of effect of the scale used for the judgment and the R^2. For contrast, between `Llama-8B-Inst` and `OLMo`. They have different scales, but similar correlations.

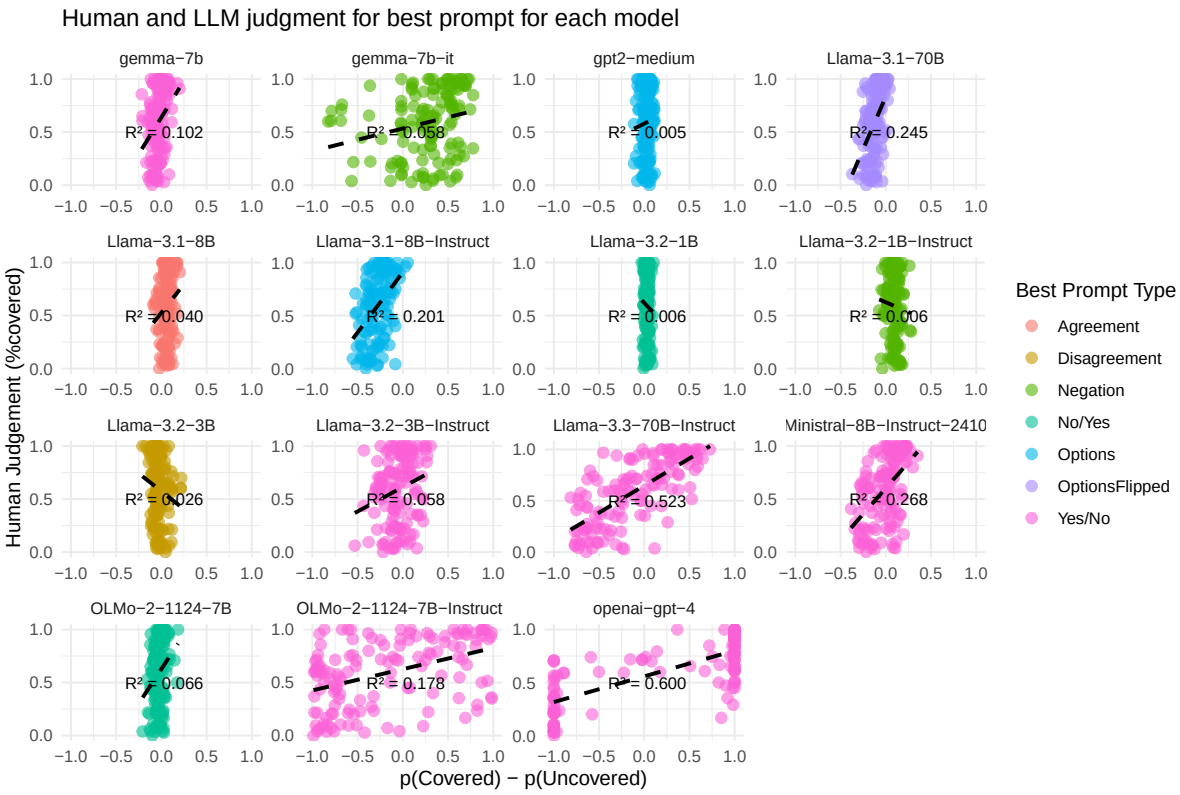

Fig. 7. Each model's judgment probability versus human consensus with its prompt type reporting strongest correlation to human consensus. Dotted lines and the corresponding R^2 are best-fit lines.

Model	R^2_{diff}	R^2_{cov}	R^2_{rel}	R^2_{norm}	Model	R^2_{diff}	R^2_{cov}	R^2_{rel}	R^2_{norm}
Llama-70B	0.25	0.21	0.26	0.26	Ministral	0.27	0.25	0.27	0.27
+Inst	0.52	0.50	0.53	0.53	OLMo-7B	0.07	0.01	0.06	0.06
Llama-8B	0.04	0.02	0.04	0.04	+Inst	0.18	0.17	0.18	0.18
+Inst	0.20	0.19	0.20	0.20	gemma-7b	0.09	0.07	0.11	0.11
Llama-3B	0.03	0.02	0.03	0.03	+it	0.06	0.06	0.06	0.06
+Inst	0.06	0.06	0.06	0.06	GPT-4	0.60	0.60	0.60	0.60
Llama-1B	0.01	0.00	0.01	0.01	GPT-2-medium	0.01	0.01	0.01	0.00
+Inst	0.01	0.00	0.00	0.00					

Table 14. R^2 values for all models and their highest human correlation. Each column represents an operationalization of LLM judgment.

H Limitations

Legal interpretation and 'ordinary meaning' are complex topics of theory and practice in the legal field. Our work only looks at a specific aspect of LLM usage, posing direct queries for ascertaining ordinary meaning with

a binary-choice QA task. This does not represent other mechanisms of using LMs for legal interpretation, such as producing arguments for and against an interpretation [75], or eliciting examples [2]. Another concern is focused on different ways of eliciting such responses; cues are necessary to get consistent answers, but such cues can also constrain and distort the output to some extent as discussed in Röttger et al. [54].

We use data from a previous study of consensus in legal interpretation. The authors of that study make no claims as to the overall representativeness of their experimental stimuli to questions that come up in day-to-day contractual interpretation. They also explicitly discuss the role of researcher subjectivity in the constructing the stimuli. We do not report any representativeness or coverage information for the data. Hence, despite the diversity of scenarios compared to other related works, it is currently unclear how representative it is of LLM use for legal interpretation in practice.

Our evaluation utilizes a small sample of 138 unique items. This is a small dataset for testing the generalization of language model judgments for the task.

We attempt to use question variants in a controlled manner to investigate how such variation affects judgment. However, the interplay between question variants and LLM judgment may not be easily disambiguated. LLMs' ability to understand and follow task instructions cannot be guaranteed, especially without targeted post-training [65].

Our models are not chosen based on their attested performance on natural language benchmarks. Additionally, none of the models evaluated are considered "reasoning" or "thinking" models.

While we attempt to include models with varying size, architecture, and training recipe, we do not present our findings to be comprehensive or conclusive. Larger and more recent models with 'reasoning', 'retrieval' and/or search methods may be able to provide stable, reliable sources of human-like legal interpretation. We also do not use "chain-of-thought" or other prompting or in-context learning methods meant to elicit or induce intermediate or reasoning steps. These have been shown to improve model performance in many tasks. Hence, our results may represent a lower estimation of model ability.

We use first token probability as the model response; this has advantages both in computing resources and analysis, but provides a limited representation of model output [28] but is effective for larger models [63]. Additionally, the first token probability has some known drawbacks [76], especially for instruction-tuned models, which we take additional steps to mitigate. Alternatively, non-instruction-tuned models are less likely to provide appropriate responses to our QA formulation.

Due to the complexity of responding to negation and interpreting it from the first token, we check the text to ascertain the correct polarity (Covered or NotCovered) to use for the judgments. However, this does not guarantee that all scenarios get captured under the polarity judgment. For distributional judgment, we collate the probabilities to only three categories, in which the tail of the probability distribution is reduced and represented as 'Other'. This approximation reduces the precision and validity of the distance metric. Our correlation analysis uses a specific transformation of model responses and compares it to the Covered proportion. We did not perform targeted validation for establishing our linking hypothesis in connecting human responses and LLM responses for legal interpretation.

We did not perform checks for data contamination [55] in the language models we have used. Data contamination of the published materials and reference policies used in the previous study [73] could have influenced the models we studied. We catalog the cutoff and release dates in Appendix C.