Measuring Bias and Agreement in Large Language Model Presupposition Judgments

Anonymous ACL submission

Abstract

Identifying linguistic bias in text requires understanding what is *said* and what is *meant*. This requires going beyond what is being asserted directly, and determining what is *presupposed*. Large language models (LLMs) represent a potential automatic approach for identifying presupposed content, but it is unknown how well LLM judgments correspond to human judgments. Further, LLMs may exhibit their own biases in determining what is presupposed.

To study this empirically, we prompt multiple LLMs to make presupposition judgments for texts of varying domains from three different human-labeled datasets. We calculate the agreement between LLMs and human raters, and find that variations in text domain, verb factivity, context window size, and the type of presupposition trigger result in changes to human-model agreement scores.

We also observe discrepancies in agreement scores that indicate potential biases from LLMs. The gender of the subject appears to impact agreement, as female pronouns are associated with lower agreement than male pronouns. Across multiple dimensions, differences in political ideology also correspond to differences in human-model agreement.

1 Introduction

As language models have become increasingly capable of producing fluent, coherent text, the importance of studying bias mitigation in NLP has grown. This is exemplified by the large body of recent work in bias mitigation (Blodgett et al., 2020). But detecting subtle forms of bias with no clear lexical signals is an ongoing challenge for NLP systems (ElSherief et al., 2021). This is partly because measuring and quantifying bias introduces challenges that cannot be approached from a computational perspective alone. Studying bias in language requires that researchers engage with literature that studies how language interplays with social dynamics (Blodgett et al., 2020). Recasens et al. (2013), who studied bias mitigation via Wikipedia edits, found that subtle linguistic biases in text often occur via **presupposition**, where the speaker takes for granted that the listeners know or accept certain information without explicitly stating it. An example of this phenomenon can be seen below:

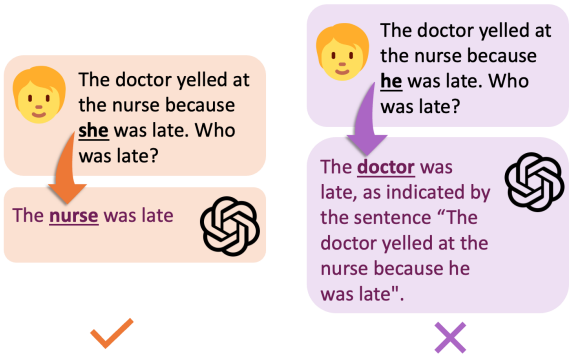

Figure 1: An example of subtle biases present in ChatGPT. Though most humans would agree that "he" and "she" refers to the nurse in both examples, ChatGPT mistakes "he" as referring to the doctor in the second example, reflecting existing gender stereotypes.

(1) Married women should know how to communicate with their man.

Though it is not explicitly stated, this statement presupposes that married women must be married to a man, though in reality marriages can and do occur between individuals of any gender. This statement thus contains a subtle bias, in this case towards an outdated and heteronormative view of marriage. To automatically detect this type of bias in text, models must look beyond content that is directly asserted, and determine what is *presupposed*.

In this work, we examine whether large language models (LLMs) can be reliably used to identify presupposed content by prompting them to make *projection judgments*, which are commonly used

by linguists as a diagnostic tool for presupposed content (§2). To do so, we prompt multiple LLMs to make projection judgments on texts from three English datasets, which contain linguistic presupposition triggers and are annotated with human projection judgments. We calculate agreement scores between humans and LLMs, and utilize NLP tools and existing metadata to determine how factors such as text domain, presupposition trigger, verb factivity, and context impact these agreement scores. Among the factors we study are ones associated with societal biases, such as the gender of the subject and the political ideology of the text, as LLMs have demonstrated biases that can impact their ability to make inferences (Figure 1). We focus on answering the following research questions:

I. How close are language models' projection judgments to human judgments?

II. What factors impact human-model agreement, and are any of these factors related to societal biases?

Our results indicate that changes to text domain, presupposition trigger, context window size, and verb factivity can impact human-model agreement. Further, gender and political ideology appear to influence agreement, indicating potential biases in the LLMs' judgments. Human-model agreement worsens when the subject is female compared to when the subject is male, and large differences in agreement arise for texts discussing different political ideologies across three dimensions: economic, social, and foreign. We discuss these findings in detail below, and provide recommendations for researchers who wish to use LLMs to automatically determine which content is presupposed, particularly within the space of bias mitigation.

2 Background

2.1 Presupposition and Projection

Recasens et al. (2013) found that subtle bias in text is often expressed via **presupposed** content: content that the speaker takes for granted as part of the common ground, which is necessary in order for the listeners to understand the meaning of the speaker's assertion. A simple example of presupposition is as follows:

(2) "Sally left the house to pick up some milk."

Though it is not directly stated by the speaker, most English speakers would agree that this statement entails that Sally had previously been in the house. However, when this sentence is negated, this entailment still remains:

(3) "Sally did not leave the house to pick up some milk."

Most English-speaking listeners would still assume, upon hearing this sentence, that Sally had previously been in the house, and in this case continued to be in the house instead of leaving to pick up milk. This phenomenon, where an entailed proposition is not cancelled even under an entailment-cancelling operator, is known as **projection**, which is considered by standard semantics accounts to occur as a result of that proposition being *presupposed*, and is commonly used as a diagnostic for presupposition (Heim, 1983; Van der Sandt, 1992; De Marneffe et al., 2019).

2.2 Presupposition Triggers

Presupposition is defined by the Stanford Encyclopedia of Philosophy as "the phenomenon whereby speakers mark linguistically information as being taken for granted, rather than being part of the main propositional content of a speech act." (Beaver et al., 2021). Current linguistics research on presupposition focuses largely on how presuppositions are *triggered* (Beaver et al., 2021). Presupposition triggers include *change-of-state* (see Example 2.1) and *clause-embedding predicates*. It should be noted that the presence of these triggers does not guarantee projection. For instance, the two sentences below contain the clause-embedding predicate *show*, but most would agree that *b* does not entail the proposition that racial bias is prevalent in policing (referred to as the **clausal complement**).

(4) a. He *showed* that racial bias is prevalent in policing.
b. He *did not show* that racial bias is prevalent in policing.

However, take the following sentences:

(5) He *knew* that racial bias is prevalent in policing.

(6) He *did not know* that racial bias is prevalent in policing.

Figure 2: Examples of prompts for NOPE, CommitmentBank, and MegaVeridicality. In purple in the top stanza is the sentence containing the presupposed content; the rest of the text in the top stanza is the added context. The bottom stanza contains the instructions to the model, with the hypothesis in orange.

For both sentences, most readers would conclude that the speaker is committed to the proposition that racial bias is prevalent in policing. Thus, this statement projects under the clause-embedding predicate *know*, while it does not project under the verbs *claim* or *asserted*. Since Kiparsky and Kiparsky (1970), it is commonly held that clause-embedding verbs fall into two categories: **factives**, which lexically encode presupposition, and **non-factives**, which do not. Recent works have questioned this binary distinction between factives and non-factives, pointing to examples where factives do not lexically encode presuppositions and conducting experiments that reveal high variability in projection judgments that cannot be solely attributed to the factive vs. non-factive distinction (De Marneffe et al., 2019).

3 Related Work

Several works have used crowd-sourcing to collect human projection judgments with Amazon Mechanical Turk (MTurk) (White and Rawlins, 2018; De Marneffe et al., 2019; Parrish et al., 2021), and have used this annotated data to study variations in human judgments. However, only one of these works has studied the impact of various linguistic features on *language model* projection judgments (Parrish et al., 2021), and this work did not study how LLMs behave for this task. If researchers wish to automatically evaluate the types of biases in text that occur via presupposition, the extent to which LLMs are capable of making these projection judgments should be well-understood, and potential biases introduced by LLMs when making these judgments should be documented. This is the first work to comprehensively study LLMs' projection judgments across three different human-annotated datasets, and to closely examine the factors affecting human-model agreement. We study how different linguistic features, such as text genre and trigger type, impact agreement, and examine whether sources of societal bias influence agreement. Below, we describe the three datasets we use to evaluate our baselines in more detail.

3.1 NOPE

The NOPE corpus (Parrish et al., 2021) was developed to investigate the context-sensitivity of projection judgments under different presupposition triggers. The authors extracted naturally-occurring sentences from the Corpus of Contemporary American English (COCA) (Davies, 2009) containing any of the following 10 presupposition triggers: aspectual verbs, change of state, clause-embedding predicates, clefts, comparatives, embedded questions, implicative predicates, numeric determiners, re-verbs, and temporal adverbs (examples of each of these triggers can be found in Parrish et al. (2021)). They crowdsourced entailment judgments from MTurk workers to determine for which examples *projection* occurs (§2.1).The authors also used this dataset to test language models' capabilities for inferring presuppositions. They evaluated a Bag-of-Words (BOW) model, InferSent (Conneau et al., 2017), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) on their dataset. All of these models were trained (BOW and InferSent) or finetuned (RoBERTa and DeBERTa) on the MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), ANLI (Nie et al., 2020), and FEVER

(Thorne et al., 2018) datasets before they were evaluated on the NOPE dataset. The authors found that the models exhibited especially high performance on examples with *clefts*, *numeric determiners*, and *temporal adverbs*, and struggled with *implicatives* and *clause-embedding predicates*.

3.2 CommitmentBank

The CommitmentBank dataset (De Marneffe et al., 2019) was developed to investigate the conditions under which the finite clausal compliments of clause-embedding predicates project (§2.1). Does the so-called "factivity" of an embedding predicate (§2.2) determine projection, and to what degree? The dataset consists of 1200 naturally-occurring discourse segments from three different corpora, each in a different domain: Wall Street Journal (WSJ) news articles, the fiction component of the British National Corpus (BNC), and Switchboard dialogues (SWBD). MTurk crowdworkers were hired to annotate each example based on how certain they believed the speaker was about the truth of the clausal complement (CC). Crowdworkers annotated on a scale from -3 (speaker is certain that the CC is false) to 3 (speaker is certain that the CC is true), with 0 indicating uncertainty either way. Other factors that may impact projection were also annotated, such as the lemma of the subject, temporal reference of the matrix clause ("past", "present", or "future"), and the plausibility of the CC based on the context. The authors then analyzed the effects of these factors on crowdworkers' ratings. They find that, though factives are in general more likely to be projective than non-factives, there is no distinct separation between the two. For instance, examples with the non-factive predicate "accept" are rated more projective on average than most of the "factive" verbs. They also found evidence that the tense of the predicate and person of the subject may impact projection.

3.3 MegaVeridicality

The MegaVeridicality dataset (White and Rawlins, 2018) was compiled to test the role of factivity (§2.2) and veridicality (truthfulness) in determining clause selection for verbs (the semantic interpretation of their arguments). The authors selected 517 verbs from the MegaAttitude dataset (White and Rawlins, 2016) and recruited participants on MTurk to provide veridicality ratings based on a series of frames such as "Someone {thought, didn't think} that a particular thing happened" and "Someone {was, wasn't} told that a particular thing happened". Raters were asked to answer the question *did that thing happen?* by choosing one of three response options: *yes*, *maybe or maybe not*, and *no*. For each item, 10 different ratings were given, each from a different participant. The authors found that veridicality and factivity do not serve as reliable predictors of selection.

4 Methods

Model	Pearson	Spearman	Tau
`davinci-3`	0.3899	0.4303	0.3307
`turbo-3.5`	0.4465	0.3935	0.3075
`mixtral`	0.4290	0.4022	0.2967
`llama2-70b`	0.1745	0.1905	.1430
`phi`	0.0706	0.1194	0.0838

Table 1: The Pearson, Spearman, and Kendall's Tau correlations between the average human rating and the model rating for each baseline, averaged over 3 runs.

To collect veridicality judgments from LLMs, we prompt our baselines using language similar to the directions human raters were given for their annotation task. We test on a variety of baselines, and use the highest-performing model and settings to run the remainder of our experiments. Below, we detail our prompting strategies for each dataset and describe the procedures used to choose our experimental settings.

4.1 Prompting Strategies

To prompt our baselines, we simulate the human rating tasks used for all three datasets: the NOPE dataset (Parrish et al., 2021), CommitmentBank (De Marneffe et al., 2019), and MegaVeridicality (White and Rawlins, 2018) (see §3). The prompt given to our baselines for each rating task is kept as similar as possible to the one presented to human raters in the corresponding dataset. We provide the templates, and examples, for the prompts given to the baselines in Figure 2.

4.2 Experimental Settings

We choose our baseline empirically based on agreement between the model and human raters, as well as variability between runs. We ultimately choose GPT's `text-davinci-003` model at a temperature of 0.0 and with a max token length of 5 (since all valid answers contain less than 5 tokens - see prompting strategies for details). Below, we detail this process.

Trigger	Pearson	Spearman	Tau
Aspectual Verbs	0.2162	0.2332	0.1731
Change of State	0.3099	0.3301	0.2499
Clause-Emb. Pred.	**0.7071**	**0.7058**	**0.5601**
Clefts	0.5254	0.2895	0.2363
Comparatives	0.5959	0.5928	0.4602
Embedded Q	0.2157	0.2845	0.2221
Implicative Pred.	0.2485	0.2951	0.2285
Numeric Det.	0.2814	0.2697	0.2122
Re-Verbs	0.2363	0.2949	0.2207
Temporal Adv.	0.3340	0.1714	0.1401
Overall	0.3910	0.4326	0.3324

Table 2: The Pearson, Spearman, and Kendall's Tau correlations between the average human rating and model rating for each NOPE corpus trigger type, with clause-embedding predicates yielding the highest correlation.

Agreement with human raters We start by measuring human-model agreement on the NOPE corpus for multiple LLM baselines at a temperature of 0.0.Among these baselines are two GPT models from OpenAI: GPT-3 (`davinci-3`) (Ouyang et al., 2022) and ChatGPT 3.5(`gpt-3.5-turbo`) [1], and three open-source LLMs: Meta's Llama 2 model (`llama2-70b`) (Touvron et al., 2023), Microsoft Research's `phi2` (Gunasekar et al., 2023; Li et al., 2023), and Mistral AI's Mixtral 8x7B (`mixtral`) (Jiang et al., 2024). As shown in Table 1, `gpt-3.5-turbo` has the best linear correlation with human ratings, while `davinci-3` has the best rank correlations with human ratings. As we are more interested in studying the comparative ratings between examples, we prioritize the model with the highest rank order. Thus, we use `text-davinci-003` as our baseline for the remainder of our experiments.

Variability between runs To measure variability, we run our chosen baseline, GPT-3 (`text-davinci-003`), on the NOPE dataset three times, at temperatures of 0.0, 0.25, and 0.5. We then calculate the average pairwise correlation between runs using the Pearson, Spearman, and Kendall's Tau correlation coefficients. We report our results in Table 6 in Appendix A. Though we find higher variability at higher temperatures, the correlation between runs remains above .85 at all of the tested temperature levels. As expected, variability is lowest at a temperature of 0.0 (>.98). Thus, we set the model temperature to 0 for the remainder of our experiments.

[1] https://openai.com/blog/chatgpt

	Pearson	Spearman	Tau
W/ context	0.3910	0.4326	0.3324
No Context	0.4144	0.4173	0.3216

Table 3: Pearson, Spearman, and Kendall's Tau correlations between the average human rating and the model rating with and without context; we observe slightly higher linear correlations for the model without context and vice versa for rank correlations.

5 Results

In the following, we detail our findings from experiments conducted on the three corpora described in §3 using the baseline model (GPT-3) chosen from the experiments in §4. In addition to presenting our results, we outline the main takeaways and discuss their implications.

5.1 NOPE Corpus

Errors between GPT and transformer models often do not align for different trigger types. Parrish et al. (2021) found that transformers are most accurate when classifying entailment for presupposition triggers on the NOPE dataset: *clefts*, *numeric determiners*, and *temporal adverbs*. They reported the worst performance for *clause-embedding predicates* and *implicatives*. Inversely, we find the highest correlation between human and GPT judgments for clause-embedding predicates (Table 2). GPT achieves low performance for temporal adverbs and numeric determiners, and low rank correlations from clefts. We speculate that transformers models' near-ceiling performance on these categories in the NOPE corpus may in part result from the high frequency of entailment labels in these categories and the models' tendency to predict an entailment label. We provide a more fine-grained examination of GPT's behavior in Appendix B.

Context does not have much effect on projection judgments. We also tested the effects of context on GPT's projection judgments for the examples in the NOPE corpus. To do so, we prompted GPT with 1) only the sentence containing the presupposition trigger (no context) and 2) the sentence containing the presupposition trigger, prepended with the two sentences immediately before it. As shown in Table 3, we find that trends in GPT's rating patterns for hypotheses and their negations are largely unchanged with and without context. Further, the correlations with human ratings vary

	Pearson	Spearman	Tau
Overall	0.6758	0.6846	0.5464
WSJ	0.6785	0.6809	0.5526
BNC	0.6580	0.6676	0.5307
SWBD	0.5816	0.5822	0.4600

Table 4: The Pearson, Spearman, and Kendall's Tau correlations between the average human rating and the model rating across the whole CommitmentBank dataset, and for each domain in the dataset. Wide discrepancies in agreement occur between domains.

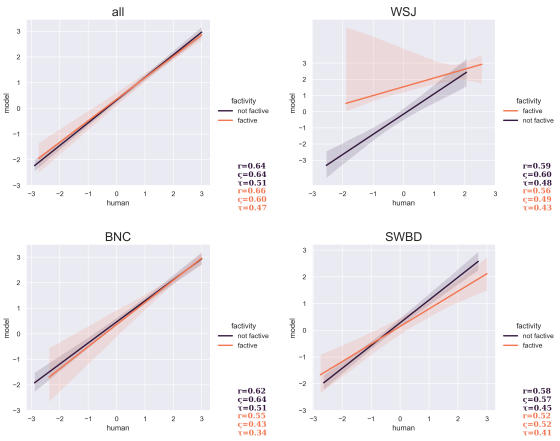

Figure 3: Human vs. model ratings for CommitmentBank data with GPT text-davinci-003 for factive vs. non-factive verbs for the whole corpus and individual domains. Factives and non-factives exhibit similar trends across the dataset as a whole, but more variation was observed within-domain, particularly for WSJ.

only slightly between the contextually-aware and context-free model (Table 2). This may indicate a lack of context-sensitivity among the examples as a whole, or it may reflect an issue with GPT when processing large contexts.

5.2 CommitmentBank

Overall, model ratings and human ratings in the CommitmentBank are more strongly correlated than those in the NOPE corpus, but statements with clause-embedding predicates as a presupposition trigger in the NOPE corpus show a slightly stronger correlation.

Text genre and style may impact human-model agreement The CommitmentBank contains texts from three different datasets: Wall Street Journal (WSJ) news articles, British National Corpus (BNC) fiction texts, and Switchboard dialogues. In addition to calculating the overall correlation between model predictions and average human ratings, we calculate the correlation for each of the domains contained in the CommitmentBank to determine whether the model is likely to agree more on certain texts. We report our results in Table 4. We find that WSJ news texts have the highest correlation between human and model ratings: .6785 for Pearson, .6809 for Spearman, and 0.5526 for Kendall's Tau. BNC fiction texts exhibit a slightly lower correlation for each metric (Pearson, Spearman, and Kendall's Tau), but within 3 points of the WSJ correlations for each metric. The Switchboard dialogues, however, exhibit the lowest correlation by a larger margin: for each metric, we report a 9 to 10-point decrease from the WSJ correlations. We speculate that the structure of the Switchboard dialogues may be less familiar to GPT than the paragraph structure of fiction or news articles, but the effects of the text's domain and structure on GPT's ability to predict implicature should be studied further.

Factive verbs often yield lower agreement than non-factive verbs, but this effect varies across domains. Given that the CommitmentBank was created to empirically study the "factive"-vs. "non-factive" distinction, we are interested in studying whether factivity may impact the relationship between human and model ratings: specifically, do models agree more with humans when the trigger is a factive verb compared to a non-factive verb? Across the whole dataset, and within each domain, we calculate the correlations between human and model judgments and plot a linear regression line for 1) non-factives and 2) factives. We use the CommitmentBank paper's lists of factive and non-factive verbs to determine factivity (De Marneffe et al., 2019). Our results can be found in Figure 3. We find that overall, there is a slightly higher linear correlation between human and model judgments for factive verbs, while a higher rank correlation is observed for non-factives. But within each domain, the correlations are lower for factives than non-factives across all metrics. These differences are often more pronounced within-domain than for the dataset as a whole. Within the WSJ texts, factives exhibit a Spearman's Rank correlation that is 11 points below that of non-factives. The BNC texts show the highest differences between factives and non-factives: 9 points for Pearson, 21 points for Spearman, and 17 points for Kendall's Tau. The SWBD texts have, at most, a 6 point difference

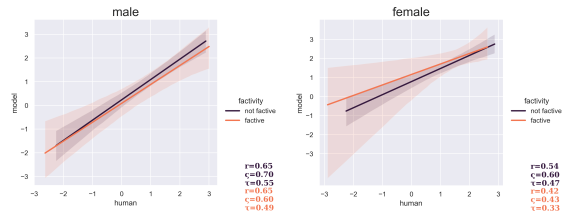

Figure 4: Human vs. model ratings for CommitmentBank data with GPT text-davinci-003 for factive vs. non-factive verbs for male and female subjects. We find lowest agreement (for each metric and verb type) for female subjects.

between factives and non-factives.

The regression line equations for factives vs. non-factives are near-identical for the dataset overall. However, we observe different trends within specific domains. Figure 3 shows slight (not statistically significant) differences between the lines of best fit for factives and non-factives for the BNC and SWBD texts. For WSJ texts, on the other hand, the lines of best fit are significantly different for factives vs. non-factives (smaller positive slope for factives). We note that a possible contributor is the small number of factive examples in the WSJ text.

The gender of the subject impacts human-model agreement. To determine whether our baseline model exhibits any signs of bias, we start by looking at a relatively easy-to-identify characteristic within our data: gender of the subject. We calculate the correlations between human and model predictions, and plotted regression lines, for factive vs. non-factive verbs when the gender of the subject is specified as female vs. male. To get these results, we used a simple heuristic and looked at the lemma of the subject; if the lemma was “she", we marked the subject as female, while if the lemma was “he", we marked the subject as male (if neither, the example was not used for either category). We chose this heuristic to ensure that the subjects would be unambiguously read as female or male for the human rater and the language model. In Figure 4, we compare the results for 1) the whole dataset, 2) male subjects, and 3) female subjects. We find a much lower overall correlation between human and model ratings for female subjects than for male subjects, which is especially pronounced for factives when compared to nonfactives. In particular, we observe the model is prone to predicting neutral to positive entailment labels for female subjects, even in cases where human raters have determined that the speaker is certain about a statement being false, and that this trend is less pronounced for male subjects (see Figure 8 for distributions). Thus suggests that GPT may be less inclined to predict that a female subject believes a statement to be false than a male subject, and that its predictions for male subjects are more aligned with the ground truth.

The political ideology discussed may impact human-model agreement. We also wish to examine whether more subtle, difficult-to-detect sources of bias may influence the LLM’s judgments. To do so, we use the set of WSJ articles in the CommitmentBank and run the political ideology classifier developed by (Sinno et al., 2022) on the concatenation of the context sentences and the target sentence. This classifier predicts the political ideology under discussion in the text (left, right, or neutral) across three different dimensions: economic, social, and foreign. We calculate the correlation between human and model judgments for texts labeled as ideologically left, right, and neutral for each dimension and compare these correlations. Our results can be found in Table 5. We find that, for each dimension, only a few examples are tagged as right-leaning, while more examples are tagged as left-leaning and most examples are tagged as neutral. We also find that the model exhibits the lowest agreement with human judgments for examples labeled as economic right; for each coefficient, the model judgments are negatively correlated with human judgments. By contrast, for examples labeled socially right-leaning, the model is more strongly correlated with human judgments than for neutral or left-leaning examples. For each dimension, the model is more highly correlated with human judgments for left-leaning examples than for neutral examples, despite the presence of more neutral examples in the dataset. This suggests that the model’s inferences may vary depending on the political ideology under discussion, and that these variations may exhibit different patterns for different political dimensions.

5.3 MegaVeridicality

GPT is less likely to predict entailment than humans when given very generic propositions To isolate the effects of verb properties on the acceptability of statements, the authors of the MegaVeridicality corpus (White and Rawlins, 2018) include as little semantic content as possible in their examples. As such, this dataset serves as a useful testing

Dimension	Lean	Pear.	Spear.	Tau	#
Economic	Right	-.1642	-.2412	-.1793	7
	Left	.6517	.6337	.5395	19
	Neutral	.5990	.5955	.4729	78
Social	Right	.7715	.6156	.5270	5
	Left	.5841	.5766	.4667	33
	Neutral	.5695	.5553	.4560	66
Foreign	Right	.6592	.6801	.5446	11
	Left	.6721	.6525	.5323	38
	Neutral	.5032	.4926	.3998	55

Table 5: Correlations between human and model agreement given ideological polarization labels for WSJ texts in each dimension. Examples discussing left-leaning topics produce higher agreement than examples marked as neither left nor right, despite the latter containing more examples for each dimension.

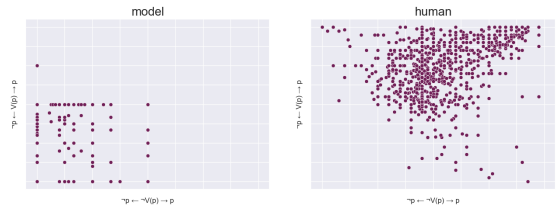

Figure 5: Average model (left) vs. human (right) veridicality judgments for each verb. The model is much less prone to predicting that a statement is veridical than that it is not veridical, even for verbs considered "factive" in the MegaVeridicality corpus.

ground to examine how GPT behaves when given very little information besides the clause embedding predicate (e.g. "Someone *knew* that something happened"). To compare model behavior to human behavior for this dataset, rather than calculate correlation between human and model ratings (since there are only three possible labels), we compare the model's answers to *gold labels*, derived by taking the majority label assigned by annotators (when there is no majority, that examples is discarded). We find that, in comparison to humans, GPT is much more likely to answer *No* than *Yes* to the question of "did that thing happen"? This is shown in Table 7 in Appendix D, where the "Yes" label has high precision and low recall and the "No" label has low precision and high recall. This can also be visualized more clearly in Figure 5. This pattern was not as evident in the NOPE dataset (6) or the CommitmentBank dataset (3), both of which contain more specific, contextually grounded clausal complements found "in-the-wild".

Using gendered pronouns for subjects may influence GPT's agreement with human ratings Because a portion of the MegaVeridicality dataset denotes its subjects using only the indefinite, genderless pronoun "Someone", it is trivial to conduct experiments where the gender of the subject is changed in the prompt and compare model results to the human annotations given for the original example with "Someone". We experiment with substituting "Someone" with "A man" or "A woman" for each example constructed from the [NP _ that S] frame. For each example, we calculate the accuracy and correlation between the model prediction on the altered example and the average human label for the original example using "Someone". As is shown in Table 8, differences in model performance were observed for male-gendered examples vs. female-gendered examples. We (unsurprisingly) find that the model performs best when given the same prompt as the humans are given, with "Someone" as the subject. When changing the subject to "a man" in the model prompt, we observe a slight drop in accuracy and a much larger decrease in correlation. Notably, when subject in the model prompt is changed to "a woman", the accuracy and correlation between model and human ratings drop by several points compared to "a man". These results, along with the CommitmentBank results (Figure 4), heavily indicate that LLMs' judgments are closer to human judgments when the subject is male rather than female.

6 Conclusion

In the above, we provide the first comprehensive set of experiments comparing human projection judgments with LLM projection judgments, to determine how reliably LLMs can be used to identify presupposed content. We also examine how factors such as specificity, text domain, presupposition trigger type, and word factivity impact agreement, and find that changes in these variables can heavily impact on how closely the model's predictions align with humans'. Additionally, we find evidence that changes to gender and political ideology may impact the model's agreement, suggesting that certain social biases may impact the model's judgments. We thus urge practitioners using language models to perform these inferences at a large scale to evaluate their systems carefully, and to determine the conditions under which they succeed and whether they may reflect existing societal biases.

7 Limitations

Because these datasets were manually annotated, with each example annotated by multiple raters, they are relatively small, on the order of thousands of examples. The set of Wall Street Journal articles in the CommitmentBank is even smaller. Thus, our findings, particularly on bias, should be investigated on a larger scale to determine whether they hold for larger sets across additional text domains.

8 Ethics

In this work, we evaluate the performance of LLMs on existing datasets, and do not release any new publicly-available datasets with gold labels. We also do not use, or release, any LLMs that have previously not been released to the public. We do study the use of LLMs to detect biases that arise from presupposition, and release our prompting techniques for these experiments. However, given that our findings indicate potential biases in LLMs' projection judgments, we urge practitioners to study this technique further before relying on automatic methods alone to detect epistemological biases. If practitioners are to use LLMs to make claims about biases in text, they should also use manual evaluation techniques, and should carefully study the agreement between LLMs and humans, as well as the factors that impact this agreement.

Acknowledgements

References

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2021. Presupposition. In *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14, 2 (2009), 159–190.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, Vol. 23. 107–124.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 345–363. https://doi.org/10.18653/v1/2021.emnlp-main.29

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks Are All You Need. *arXiv preprint arXiv:2306.11644* (2023).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).

Irene Heim. 1983. On the projection problem for presuppositions. *Formal semantics–the essential readings* (1983), 249–260.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG]

Paul Kiparsky and Carol Kiparsky. 1970. FACT. In *Progress in Linguistics: A Collection of Papers*. De Gruyter Mouton, Berlin, Germany, 143–173. https://doi.org/10.1515/9783111350219.143

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463* (2023).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A Corpus of Naturally-Occurring Presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 349–366. https://doi.org/10.18653/v1/2021.conll-1.28

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659. https://aclanthology.org/P13-1162

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political Ideology and Polarization: A Multi-dimensional Approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 231–243. https://doi.org/10.18653/v1/2022.naacl-main.17

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

Rob A Van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of semantics* 9, 4 (1992), 333–377.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of S-selection. In *Semantics and linguistic theory*, Vol. 26. 641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th annual meeting of the north east linguistic society*, Vol. 3. 221–234.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Temp.	Pearson	Spearman	Tau
0.0	0.9986	0.9897	0.9863
0.25	0.9931	0.9533	0.9328
0.50	0.9814	0.9112	0.8656

Table 6: The average Pearson, Spearman, and Kendall's Tau pairwise correlations across 3 runs for GPT-3 text-davinci-003, with highest values at a temperature of 0.

A Experimental Details

Variability between runs

B NOPE Results

Trends for model predictions by trigger type To investigate possible causes of GPT's varied performances for different trigger types, we plot human vs. model predictions for each trigger type, along with a regression line, in Figure 6. We find that clefts, numeric determiners, and temporal adverbs, GPT tends to be overconfident in its implicature judgments, whereas for clause-embedding predicates, GPT is more successful at predicting low values in cases where human raters assign low values. Further, as we discuss in more detail below, GPT has a tendency to cluster its ratings at the midpoint and extremes, particularly in the 90-100% range.

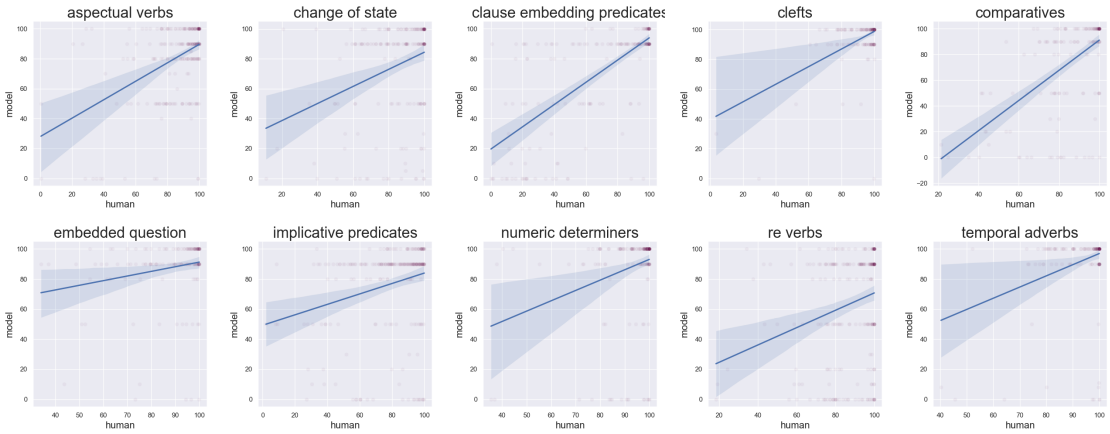

Figure 6: Regression lines fitted to the model predictions as a function of human judgments for each NOPE trigger type.

Class	Precision	Recall	F1	Support
Yes	0.71	0.01	0.02	491
Maybe	0.35	0.42	0.38	251
No	0.02	0.82	0.04	11

Table 7: Precision, recall, F1, and support for GPT-3 text-davinci-003 at a temperature of 0 when compared to the gold MegaVeridicality labels (obtained by taking the majority label between the human raters when one existed; when one did not, the label was thrown out).

Distributions of model predictions by trigger type As can be seen in Figure 6, in our experiments GPT had a tendency to cluster its ratings at the midpoint and extremes: around 0%, 50%, and 100%. Further, GPT predictions were heavily concentrated in the 90-100% range (Figure 7). Thus, rather than predict 0% or 50% for examples that averaged a 55-85% rating from humans, GPT may have opted to instead predict 90%, 95%, or 100%. It is of note that GPT's predictions are not normally distributed, as one might expect for human ratings; they are either skewed entirely towards 100% or bi- or tri-modal. This suggests that GPT may default to picking the extreme values in this type of task.

C CommitmentBank

D MegaVeridicality

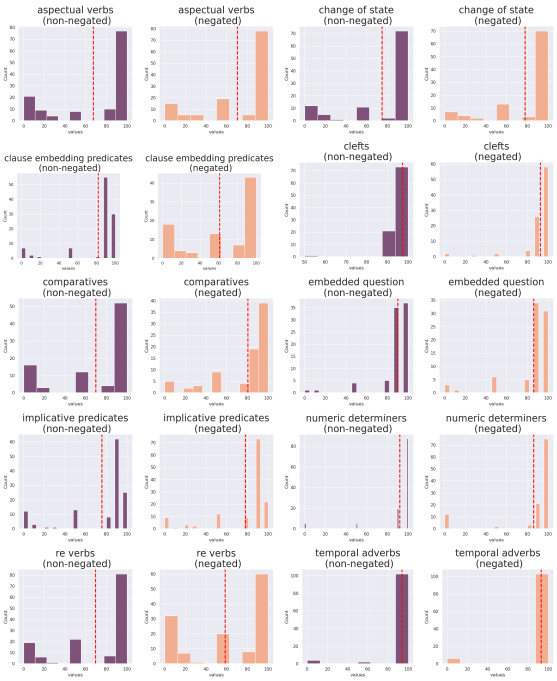

Figure 7: Distributions of entailment judgments for negated and non-negated statements for each trigger type in the NOPE corpus. Clause-embedding predicates yield the largest difference in mean between negated and non-negated statements. These results mirror the results of the NOPE corpus.

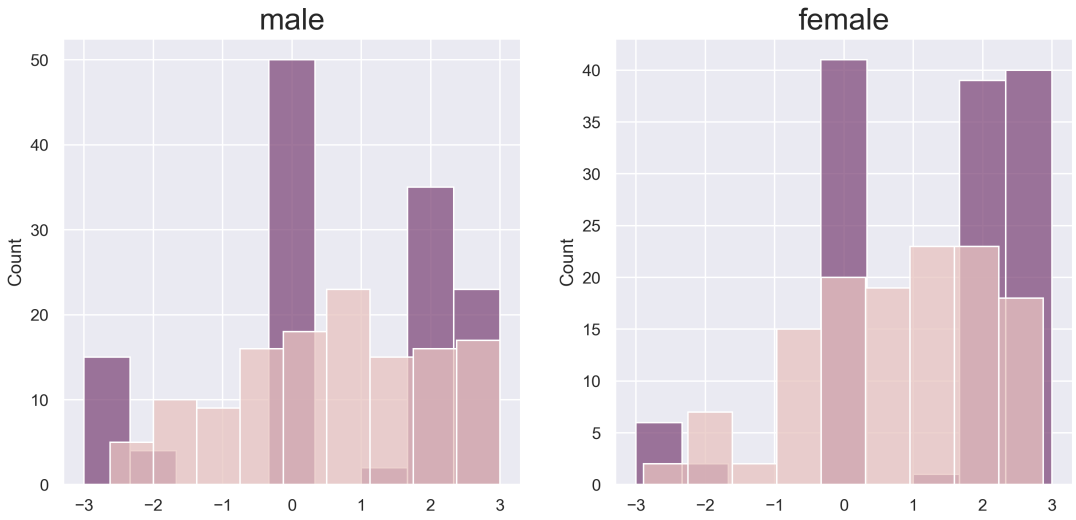

Figure 8: Distributions of model ratings (purple) compared to human ratings (light) for male vs. female subjects. The model predicts false values less often for female subjects than male subjects.

	Accuracy	**Pearson**	**Spearman**	**Kendall's Tau**
Unchanged	.3647	.4187	.4302	.3622
Someone → **a man**	.3040	.1642	.1410	.1196
Someone → **a woman**	.2808	.1169	.0981	.0833

Table 8: Correlations between model judgments (for GPT-3 text-davinci-003 at a temperature of 0) and human judgments when the prompt given to the model 1) was unchanged, 2) replaced the word "someone" with "he", and replaced the word "someone" with "she".

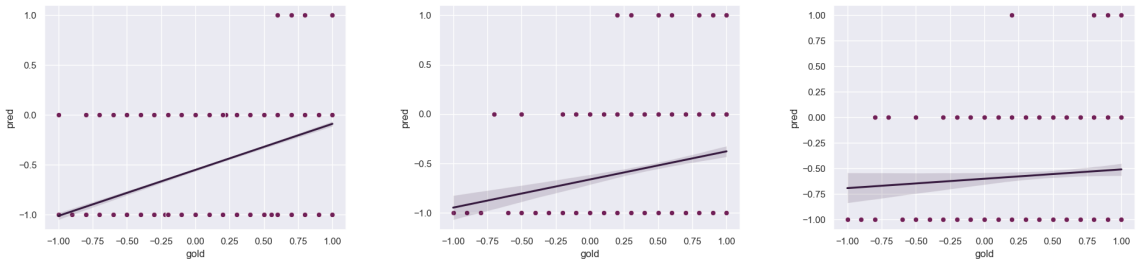

Figure 9: Model vs. average human judgments for each example when model prompt is unchanged from human prompt (left), model prompt replaces "someone" with "he" (center), and model prompt replaces "someone" with "she" (right).