Predicting consensus in legal document interpretation

Brandon Waldon,[1] **Madigan Brodsky,**[1] **Megan Ma,**[2] **and Judith Degen**[1]
bwaldon@stanford.edu, madiganb@stanford.edu, meganma@law.stanford.edu, jdegen@stanford.edu
[1]Department of Linguistics, Stanford University [2]Center for Legal Informatics, Stanford Law School
Stanford, CA 94305 USA

Abstract

We present a large-scale conceptual replication of an experiment that provided evidence of *false consensus biases* in legal interpretation: when reading a legal contract, individuals tend to over-estimate the extent to which others would agree with their interpretation of that contract (Solan, Rosenblatt, & Osherson, 2008). Our results are consistent with this previous finding. We also observe substantial unexplained item-level variation in the extent to which individuals agree on contract interpretation, as well as unexplained variation in the extent to which the false consensus bias holds across different contexts.

In a first step towards understanding the source(s) of this variability, we show that a state-of-the-art large language model (LLM) with zero-shot prompting does not robustly predict the degree to which interpreters will exhibit consensus in a given context. However, performance improves when the model is exposed to data of the form collected in our experiment, suggesting a path forward for modeling and predicting variability in the interpretation of legally-relevant natural language.[1]

Keywords: experimental jurisprudence, large language models, false consensus, linguistic ambiguity

Introduction

Legal decision-making often depends on how legal practitioners interpret the language in which a document is written (e.g. in a homeowner's insurance contract, a provision that the policyholder must take *reasonable care* of her property to receive benefits), as well as on whether the document's meaning is judged to be 'clear.' For example, the principle of *contra proferentem* dictates that when a dispute arises over the interpretation of a contract with 'unclear' terms, the preferred interpretation is the one that leads to a favorable outcome for the party that merely agreed to (but did not draft) the contract. Similarly, the *parol evidence* principle "permits the admission of extrinsic evidence to resolve ambiguity in contractual language, but prohibits evidence offered to vary the terms of a contract whose language is clear" (Solan et al., 2008: 1269).

In an early example of research in the nascent field of *experimental jurisprudence*,[2] Solan et al. (2008) demonstrated that people – both members of the lay public and professional American judges – inaccurately estimate the extent to which legal language is, in fact, clear. Solan et al. asked participants to read passages in which an insurance policyholder experiences some material loss and attempts to file a claim with an insurer on the basis of that loss. Participants reported whether they believed the claim was covered by the insurance, an assessment that hinged on the interpretation of an under-specified term in the contract (either *pollution* or *earth movement*); they then estimated the extent to which other participants would agree with their individual assessment. Participants demonstrated a consistent *false consensus bias* (Ross et al., 1977; C. E. Brown, 1982; Krueger & Clement, 1994): though participants were roughly split as to whether they believed the claims were covered, they tended to believe they had provided the majority response.

Solan et al.'s results suggest that interpreters tend to overestimate the clarity of natural language in legal contexts. These findings have widespread implications for the legal system: as the authors write, unchecked false consensus biases could lead to "misapplications of operative legal principles" such as contra proferentem and parol evidence (2008: 1295). The authors note that similar principles susceptible to false consensus biases exist in other legal domains; for example, in criminal law, the *rule of lenity* dictates that unclear laws should be construed in favor of criminal defendants.

Solan et al.'s finding has helped to motivate modern empirical approaches to legal interpretation that de-emphasize the role of armchair linguistic intuitions (Solan & Gales, 2017; Macleod, 2019; Tobia, 2020), including corpus-assisted and experimental methodologies reviewed below. To lend more credence to Solan et al.'s finding, we report a large-scale conceptual replication of their original study (which featured just four critical items that tested interpretation of just two terms). In line with Solan et al.'s results and a much larger tradition of findings in cognitive psychology, we find strong evidence of a false consensus bias in the interpretation of legal contracts. Moreover, the results of our replication indicate that there is substantial unexplained item-level variation in the extent to which individuals agree on linguistic interpretation, as well as unexplained variation in the extent to which the false consensus bias holds across interpretive contexts. We argue that this observed variation motivates a larger effort to identify the features of linguistic context that modulate both interpretive consensus and the false consensus bias.

In a first step towards this broader enterprise, we review the strengths and weaknesses of current empirical jurisprudence methods before assessing the ability of a state-of-the-art large language model (LLM) to predict the observed patterns of interpretive variation from our experiment. Though the model we assess has demonstrated remarkable performance on a number of downstream natural language tasks in zero-shot settings in which the model is exposed to no training data (Ouyang et al., 2022), our results suggest that zero-shot prompting is insufficient to robustly predict when and to what degree human beings will exhibit interpretive consensus. Model performance improves markedly with few-shot prompting, though performance is still far from robust. We

[1]Corresponding author: B. Waldon bwaldon@stanford.edu
[2]See Tobia (2022) for a recent review.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

conclude with implications for language scientists, scholars of empirical jurisprudence, and researchers interested in the behavior of large artificial neural network systems.

Experiment

The experiment extends the methodology reported by Solan et al. (2008), in which participants read a short passage of text and judged whether an insurance claim described in the passage was covered by an insurance policy specified in the text. To investigate the presence of false consensus biases, participants then estimated how many other participants would agree with their individual judgment.

Methods

Participants: We recruited 1380 participants through Prolific (US-Based, native English speakers, >99% approval rating on Prolific). Participants were paid $0.85 and median task completion time was roughly 3 minutes 15 seconds, for a median hourly compensation rate of $\sim$ $16/hr.[3]

Materials and procedure: An example trial is reproduced in Fig. 1. Each trial introduced participants to a policyholder (e.g. Tom in Fig. 1)[4] who, in light of their insurance and in light of an event described in the text, files a claim with their insurance company. The experiment featured 46 groups of 3 items, where each item in a group targeted interpretation of the same linguistic expression in one of 3 variations of the same hypothetical scenario. These variations corresponded to 3 experimental conditions:

- *Covered*: as researchers, we believed prior to data collection that most people would think that the claim described in the text is covered by the individual's insurance.
- *Not covered*: we believed that most people would think that the claim described in the text is *not* covered.
- *Controversial*: we believed there to be two plausible interpretations of the text: one that suggests the individual is covered, and one that suggests the individual is not covered.[5]

Across the 46 items, we tested interpretation of 25 commonly-occurring terms in consumer insurance contracts (e.g. *Wind Damage*). Our methodology departs from that of Solan et al. (2008) in that their items described insurance coverage inclusion/exclusion clauses that hinged on the interpretation of a single under-specified linguistic expression (*pollution* or *earth movement*), with no definition or elaboration of the meaning of that expression in context. To construct our stimuli, we drew from publicly-available, consumer-facing materials published by a global provider of home and vehicle insurance, in which insurance terms are explicated for (prospective) policyholders. We adapted these elaborations for our stimuli, each of which featured one of three possible elaboration types (presented in **bold** for the participant):

- *Exhaustive definition*: The term of interest was explained in a manner that suggested the term is incompatible with meanings not explicitly specified in the text. (The *Wind Damage* example in Fig. 1 contains such a definition).
- *Non-exhaustive definition*: These elaborations featured an explicit 'includes' clause, e.g. *Peter has insurance that covers "Loss or Damage to a Goods Carrying Vehicle," which includes* ***"key replacement in the case of theft."*** The term was thus explained in a manner that suggested the term is compatible with meanings not explicitly specified in the text.
- *Exclusion*: These elaborations featured an explicit 'excludes' clause, e.g. *Dillon's car insurance policy includes coverage for "Vehicle Theft," which excludes* ***"loss or damage caused by theft or attempted theft if your car was taken by a member of your family or household, or taken by an employee or ex-employee."*** In this sense, similar to non-exhaustive definitions, exclusion definitions were presumed to be compatible with meanings not specified in the text.

The locus of linguistic uncertainty – that is, the expression that we expected would lead (in controversial cases) to substantial population-level variation in the overall evaluation of the insurance coverage – was a term in the elaboration. The 'Covered' and 'Not covered' items kept these elaborations constant but differed in that they described events that we believed were uncontroversially within (outside) the scope of the described coverage.

Participants were randomly assigned to 3 of 46 item groups in a manner that guaranteed that no participant ever provided a judgment about the same contract term more than once. Each participant was then assigned to the 'Covered,' 'Not covered,' and 'Controversial' conditions exactly once, such that they saw one item in exactly one of these 3 conditions.[6]

The 'Covered' and 'Not covered' conditions served as attention checks for participants; those who provided the 'unexpected' response to Question 1 on both items seen in these conditions ('No' for 'Covered'; 'Yes' for 'Not covered') were excluded from the analysis.

Results

34 participants were excluded on the exclusion criterion described above, and data from a further 8 participants were excluded because they reported a native language other than English. Following Solan et al. (2008), for each participant p's response, we recorded p's 'error' in estimating population level consensus in interpretation of item i by subtracting [the proportion of participants who provided p's response to Question 1 on i] from [p's response to Question 2 on i]. Thus, a positive error indicates that p over-estimated the level of population-level agreement with p's judgment on i.

To determine whether there was an overall consensus bias among participants when evaluating Controversial items, we

[3]Methods, materials, exclusions, and analyses for the experiment were pre-registered through the Open Science Foundation at https://osf.io/vtqg8/. Materials, data, and code are available at https://github.com/madiganbrodsky/vague_contracts.

[4]Half of the items featured conventionally male names; half featured conventionally female names.

[5]These conditions corresponded, respectively, to the two 'prototypical situation' control conditions and one 'experimental' condition originally reported by Solan et al. (2008).

[6]In Solan et al.'s study, participants saw exactly one item.

Tom's home insurance policy includes coverage for "Wind Damage," defined as **"damage from wind speeds of at least 55 mph."**

'Covered' condition:	*'Not covered' condition*:	*'Controversial' condition*:
Tom's house is located near a large lake. One day, strong winds in excess of 55mph blow across the lake and towards Tom's house, damaging the roof. Tom files a claim with his insurance company for the damage.	Tom's house is located near a large lake. One day, strong winds in excess of 55mph blow across the lake while Tom is working on his roof, but it's a loud and surprising ring from his cell phone that causes him to drop his heavy toolbox and thereby damage the roof. Tom files a claim with his insurance company for the damage.	Tom's house is located near a large lake. One day, strong winds in excess of 55mph blow across the lake, causing waves to crash into Tom's house and thereby damaging the roof. Tom files a claim with his insurance company for the damage.

1. Do you think that the claim is covered under Wind 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]

Figure 1: Example materials on a trial of the experiment. The locus of interpretive uncertainty in this example is causative *from* in "damage from wind speeds..." – which is consistent with both proximate and distal causation.

conducted a one-sided, one-sample Fisher test on the distribution of participant errors for those items. The results of this test indicated strong evidence that participant errors were above zero ($\mu = 15.32$, 95% CI = [14.06, 16.49][7], $p < 0.001$); that is, that participants overall tended to overestimate the extent to which others would agree with their individual judgments.

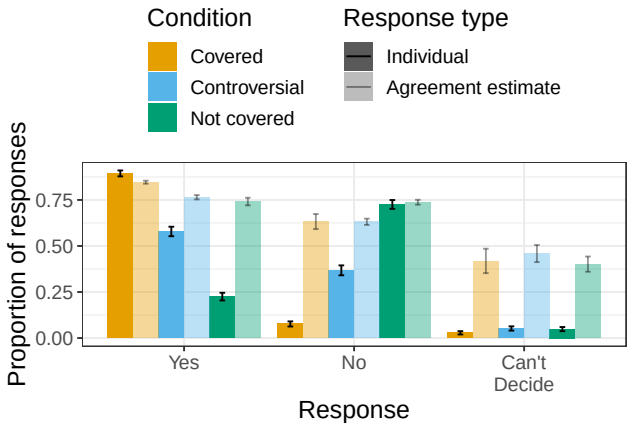

Figure 2: Response proportions alongside agreement estimates by response type and condition. Error bars indicate 95% bootstrapped confidence intervals.

As can be seen in Fig. 2, estimated agreement exceeded observed levels of agreement in not only the 'Controversial' condition but also in the 'Covered' and 'Not covered' conditions when participants were in the minority of respondents. Across all three conditions, participants providing a 'Can't Decide' response on Question 1 also vastly overestimated the extent to which others would share in their assessment.

When examining behavior in the 'Not covered' condition (the green bars of Fig. 2), it is notable that response patterns are visibly far less uniform than in the 'Covered' condition (yellow bars): whereas 'Yes' responses are close to ceiling (89%) in the former condition, 'No' conditions are less frequent in the latter (72%). These results are notable in light of those reported originally by Solan et al., who tested interpretation in a between-subjects manipulation that we did not pursue: in the first condition, the policyholder received insurance benefits if the event described in the vignette as was deemed to count as an instance of *pollution / earth movement*; in a second condition, the policyholder did not receive benefits if this was the case. Solan et al. introduced this manipulation to "control[] for result-oriented responses reflecting a possible bias against either insurance companies or plaintiffs" (2008: 1268), but they report that "There was no evidence... that people respond differently to the scenario depending on whether saying 'yes' meant triggering insurance or excluding insurance" (ibid: 1269).

Conversely, we find that in scenarios where we believed an insurance policy cannot be plausibly construed in favor of the policyholder, a sizable minority of participants believe policyholders to be covered. In a post-hoc analysis, we further explored this behavior by coding responses in the 'Covered' and 'Not covered' conditions according to whether the response was the expected majority response. ('Yes' for 'Not covered' items; 'No' for 'Covered' items). Using the `brms` package in R, we ran a Bayesian mixed effects logistic regression predicting log odds of expected majority response from a fixed effect of condition (reference level: 'Covered') with random by-participant and by-item random intercepts as well as a by-item random slope for condition, the maximal random effects structure that allowed for convergence. We found strong evidence for a main effect of condition ($\hat{\beta}$ = -1.43, 95% CI = [-1.88, -1.03]), consistent with a bias against answering 'No' in the 'Not covered' condition.

At the by-item level, we observe variation in levels of interpretive consensus, as Fig. 3 shows. Notably, though rates of 'Yes' responses generally tend to be highest in the 'Covered' condition and decrease progressively in the 'Controversial' and 'Not covered' conditions, there are exceptions to this generalization (e.g. *Escape of Water I*, *Identity Theft I*) in which

[7]Confidence intervals were computed via bootstrap sampling.

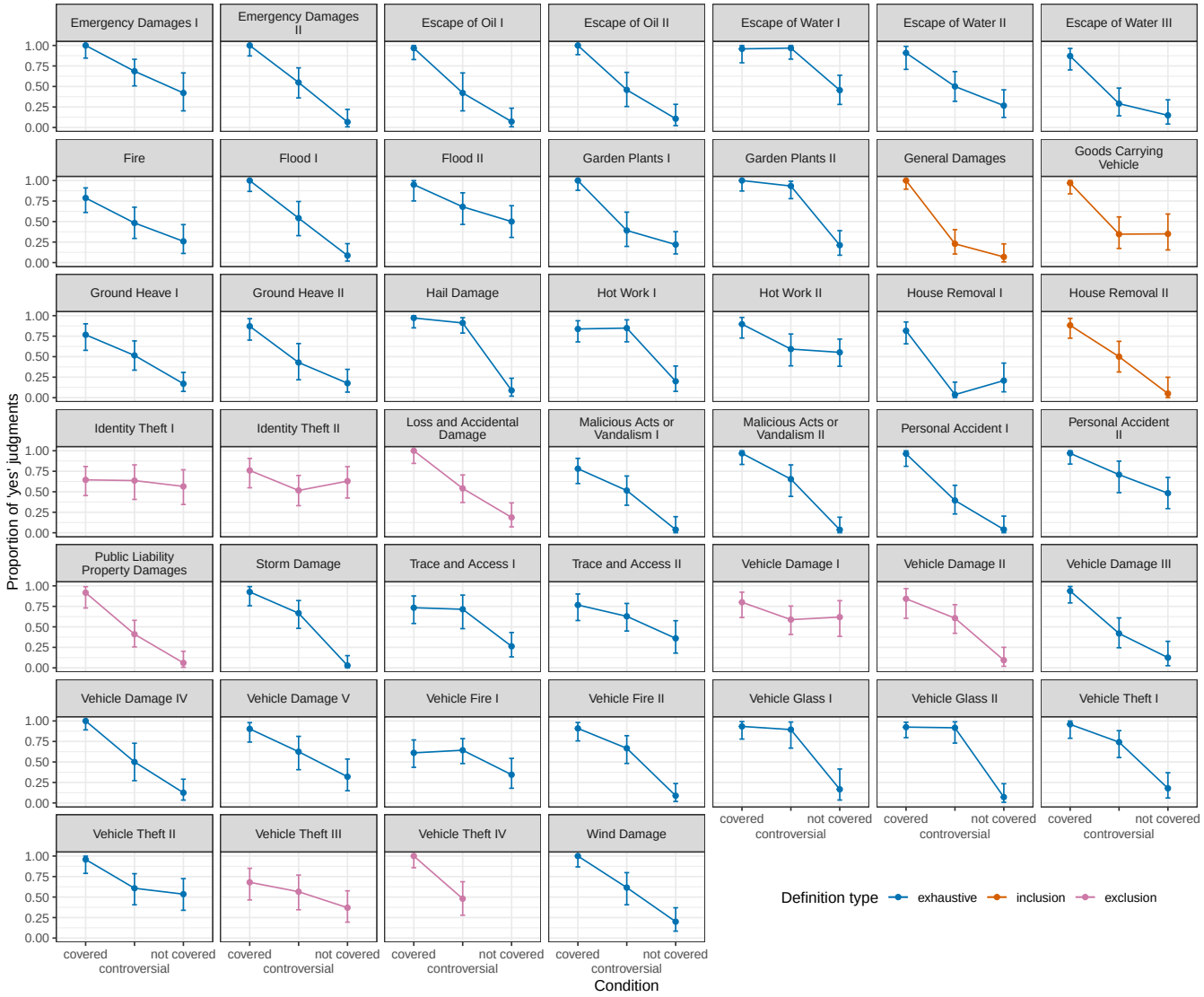

Figure 3: By-item proportions of 'Yes' responses by condition. Error bars indicate 95% bootstrapped confidence intervals.

patterns of response are largely indistinguishable across conditions. Indeed, that rates of 'Yes' response are not uniformly at ceiling for 'Covered,' around 50% for 'Controversial,' and at floor for 'Not covered' likely points to the fact that, as researchers, we exhibited some 'false consensus biases' of our own in the design of our experimental stimuli.

Lastly, as seen in Fig. 4, there is substantial unexplained item-level variation in the strength of the false consensus bias. A Bayesian mixed-effects linear regression predicting mean agreement estimate from centered fixed effects of proportion of individual judgment, judgment type ('Yes' or 'No'), their interaction, and random by-item intercepts, yields strong evidence that agreement estimates increase with increasing frequency with which that judgment was provided ($\hat{\beta} = .19$, 95% CI = [.16, .23]). Moreover, 'yes' responses were consistently associated with greater agreement estimates than 'no' responses ($\hat{\beta} = 10.01$, 95% CI = [7.61, 12.34]). Notably, in some cases where participants largely agree on an interpretation, participants tend to, if anything, slightly underestimate population-level consensus. This finding merits further inves-

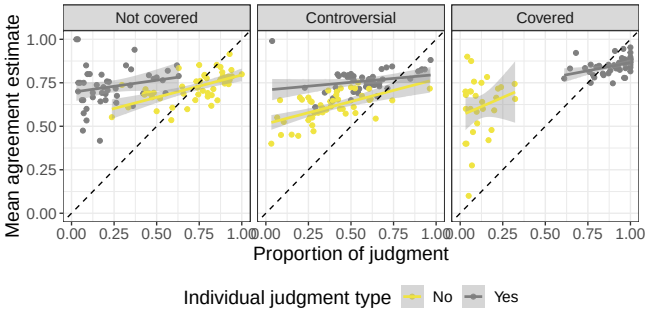

Figure 4: By-item agreement estimates against response proportions for the responses 'No' and 'Yes'. Dashed line indicates expected pattern if participants' agreement estimates perfectly tracked the true population agreement.

tigation in future research.

Can LLMs predict interpretive variability?

The results of our experiment suggest that there is considerable unexplained variation when it comes to contract interpretation: the degree of interpretive consensus (and of false consensus bias) varies considerably by-item. Is there systematicity to this variation, and if so, how can we accurately predict these interpretive patterns?

Both questions are of considerable importance to legal practitioners because, as discussed above, biases in legal interpretation can have immense real-world consequences. Aware of the pitfalls of textual analysis grounded exclusively in introspective judgements, lawyers and judges have increasingly looked to external sources of linguistic evidence to supplement their interpretive activities. Existing scholarship has questioned the utility of analyses informed by dictionary definitions, a common practice in American courts which allows for several degrees of user freedom: the lawyer or judge is free to choose the dictionary and definition that they believe to be most germane to understanding a word or phrase as it appears in the context of a statute or other legal text. Legal corpus linguistics – the practice of triangulating the meaning of legal language through linguistic corpora, including through concordance and raw frequency analyses – offers more datapoints than do dictionaries but, as Macleod (2019) notes, similarly risks relying on data drawn from linguistic contexts that bear little resemblance to the particular legal document context on which a legal dispute may hinge.

Results such as ours underscore these risks. When parties draft and analyze legal documents, they are likely to overestimate the extent to which the population at large would concur with their individual interpretive judgments. Overly subjective methods for triangulating legal meaning may do little to attenuate such biases: as Brudney and Baum (2013) argue, judges likely often deploy dictionary definitions to simply *confirm* prior beliefs regarding the so-called 'ordinary' linguistic meaning of legal texts. To address these concerns, Macleod (2019) advocates for the use of experimental survey methods in which researchers directly ask members of the lay public to interpret relevant natural language in contexts that tightly match the relevant features of a legal dispute.

Before we consider the extent to which LLMs might usefully supplement the legal practitioner's empirical toolkit, it is worth emphasizing that LLMs have been repeatedly demonstrated to reify human biases that are reflected in training data (Bordia & Bowman, 2019; Schramowski et al., 2022, inter alia), and that 'de-biasing' such models is an active area of NLP research in which researchers attempt to align model behavior with some particular ideal. Moreover, as with the tools mentioned above, we acknowledge that LLMs could, in principle, be used in a myriad of unsophisticated ways to simply confirm the interpretive biases of the end user.

On the other hand, LLMs possess the potentially useful feature of having been trained on text produced by billions of human beings in many billions of individual contexts of use; moreover, they have demonstrated some ability to predict population-level variation for a variety of linguistic judgments, including scalar inferences from *some* and *or* (Schuster et al., 2020; Li et al., 2021). LLMs also possess the striking feature of inherently encoding high dimensional, highly contextual representations of lexical units (Petersen & Potts, 2023); thus, similar to experimental methods, they are well-posed to facilitate linguistic analyses that respect (rather than abstract away from) the role of context in interpretation.

Therefore, we considered whether a state-of-the-art LLM can predict the sort of interpretive variation that we observed in the experiment. We employ OpenAI's `text-davinci-003`, a 175 billion-parameter LLM fine-tuned with human feedback on a variety of tasks (Ouyang et al., 2022) and made available through OpenAI's API.

Methods

We assessed the performance of `text-davinci-003` on a text insertion task, in which the model receives a prompt consisting of an experimental item (as it appeared to participants). The model then predicts a token to insert within a desired completion. That completion is a statement that closely resembles Question 1 of the experiment, e.g.

```
COMPLETION: Out of 100 randomly-sampled English
speakers, it is estimated that [INSERT] would
believe that the claim is covered under Wind
Damage as it appears in the policy.
```

To create gold labels for each of the 138 items of the experiment, we took the percentage of participant "Yes" responses to Question 1, rounded to the nearest ten, and converted the resulting value to an alphabetical (i.e. non-numeric string). For example, if an item received 62% "Yes" responses, the gold label was `sixty`. (For 0 and 100, the gold labels were `none` and `everyone`, respectively). When prompting the model, we set a logit bias on tokens corresponding to our gold labels, which uniformly increased the probability of the model producing those tokens in inference.

To assess whether our experimental data can augment the predictive capabilities of the model, we used two prompting regimes: a 'zero-shot' regime featuring just the target item; and a 'few-shot' regime that additionally featured examples (with gold label completions) drawn from 3 randomly-selected sets of items, presented in each of the 3 conditions in which the items appeared in the experiment (9 examples total).[8] To compare model performance after few-shot prompting, we assessed prediction accuracy only on the remaining items not included as examples in the few-shot prompt.

Results

We ran the above procedure 15 times (with separate random seeds) to assess model performance given a variety

[8]We pursue zero and few-shot prompting regimes in light of the fact that LLMs including `text-davinci-003` have demonstrated robust performance on a variety of tasks in such settings, see T. Brown et al. (2020) and Ouyang et al. (2022) for more discussion.

of randomly-selected sets of examples in few-shot prompting; we then compared mean-pooled model predictions under each of the two regimes. Those predictions are evaluated against our empirical data in Fig. 5. Compared to the model prompted under the zero-shot regime, the few-shot model makes a wider range of quantitative predictions and also exhibits greater overall predictive accuracy ($R^2 = 0.33$ for few-shot; vs. $R^2 = 0.19$ for zero-shot), suggesting that data of the form collected in the experiment can help to augment the predictive capabilities of LLMs on this task.

Error analysis

With zero-shot prompting, the predictions of `text-davinci-003` fall within a restricted band of high values: `ninety` is the most frequently-produced label at 64% followed by `eighty` at 34%.[9] Few-shot prompting attenuates this prior bias, but the model still tends to over-predict high values: `ninety` is still most frequent (35%) followed by `everyone` (20%) and `eighty` (16%). As a result, model predictions are especially degraded for items where few participants provided a 'Yes' response to Question 1 of the experiment, for example:

```
PROMPT: Cam's home insurance policy includes
coverage for "Storm Damage," defined as
<b>"damage caused by a storm."</b> One
day, a lightning storm passes through
Cam's neighborhood and lightning hits his
neighborhood's power plant, causing a power
outage.  Cam goes to his attic with a candle
to reset the circuit breaker, but he drops the
candle and starts a fire so large that almost
all of his roof burns before firefighters
manage to extinguish it.  Cam files a claim
with his insurance company for the damage.
COMPLETION: Out of 100 randomly-sampled English
speakers, it is estimated that [INSERT] would
believe that the claim is covered under Storm
Damage as it appears in the policy.
```

Across 15 runs, the most frequent few-shot prediction label is `ninety` but the target label is `none`. (Just 3% of participants indicated that the claim is covered in this scenario). In ongoing work, we are investigating whether more data-intensive training regimes (i.e. fine-tuning) can further improve the predictive accuracy of the model over few-shot prompting.

General Discussion

In line with previous findings – not only those of Solan et al. (2008) but across many contexts and modalities in cognitive psychology – we find evidence of a false consensus bias, in our case in the domain of legal interpretation. As the previous section demonstrates, predicting observed interpretive consensus in such contexts is also a non-trivial task for a state-of-the-art neural network language model, suggesting that we have identified a new, challenging benchmark for assessing such systems. The results of our model comparison furthermore suggest that we have identified a form of training data that can ultimately improve model performance.

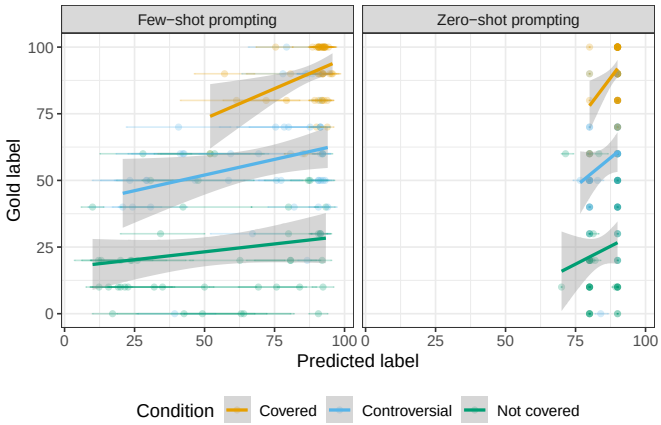

Figure 5: By-item model predictions of `text-davinci-003`, evaluated under few-shot and zero-shot prompting regimes. For the purposes of visualization and to assess model accuracy (R^2), labels are transformed into integers. Error bars are 95% bootstrapped confidence intervals of model predictions.

Our results also suggest avenues for future empirical research. For example, though we find evidence that participant interpretations may be influenced by a desire to construe legal documents in favor of particular parties, it is also possible (due to our own researcher biases) that many scenarios we assumed were certainly 'Not covered' are, in fact, of a less determinate status. To adjudicate between these two possibilities, we intend to more systematically investigate the role of result-oriented biases in legal document interpretation.

Finally, recall that items of our experiment also varied as to how contractual terms of interest were elaborated for the participants. As can be seen in Fig. 3, a number of items which featured an 'exclusion' elaboration (e.g. *Identity Theft I* and *II*) do not pattern as expected, with similar proportions of participants indicating across conditions that the claimant is covered by her insurance.

Consistent with this preliminary finding, it has long been recognized that negative constructions – of which exceptives are a subclass – are more difficult to verify and more costly to process than are positive-form constructions (Clark & Chase, 1972; Fischler et al., 1983). Moreover, though they do not specifically examine exceptives, Martinez et al. (2022a) demonstrate experimentally that contracts are more difficult to interpret when they contain linguistic features that are hard to process. It remains to be seen whether the presence of such features modulates false consensus biases in interpretation, a question we leave to future work.[10]

[9] We see a similar lack of variance in label predictions with smaller OpenAI LLMs prompted under similar conditions.

[10] This future research direction is additionally motivated by the fact that both legal contracts and statutory laws contain far more difficult-to-process constructions than is generally observed in other contexts (Martinez et al., 2022a, 2022b).

Acknowledgements

We would like to thank Cleo Condoravdi, Zehua Li, Julian Nyarko, members of the Stanford ALPS Lab, and the audience at the Stanford CodeX Computable Insurance Workshop for valuable feedback and discussion. We are grateful as well to our anonymous CogSci reviewers for their insightful commentary. This work is supported by a Seed Research Grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

Bordia, S., & Bowman, S. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 7–15).

Brown, C. E. (1982). A false consensus bias in 1980 presidential preferences. *Journal of Social Psychology*, *118*(1), 137–138.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

Brudney, J. J., & Baum, L. (2013). Oasis or mirage: The Supreme Court's thirst for dictionaries in the Rehnquist and Roberts eras. *William & Mary Law Review*, *55*, 483–580.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: an ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*(4), 596–609.

Li, E., Schuster, S., & Degen, J. (2021). Predicting scalar inferences from "or" to "not both" using neural sentence encoders. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 446–450).

Macleod, J. A. (2019). Ordinary causation: a study in experimental statutory interpretation. *Indiana Law Journal*, *94*, 957–1029.

Martinez, E., Mollica, F., & Gibson, E. (2022a). Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, *224*.

Martinez, E., Mollica, F., & Gibson, E. (2022b). So much for plain language: An analysis of the accessibility of United States federal laws (1951-2009). In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 397–403).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *arXiv:2203.02155*.

Petersen, E., & Potts, C. (2023). Lexical semantics with large language models: A case study of English break. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 490–511).

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Schuster, S., Chen, Y., & Degen, J. (2020). Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5387–5403).

Solan, L., & Gales, T. (2017). Corpus linguistics as a tool in legal interpretation. *BYU Law Review*, 1311–1358.

Solan, L., Rosenblatt, T., & Osherson, D. (2008). False consensus bias in contract interpretation. *Columbia Law Review*, *108*, 1268–1300.

Tobia, K. (2020). Testing ordinary meaning. *Harvard Law Review*, *134*, 726–806.

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