Sense and Sensitivity: "Reasoning" Models are More Robust, but can Diverge from Human Consensus in a Legal Interpretation Task

Dawson Petersen **Abhishek Purushothama** **Nathan Schneider**

Georgetown University

{dawson.petersen, ap2089, nathan.schneider}@georgetown.edu

Abstract

Can LLMs make metalinguistic judgments? While LLM embeddings are often regarded as high-quality semantic representations, it is not clear that prompting an LLM is a useful way to obtain metalinguistic insights (e.g., whether a DIY gun kit is a "firearm"). While some prior work has suggested LLM prompting can simulate surveys with human participants, computational studies in the domain of legal interpretation have found that LLMs are unreliable for metalinguistic judgments due to prompt sensitivity. However, these studies did not directly compare humans and LLMs on identical tasks, nor did they test so-called "reasoning" models. The current study addresses these gaps by directly comparing the robustness of human and LLM judgments (with and without `reasoning`) in an English-language legal interpretation task. Our results show that LLMs were more sensitive to irrelevant prompt features compared to human participants. Enabling `reasoning` improved the stability of LLM responses. However, even `reasoning` model outputs had only moderate correlations with human judgments, and all models sometimes output interpretations that no humans reached in response to the same prompt. We conclude that while `reasoning` decreases prompt sensitivity, LLMs are still poor proxies for human metalinguistic judgments.

1 Introduction

It is highly tempting to see querying LLMs as a faster and cheaper alternative to traditional survey methods (Cho et al., 2024). This is especially true when the goal is to draw inferences about the subject which LLMs purportedly know best—language. LLMs train on vast stores of language data, collected from books and the web, and exhibit high performance on a wide range of language tasks. However, these facts on their own do not guarantee that LLMs can provide valid proxies for human survey responses, not even in language tasks.

Previous work has shown that although LLMs encode high-quality semantic representations in the form of word embeddings, LLM-generated metalinguistic judgments often do not reflect internal model parameters (Hu and Levy, 2023). Related work has shown that LLM-generated semantic categorizations show only moderate correlations with human judgments (Heyman and Heyman, 2019, 2024; Pedrotti et al., 2025) and can fail entirely when those judgments rely on nuanced semantic knowledge (Misra et al., 2023). Indeed, even when LLMs generate plausible responses to survey questions, they tend to be prompt-sensitive (Bisbee et al., 2024; Chen, 2024; Choi, 2025), and because optimal prompting strategies can be model-dependent, it is difficult to know, a priori, which prompting strategies are likely to produce valid responses.

Although LLMs are clearly not perfect, it is worth considering whether or not they are *good enough* at these tasks to approximate human participants. After all, human participants are not always reliable. Indeed, Heyman and Heyman (2024) found that the difference between their human subjects' responses on day one and day two in a test-retest paradigm was almost as great as the difference between the human and LLM responses. Furthermore, humans exhibit certain types of "prompt sensitivity", like acquiescence bias (the tendency to agree rather than disagree with survey statements; Moss, 2008) which can harm validity. Therefore, in order to understand whether LLMs can serve as useful proxies of human metalinguistic judgments, it is necessary to directly compare the performance of humans and LLMs under the same task conditions. While some prior work has tested whether or not LLMs show humanlike response biases, such work has either compared LLM outputs to hypothetical "expected" human responses

(Tjuatja et al., 2024; Chen, 2024) or attempted to replicate previous human subjects experiments with LLMs (Purushothama et al., 2026; Engel and McAdams, 2024; Martínez, 2025). None of these studies collected original data from both humans and LLMs to compare their responses to identical prompt variations. Additionally, these studies did not test whether so-called `reasoning` methods (OpenAI et al., 2024b; Welleck et al., 2024; Snell et al., 2025) could improve model performance. The current study fills these gaps by directly comparing the robustness of human and LLM responses (with and without `reasoning`) to identical prompt manipulations (see Table 1). Specifically, we investigate these issues using an extensional judgment task for legal interpretation (i.e., the analysis of legal texts to determine their meaning, Kurzon, 2006), due to the high demand for reliable metalinguistic judgments in this context.

Our results show:

- Humans are less prompt sensitive than LLMs.
- Enabling `reasoning` decreases LLM prompt sensitivity.
- Even `reasoning` LLMs do not consistently match human judgments.

2 Background

Determining the "ordinary meaning" (Scalia and Garner, 2011) of ambiguous legal texts is a central problem for the U.S. judicial system. Courts have devised a repertoire of canons (i.e., interpretive heuristics) which are used to approximate ordinary meaning, alongside supplementary tools like dictionaries. More recently, however, the field of legal interpretation has undergone an empirical turn (Tobia, 2022) as scholars have begun to question the extent to which these traditional tools reflect ordinary meaning (Chen, 2025). While scholars have proposed a number of new empirical methods, including surveys (Tobia, 2020) and corpus linguistics (Solan and Gales, 2017; Tobia, 2021), courts have generally been reluctant to adopt them, due in part to both the financial costs and expertise required (Hoffman and Arbel, 2024).

In light of these concerns, some judges (Newsom, 2024a,b) and legal scholars (Hoffman and Arbel, 2024; Engel and McAdams, 2024; Datzov, 2025; Martínez, 2025) have proposed that LLMs might serve as a cheaper and easier way to acquire empirical data to inform interpretive questions.[1] However, this proposal has not been without controversy (Wilf-Townsend and Tobia, 2025). Critics have raised concerns about the nonrepresentativeness of training data (Waldon et al., 2025), metalinguistic access (Waldon et al., 2025), transparency (Lee and Egbert, 2024), interface design (Nielsen et al., 2025), response reliability (Chen, 2024; Choi, 2025; Purushothama et al., 2026), and the possibility that corporate engineering decisions could unduly influence court rulings.[2]

As such, it is critically important to investigate the validity of using LLMs as proxies for human judgments in such legal interpretation tasks. While all of these validity questions will eventually need to be answered before LLMs can be safely used in the courtroom, in this work, we focus specifically on the issues of response reliability and approximation of human judgments.

Because prevailing legal theories require that interpretations should align with the judgments of "ordinary speakers" (Scalia and Garner, 2011), we tested LLM interpretations against a representative U.S. sample of English speakers (Tobia, 2022) as an interpretive gold standard.

3 Methodology

In order to evaluate the comparative prompt sensitivity of humans and LLMs, we adapted the stimuli set developed by Waldon et al. (2023). The stimuli set consists of 46 paragraph-long stories about insurance claims. There are three variants of each story with different expected interpretations: covered, not covered, and controversial.

We followed each item with either a positively or negatively framed statement about whether or not the claimant was covered (e.g., "John is covered" vs. "John is *not* covered") and instructions to select "Agree" or "Disagree". We systematically varied whether each item was paired with a negative or positive statement and the order of the answers, resulting in 552 distinct item–prompt variations. The experiment followed a 3 [version: covered, controversial, uncovered] × 2 [frame: positive, negative] × 2 [order: agree first, disagree first] design. This experimental design is illustrated in Table 1.

Human subjects We recruited a representative U.S. sample of 300 participants from the online re-

[1] This usage is distinct from using LLM as automatic evaluators for NLP tasks (popularly referred to as LLM-as-a-judge; Bavaresco et al., 2025).

[2] We recommend "Judges Shouldn't Rely on AI for the Ordinary Meaning of Text" by Lawfare as an accessible reading on this topic.

Prompt		
Chris's car insurance policy includes coverage for "Vehicle Damage," defined as "loss or damage to the policy holder's 1) car; or 2) audio equipment (while in the car)." One day, Chris is involved in a minor accident.		
uncovered version	**controversial version**	**covered version**
His work laptop, which was in the car at the time, was damaged.	His portable Bluetooth speaker, which was in the car at the time, was damaged.	His car stereo system and speakers were damaged.
Chris files a claim with his insurance company for the damage. Based on this definition, select your response to the following statement:		
positive frame		**negative frame**
Chris is covered by the Vehicle Damage policy.		Chris is *not* covered by the Vehicle Damage policy.
agree first order		**disagree first order**
◯ Agree ◯ Disagree		◯ Disagree ◯ Agree

Table 1: Sample experimental item with prompt variations. For each item in the stimuli 12 (3 versions × 2 frames × 2 orders) experimental items are constructed. (For reference, an example conditioned item is included in Appendix A.)

cruitment platform Prolific.[3] The median age was forty-six. The sample was 51% female, 49% male. Ethnically, 62% of the sample self-identified as White, 12% as Black, 11% as Mixed, 6% as Asian, and the rest as Other. After indicating informed consent, participants completed a Qualtrics survey. Each participant saw six pseudorandomly selected items in a pseudorandom combination of conditions such that the number of observations across items and conditions was approximately balanced, with a median of three observations per prompt. Median completion time was 4 minutes and 22 seconds. We excluded twelve participants who finished the task in less than two minutes, resulting in a final sample size of 288. Each participant who completed the experiment was paid $1.50,[4] regardless of exclusion. This research was approved by the Georgetown University Institutional Review Board (Approval #00010299).

LLMs We performed five runs of the experiment with LLMs (Llama-3.3-70B, GPT-4.1, GPT-5.2 with `reasoning` disabled, and finally GPT-5 Mini and GPT-5.2 with `medium effort reasoning` enabled[5]). We chose models which we believed, based on prior work (Purushothama et al., 2026), to have a reasonably high chance of success and that would allow for fair comparisons between `reasoning` and non-`reasoning` approaches. Each model-generated a response to all 552 possible prompt variations.

The prompts processed by the LLMs differed only slightly from those seen by our human participants. Namely, possible answers were presented in text as "Agree or Disagree?" rather than bulleted options, and the prompt ended with the words "Final answer is:" in order to optimize the number of interpretable responses. When models successfully generated "Agree" or "Disagree" as the first word of the output, they were coded as agree/disagree based on first token. When the models failed to generate an interpretable first token, the response was coded based on later content in the output. We note these edge cases in §4. See Appendix C for additional implementation details. Model querying code is publicly available.[6]

4 Results

Responses were first coded as a *covered decision* or a *not covered decision* based on the selected answer and frame condition. Decision data were then analyzed using logistic regression in R (4.4.1, R Core Team, 2024).[7] For each sample, we first defined the maximal model below and then tested whether each term significantly improved model fit using likelihood ratio tests[8] (results in Appendix D). This allowed us to statistically test whether or not differences in coverage decisions between order and

[3] We used the "standard representative U.S. sample" option in Prolific which controls for age, sex, and ethnicity based on the U.S. Census. See "What are representative samples on Prolific" for more information.

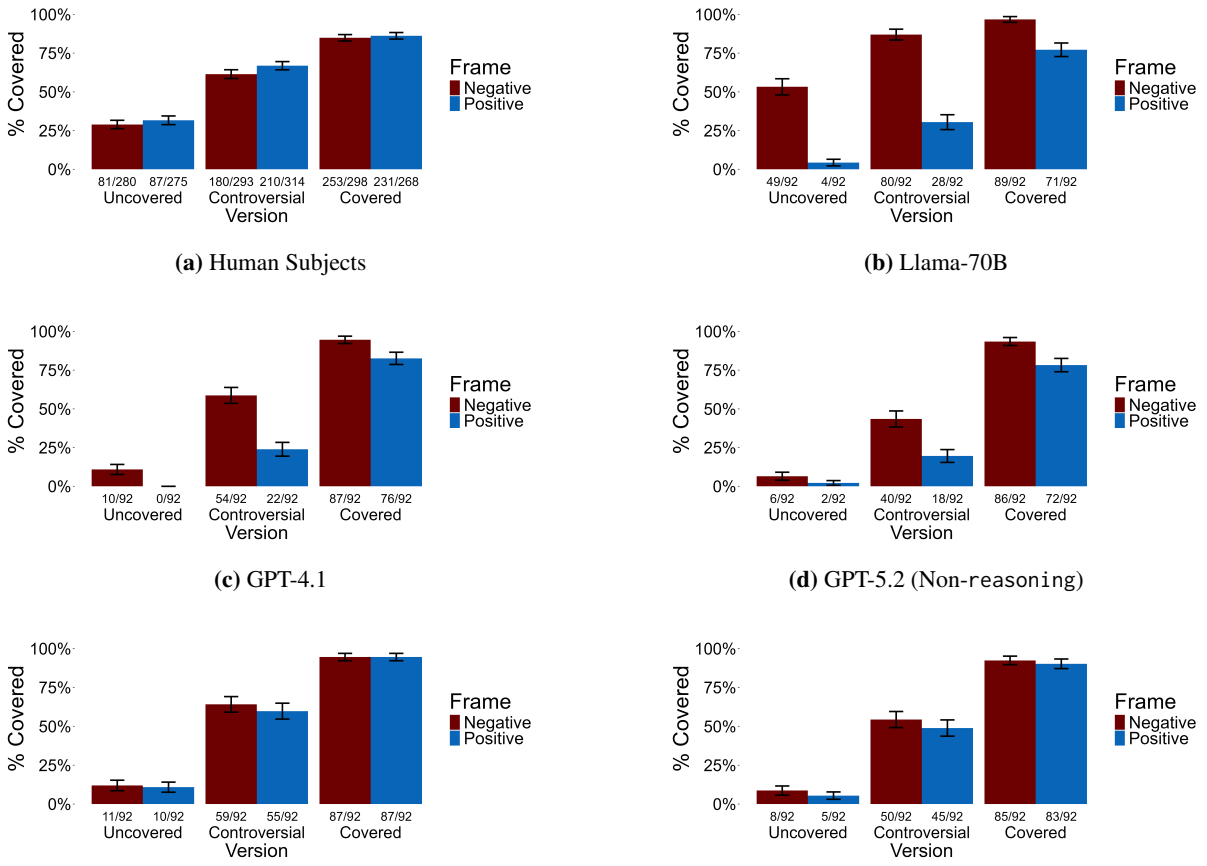
[4] Average of $20.61 per hour

[5] Appendix B contains full reference for the models.

[6] `https://github.com/Dawson-Petersen/Sense-and-Sensitivity`

[7] For an introduction to logistic regression see Hilbe (2009)

[8] For an introduction to model selection see Leeb and Pötscher (2009)



(a) Human Subjects

(b) Llama-70B

(c) GPT-4.1

(d) GPT-5.2 (Non-reasoning)

(e) GPT-5 Mini (reasoning)

(f) GPT-5.2 (reasoning)

Figure 1: Coverage determinations by frame and version. The y-axis shows the percent of trials where the covered option (i.e., agree with positive framing, disagree with negative framing) was selected for each combination of conditions. The number of covered trials over total trials is displayed at the bottom of each bar. The non-reasoning models (Llama-70B, GPT-4.1, GPT-5.2) show a framing effect in this task (i.e., a difference in % Covered between negative and positive frames), while human subjects do not. Like humans, reasoning models show no effect of frame.

framing conditions likely reflect real differences in the population rather than sample variance. We report effects in terms of odds ratio (OR, Bland and Altman, 2000) and 95% confidence intervals. Data and analysis code is publicly available.[9]

The maximal regression model was specified as:

```
glm(decision ~ version + frame + order,
    family = binomial)
```

4.1 Human Subjects

For the human subjects sample, results showed that only the main effect of version significantly improved model fit. The effect was such that participants made a covered decision less often for the uncovered version compared to the controversial (OR = .242, 95% CI [.189, .309]) and covered versions (OR = .074, 95% CI [.055, .099]) (see Figure 1a).

Overall, these results indicate that human participants' interpretations were sensitive to substantial changes in story content but not to irrelevant prompt variations due to frame (positive, negative) or ordering of the choices.

4.2 Non-reasoning Models

Llama-70B Llama-70B successfully generated "agree"/"disagree" as the first word of all 552 trials. Results showed that the main effects of version, frame, and order all significantly improved model fit.

The effect of version was such that Llama-70B generated a covered decision less often for the uncovered version compared to the controversial (OR = .285, 95% CI [.185, .439]) and covered versions

[9] https://georgetown.box.com/v/sense-and-sensitivity-analysis

(OR = .061, 95% CI [.036, .104]). However, Llama-70B was also highly sensitive to both irrelevant prompt variations (frame and order), being somewhat less likely to generate a covered response when the "agree" option was listed first (OR = .709, 95% CI [.505, .996]) and much less likely to generate a covered response to the positive frame compared to the negative frame (OR = .158, 95% CI [.108, .231]) (see Figure 1b).

Overall, while Llama-70B showed some sensitivity to substantial changes in story content, it was also highly sensitive to irrelevant prompt variations. Furthermore, these sensitivities did not follow the same patterns as previously documented human acquiescence bias (Moss, 2008). Rather, Llama-70B showed an unpredicted *dis*acquiescence bias, being more likely generate a covered response for negatively framed prompts.

GPT-4.1 GPT-4.1 successfully generated "agree"/"disagree" as the first word of only 515 out of 552 trials. The remaining 37 outputs were in the form "Final answer is: Agree/Disagree" and were coded manually. Results showed that the main effects of version and frame significantly improved model fit.

The effect of version was such that GPT-4.1 generated a covered decision less often for the uncovered version compared to the controversial (OR = .082, 95% CI [.040, .165]) and covered versions (OR = .007, 95% CI [.003, .016]). The effect of frame was such that GPT-4.1 was less likely to generate a covered response to the positive frame compared to the negative frame (OR = .456, 95% CI [.324, .642]) (see Figure 1c).

Overall, GPT 4.1 was more robust compared to Llama-70B, but was still sensitive to the irrelevant framing prompt variation. Like Llama-70B, GPT-4.1 demonstrated a disacquiescence bias, being more likely generate a covered response for negatively framed prompts.

GPT-5.2 (Non-`reasoning`) GPT 5.2 successfully generated "agree"/"disagree" as the first word of 538 out of 552 trials. The remaining 14 outputs were in the form "Final answer is: Agree/Disagree" and were coded manually. Results showed that the main effects of both version and frame significantly improved model fit.

The effect of version was such that GPT-5.2 generated a covered decision less often for the uncovered version compared to the controversial (OR = .099, 95% CI [.046, .214]) and covered versions (OR = .007, 95% CI [.003, .017]). The effect of frame was such that GPT-5.2 was less likely to generate a covered response to the positive frame compared to the negative frame (OR = .545, 95% CI [.387, .769]) (see Figure 1d).

Like the smaller models, GPT-5.2 showed sensitivity to the irrelevant framing prompt variation in the form of a disacquiescence bias.

4.3 `reasoning` Models

GPT-5 Mini (`medium effort reasoning`) GPT-5 Mini successfully generated "agree"/"disagree" as the first word of 548 out of 552 trials. The remaining four outputs were in the form "Final answer is: Agree/Disagree" and were coded manually. Results showed that only the main effect of version significantly improved model fit.

The effect of version was such that GPT-5 Mini generated a covered decision less often for the uncovered version compared to the controversial (OR = .079, 95% CI [.046, .136]) and covered versions (OR = .007, 95% CI [.003, .016]) (see Figure 1e).

Overall, with `medium effort reasoning` enabled, GPT-5 Mini was more robust compared to the non-`reasoning` models, being sensitive to substantial changes in story content but not to irrelevant prompt variations.

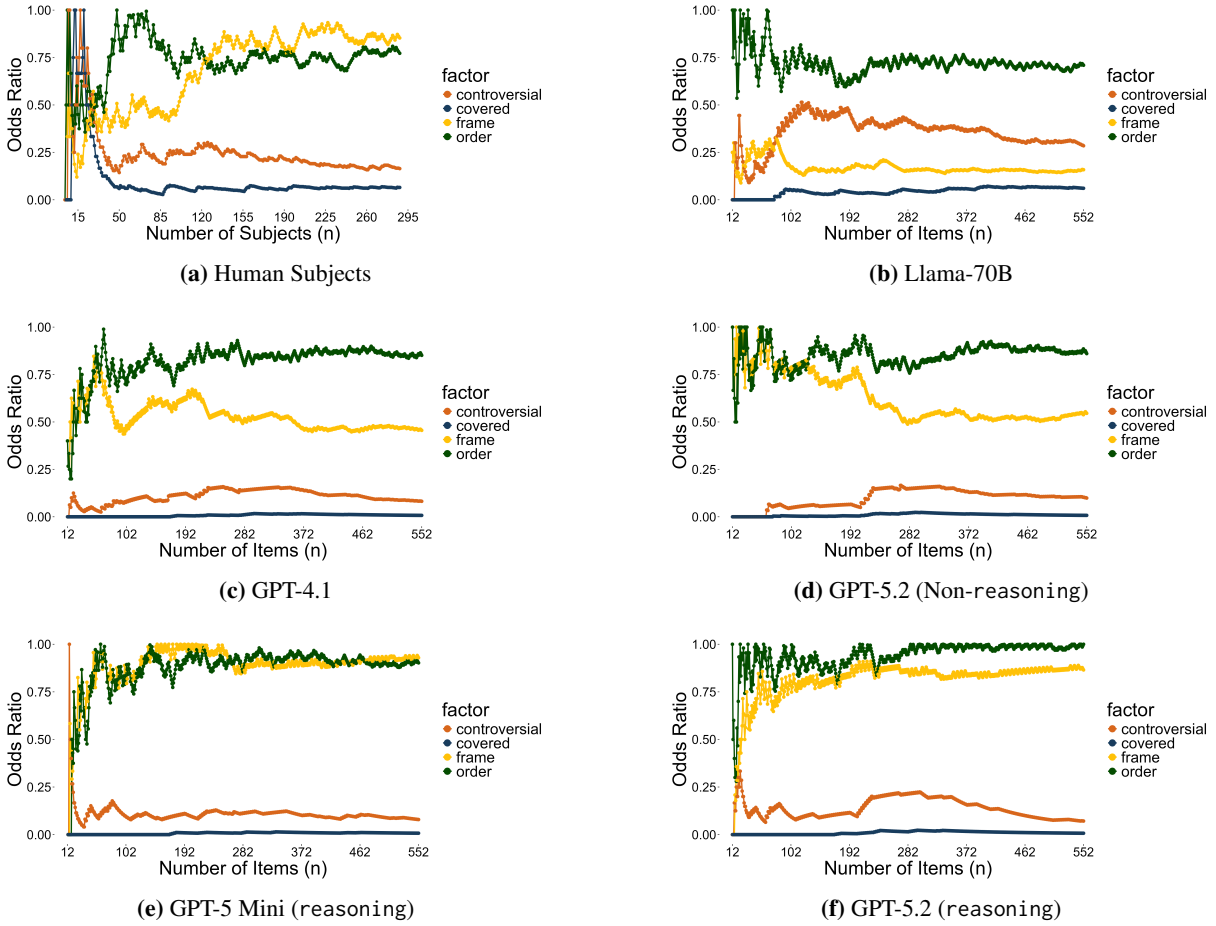
GPT-5.2 (`medium effort reasoning`) Out of 552 trials, GPT-5.2 successfully generated "agree"/"disagree" as the first word of 540 outputs. The remaining 12 outputs were in the form "Final answer is: Agree/Disagree" and were coded manually. Results showed that only the main effect of version significantly improved model fit.

The effect of version was such that GPT-5.2 generated a covered decision less often for the uncovered version compared to the controversial (OR = .071, 95% CI [.038, .134]) and covered versions (OR = .007, 95% CI [.003, .016]) (see Figure 1f).

Overall, with `medium effort reasoning` enabled, GPT-5.2 was more robust compared to the non-`reasoning` models, being sensitive to substantial changes in story content but not to irrelevant prompt variations.

4.4 Statistical Power of Effects

One potential concern with our results is that the lack of frame or order effects in our human subjects and `reasoning` models may have resulted from a lack of statistical power. In order to address this concern, we performed an effect size stabilization



(a) Human Subjects

(b) Llama-70B

(c) GPT-4.1

(d) GPT-5.2 (Non-`reasoning`)

(e) GPT-5 Mini (`reasoning`)

(f) GPT-5.2 (`reasoning`)

Figure 2: Effect sizes (odds ratio) by number of participants (human sample) or number of items (LLM samples) for four effects (frame, order, version: controversial, and version: covered). Results show that all effects were stable when data collection was stopped. Lower odds ratios indicate stronger effects.

analysis (Anderson et al., 2022) for our human and LLM results.

We calculated odds ratios for each of the four effects defined by the maximal model (frame, order, version: controversial, and version: covered) continuously as we added more participants[10] (for the human subjects sample) or items (for the LLM samples). Over the final ten observations, the maximum effect size change was .016 for the human sample and .008 for the LLM samples. These results (see Figure 2) indicate that the effect sizes of all fixed effects were stable when data collection was stopped, suggesting that they are likely good estimates of the true effect sizes in the population.

4.5 Correlations with Human Judgments

While `reasoning` models showed greater robustness to irrelevant prompt variations, this does not guarantee that they accurately approximate human judgments under the same conditions. In order to answer this question, we calculated the percentage of our human participants who made a covered decision for each version of each item (averaged across frame and order conditions), and we then performed a point-biserial correlation (Kornbrot, 2014) with the outputs from each model. The results (see Table 2) show that while `reasoning` models outperformed non-`reasoning` models, none of them achieved better than a moderate correlation with human judgments, and all of them generated some judgments that no humans agreed with (see Figure 3 and example item in Figure 4).

In relation to Waldon et al.'s (2023) study For comparison, we accessed Waldon et al.'s (2023) human subjects data and found that their participants' responses were much more closely aligned to ours (r = .87, r^2 = .76), compared to any of the models (r = .57, r^2 = .32 being the highest). This demonstrates that the difference between LLMs and our human sample is much greater than the difference

[10] For an introduction to effect size stabilization see Anderson et al. (2022)

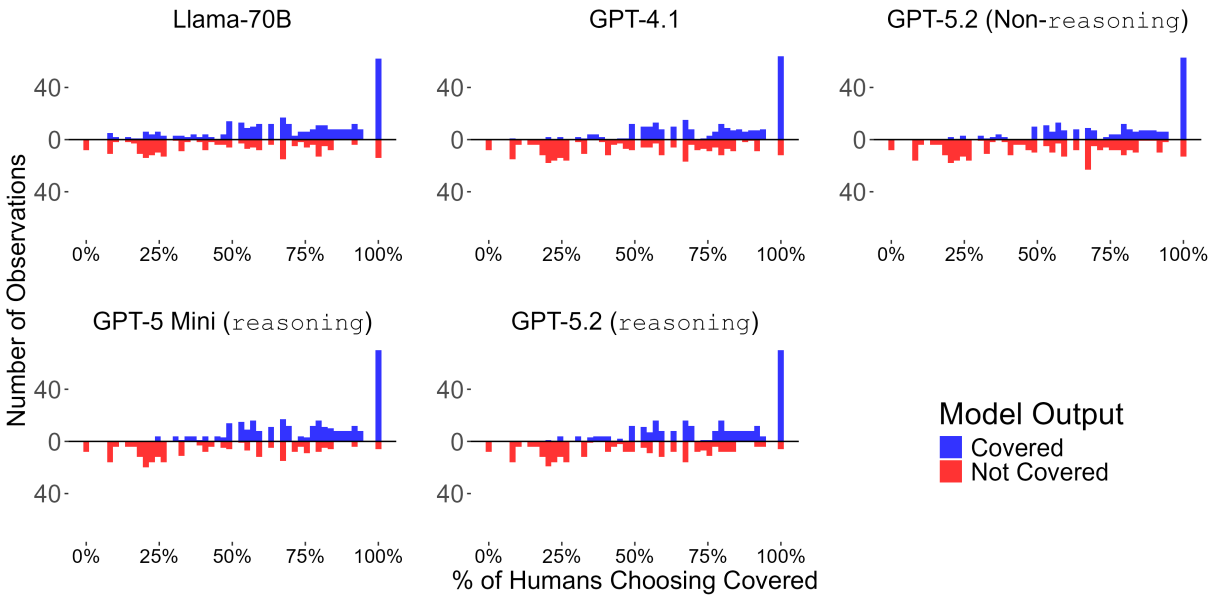


Figure 3: Relationship between LLM-generated and human made coverage determinations. The x-axis shows the percent of humans who made a covered decision for each specific prompt. The y-axis shows the number of LLM observations (e.g. prompts). The color and direction of each bar indicates the model generated decisions for those prompts. The red bars at the right side of each panel indicate prompts for which models responded with *not covered* decisions, despite humans unanimously agreeing they should be *covered*. One such prompt is illustrated in Figure 4

Model	r	r^2
Llama-70B	.38	.14
GPT-4.1	.48	.23
GPT-5.2 (Non-reasoning)	.47	.22
GPT-5 Mini (reasoning)	.57	.32
GPT-5.2 (reasoning)	.52	.27

Table 2: Point-biserial (Kornbrot, 2014) correlations between averaged human coverage decisions and model generated coverage decisions. Even large reasoning models only achieved moderate correlations.

between these two human samples.

5 Discussion

The goal of the current study was to investigate whether querying LLMs can provide a good approximation of human metalinguistic judgments. Particularly, we investigated how irrelevant prompt variation (positive vs. negative framing and answer order) affected both human and LLM responses to questions of legal interpretation.

We found that humans showed a high level of robustness in this task. We found no evidence of acquiescence bias (i.e., frame effects) or order effects on human judgments in our sample. Furthermore, the effect size stabilization analysis suggested that these minimal effect sizes of frame and order were good estimates of the true effect sizes in the population.

Sample item where LLMs diverged from human judgments:

Vivian has home insurance for "Loss and Accidental Damage," which only applies if the damage occurs "while temporarily away from home anywhere in the world." Vivian lives alone in a single-family house in the suburbs. She decides to take a vacation abroad, and her house is completely vacant while she's away. While abroad, her house suffers from an electrical short circuit that causes a severe fire, damaging nearly the entire home. Vivian files a claim with her insurance company for the damages.

Select your response to the following statement: Vivian is / is *not* covered by the Loss and Accidental Damage policy.

Figure 4: Across frame and order variations, models generated a *not covered* response to 13/20 instances of the covered version of this prompt (across 2 frames, 2 orders, and 5 systems for each frame+order combination). Humans (n = 13) universally made a *covered* decision.

In contrast, LLMs were more sensitive to these factors. Llama-70B showed a significant effect of order, and all of the non-`reasoning` models (including GPT-5.2) showed significant effects of frame, in the form of a *dis*acquiescence bias (i.e., they were more likely to generate a covered decision in the negative frame compared to the positive frame).

We also found that enabling `reasoning` (Singh et al., 2025) helped to mitigate this prompt sensitivity. With `medium effort reasoning` enabled, GPT-5.2 and GPT-5 Mini showed no significant effects of frame or order. However, despite these improvements, our correlation analysis demonstrated that none of the models we tested had more than moderate correlations with human judgments, and there were some scenarios for which all five models denied coverage when none of our human participants did.

These findings have implications both for the use of LLMs in survey paradigms and for legal interpretation more specifically. One of the key motivations for using LLMs to approximate human metalinguistic judgments is that LLMs train on vast amounts of natural language data. In the words of Judge Newsom, "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" (Newsom, 2024a). Proponents for the use of LLMs to approximate survey results in social science research make similar big data justifications (Cho et al., 2024). Our results call this line of reasoning into question.

Instruction tuning (Peng et al., 2023), reinforcement learning from human feedback (Bai et al., 2022), and `reasoning` (OpenAI et al., 2024b) all alter model behavior in ways that do not reflect the structure of the underlying pretraining data. We found that such adjustments are necessary to achieve adequate performance on these metalinguistic tasks, which challenges the motivation for using LLMs in the first place. Whatever these judgments rely on, it is not merely statistical predictions about how "ordinary people ordinarily use words". Even worse, from a pragmatic perspective, we found that even cutting-edge models using these techniques do not provide high-quality approximations of human consensus judgments.

Based on these findings, we do not recommend querying LLMs as a substitute for human subjects experimentation in applications like legal interpretation where alignment with the metalinguistic judgments of ordinary people is key.

6 Conclusion

In this study, we directly compared the robustness of human and LLM metalinguistic legal judgments under prompt variation. We found (1) that LLMs were prompt sensitive in this task, while humans were not, (2) that enabling `reasoning` improved LLM stability, and (3) that despite this improvement, even `reasoning` LLMs showed only moderate correlations with human consensus judgments. Overall, we conclude that despite the improvements to judgment stability made possible by `reasoning`, LLMs remain poor proxies for human interpretive judgments about language.

Our findings raise concerns about the motivating factors for LLM use in legal interpretation and demonstrate that even sophisticated LLMs do not provide high-quality approximations of human consensus in this task.

Limitations

Our experimental stimuli focused specifically on insurance coverage contracts written in English. We do not claim that they are representative of all legal interpretation tasks or metalinguistic judgments, more broadly.

We tested only a limited range of models. Based on this data alone, we cannot be certain how our findings might generalize to other models.

Our study focused primarily on the *reliability* concern with the use of LLMs for legal interpretation. As we discussed previously, there are many other concerns which ought to be addressed before LLMs can be safely used in the courtroom. Of particular concern is the possibility that an LLM developer could tune its model's outputs in a way that benefits their own interests when it is used for legal interpretation. While some courts have dismissed this concern (Newsom, 2024a), we are not convinced that such fine-tuning would be as detrimental to general model performance as those courts have suggested. This risk is especially worth considering because the `reasoning` models, which showed the highest reliability in our tests, are proprietary, closed-source models. The improvements to prompt sensitivity that we observed are likely not sufficient cause to justify the use of closed-source models for real-world legal interpretation.

Acknowledgments

This research was supported in part by NSF award IIS-2144881 and by the Fritz Family Fellowship. This work used DeltaAI at National Center for Supercomputing Applications and Anvil at Purdue University through allocation CIS250932 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. We thank Brandon Waldon, Junghyun Min, Minh Duc Bui, members of the NERT lab, and our anonymous reviewers for their thoughtful recommendations.

References

Richard B Anderson, Jennifer C Crawford, and Michael H Bailey. 2022. Biasing the input: A yoked-scientist demonstration of the distorting effects of optional stopping on bayesian inference. *Behavior Research Methods*, 54(3):1131–1147.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416.

J Martin Bland and Douglas G Altman. 2000. The odds ratio. *BMJ*, 320(7247):1468.

Benjamin Minhao Chen. 2025. Do linguistic canons matter? *Connecticut Law Review (Forthcoming)*.

Carissa Chen. 2024. The textualist test for large language models. *Preprint*, Social Science Research Network:5495399.

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. LLM-based doppelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*.

Jonathan H Choi. 2025. Off-the-shelf large language models are unreliable judges. *Preprint*, Social Science Research Network:5188865.

Justin Curl, Peter Henderson, Kart Kandula, and Faiz Surani. 2025. Judges shouldn't rely on AI for the ordinary meaning of text. Lawfare.

Nikola L Datzov. 2025. AI jurisprudence: Toward automated justice. *Northwestern Journal of Technology and Intellectual Property*, 23(1):1.

Christoph Engel and Richard H McAdams. 2024. Asking GPT for the ordinary meaning of statutory terms. *Journal of Law, Technology & Policy*, page 235.

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, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tom Heyman and Geert Heyman. 2019. Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, 72(8):2084–2109.

Tom Heyman and Geert Heyman. 2024. The impact of ChatGPT on human data collection: A case study involving typicality norming data. *Behavior Research Methods*, 56(5):4974–4981.

Joseph M Hilbe. 2009. *Logistic regression models*. Chapman and hall/CRC.

David A Hoffman and Yonathan Arbel. 2024. Generative interpretation. *New York University Law Review*, page 451.

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*.

Diana Kornbrot. 2014. Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*.

D. Kurzon. 2006. Law and language: Overview. In *Encyclopedia of Language & Linguistics (Second Edition)*, pages 728–731. Elsevier, Oxford.

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, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Thomas R. Lee and Jesse Egbert. 2024. Artificial Meaning? *Preprint*, Social Science Research Network:4973483.

Hannes Leeb and Benedikt M Pötscher. 2009. Model selection. In *Handbook of Financial Time Series*, pages 889–925. Springer.

Eric Martínez. 2025. Traditional and Computational Canons. *Preprint*, Social Science Research Network:5155444.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Simon Moss. 2008. Acquiescence bias.

Kevin Newsom. 2024a. Concurring opinion in *Snell v. United Specialty Insurance Co.*

Kevin Newsom. 2024b. Concurring opinion in *United States v. Deleon*.

Aileen Nielsen, Chelse Swoopes, and Elena Glassman. 2025. Law is vulnerable to AI influence; Interface design can help. *Preprint*, Social Science Research Network:5387231.

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, and 262 others. 2024a. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024b. OpenAI o1 system card. *Preprint*, arXiv:2412.16720.

Andrea Pedrotti, Giulia Rambelli, Caterina Villani, and Marianna Bolognesi. 2025. How humans and LLMs organize conceptual knowledge: Exploring subordinate categories in Italian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4464–4482.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *Preprint*, arXiv:2304.03277.

Prolific. 2025. What are representative samples on prolific.

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 *FAccT '26: The 2026 ACM Conference on Fairness, Accountability, and Transparency*.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Antonin Scalia and Bryan A. Garner. 2011. *Reading Law: The Interpretation of Legal Texts*. West Group.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Lawrence M Solan and Tammy Gales. 2017. Corpus linguistics as a tool in legal interpretation. *Brigham Young University Law Review*, 2017(6):1311–1357.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Kevin Tobia. 2020. Testing ordinary meaning. *Harvard Law Review*, 134:726.

Kevin Tobia. 2021. The corpus and the courts. *The University of Chicago Law Review*.

Kevin Tobia. 2022. Experimental jurisprudence. *The University of Chicago Law Review*, 89(3):735–802.

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).

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. *The Georgetown Law Review*, 114.

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. *Preprint*, arXiv:2406.16838.

Daniel Wilf-Townsend and Kevin Tobia. 2025. Generative AI and courts in the United States. *Preprint*, Social Science Research Network:5243402.

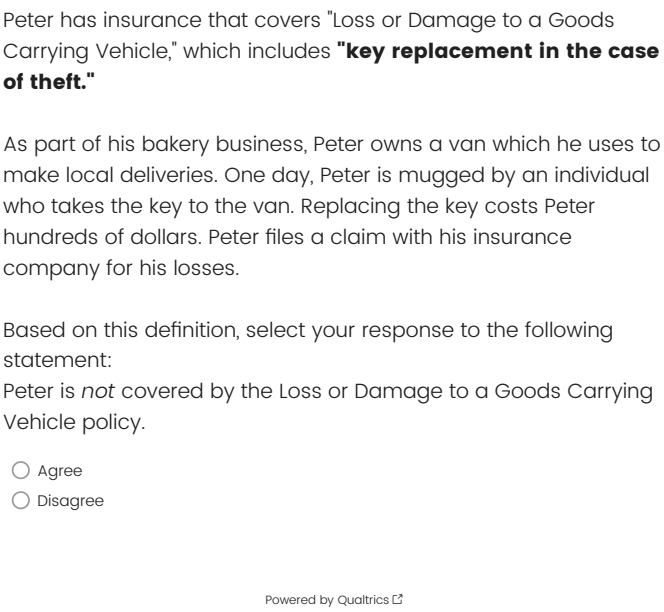


Figure 5: A reference example question from the survey. This is not an exact replication of how it was shown to the subjects in the interactive web interface.

A Survey

The survey was run with Qualtrics, and collected 1728 judgments from 288 subjects (§3), by showing 6 conditioned items to each subject. For the 46 items with 12 conditions each, we collected a median of 3 observations per conditioned item, and median 37 observations per item. An example conditioned item is shown in Figure 5 for reference.

B Models

We used both open-weight and closed-source models in this paper.

Model	Technical Report
Llama-3.3-70B-Inst	Grattafiori et al. (2024)
GPT 4.1	OpenAI et al. (2024a)
GPT 5 Mini, 5.2	Singh et al. (2025)

Table 3: Models used for our experiments and their corresponding technical reports.

`reasoning` in GPT models. In this paper we use GPT-5.2 and GPT-5 as reference `reasoning` models. Singh et al. (2025) notes that the `reasoning` in these models is trained with reinforcement learning, and utilize 'parallel test-time compute'. We point the readers to Welleck et al. (2024), Snell et al. (2025), and OpenAI et al. (2024b) for more on test-time compute, including scaling and selection.

C Implementation

We used the OpenAI Python SDK[11] for implementation of our trials with OpenAI models.

OpenAI provides an `effort` parameter to influence the `reasoning` effort. We utilized these model-specific parameters accordingly in our trials. We configured `max_output_token` to be 2064 for models with reasoning tokens and did not vary this across the different effort settings. We set the much lower 16 token setting for models without test-time compute, such as GPT-4.1. We also additionally used `temperature=0` for the models that support it, namely 4.1 mini. Since we used Batch API,[12] we will also make the responses available for reproducibility when publishing the code.

We use `vllm` (Kwon et al., 2023) to implement our inference pipeline for open-weight models and use the model implementations available on Huggingface Model Hub.[13] We utilized two Nvidia GH200 GPUs for this inference.

D Likelihood Ratio Test Results

Significance results from the likelihood ratio tests of regression model fit are shown below.

[11] https://pypi.org/project/openai/

[12] https://developers.openai.com/api/docs/guides/batch

[13] https://huggingface.co/models

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			1727	2321.70		
version	2	381.25	1725	1940.40	< .001	**
frame	1	2.35	1724	1938.00	.125	
order	1	0.40	1723	1937.70	.528	

Table 4: Likelihood ratio test results for coverage decisions made by human subjects recruited from Prolific. Results show that *only* version significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			551	750.49		
version	2	137.57	549	612.92	< .001	**
frame	1	141.75	548	471.17	< .001	**
order	1	6.98	547	464.19	.008	*

Table 5: Likelihood ratio test results for coverage decisions generated by Llama-70B. Results show that version, frame, and order significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			551	759.94		
version	2	302.10	549	457.84	< .001	**
frame	1	40.38	548	417.46	< .001	**
order	1	1.86	547	415.61	.173	

Table 6: Likelihood ratio test results for coverage decisions generated by GPT-4.1. Results show that version and frame significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			551	745.52		
version	2	300.47	549	445.05	< .001	**
frame	1	23.64	548	421.41	< .001	**
order	1	1.52	547	419.89	.217	

Table 7: Likelihood ratio test results for coverage decisions generated by GPT-5.2 without `reasoning` enabled. Results show that version and frame significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			551	757.32		
version	2	304.51	549	452.81	< .001	**
frame	1	0.35	548	452.46	.554	
order	1	0.69	547	451.77	.407	

Table 8: Likelihood ratio test results for coverage decisions generated by GPT-5 Mini, with `medium effort reasoning`. Results show that *only* version significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05

	Df	Deviance	Resid. Df	Resid. Dev	p	
NULL			551	765.23		
version	2	307.67	549	457.56	< .001	**
frame	1	1.38	548	456.18	.240	
order	1	0.00	547	456.18	1.000	

Table 9: Likelihood ratio test results for coverage decisions generated by GPT-5.2, with `medium effort reasoning`. Results show that *only* version significantly improved model fit. Signif. codes: '**' < .001, '*' < 0.05