Do Large Language Models Show Negation Bias? A Replication of Beukeboom et al. (2010)

Leonie-Marie Allar (leonie-marie.allar@student.uni-tuebingen.de), Ann-Sophie Berg, Barbara Kaup & Francesca Capuano
Department of Psychology, University of Tübingen

Abstract

LLM bias research has largely focused on content-level associations, overlooking biases arising from linguistic form and pragmatic choice. One such phenomenon is the *negation bias*, where negated expressions are preferentially used for stereotype-inconsistent behavior and trigger systematic inferences. While LLMs' handling of negation in semantic tasks is well studied, it is unclear whether they reproduce these pragmatic inferences in social contexts. We replicate Beukeboom et al. (2010) with Mistral-7B-Instruct-v0.2, varying the valence and polarity of trait descriptions. The model inferred more negative expectations from negated negative traits (e.g. *not stupid*) and more positive expectations from negated positive traits (e.g. *not smart*), and showed stronger situational but weaker dispositional attributions. These results mirror human negation-bias patterns, suggesting that instruction-tuned models can produce pragmatic-like bias in judgment tasks. The findings underscore the need to extend LLM bias evaluation beyond content to linguistic form.

Keywords: Large Language Models; Negation Bias; Pragmatic Inference; Linguistic Bias; Stereotypes.

Introduction

Large language models (LLMs) are well-documented to exhibit social biases, particularly those associated with race, gender, and other demographic categories (Bender et al., 2021; Ferrara, 2024; Gallegos et al., 2023). These biases typically manifest as associational patterns between social groups and attributes (e.g., linking women with nurturing roles) (Caliskan et al., 2017; Dhamala et al., 2021; Nadeem et al., 2021). The consequences of such biases are socially significant: LLMs can reproduce and amplify discriminatory outputs and harmful stereotypes (Benjamin, 2019; Ferrara, 2024; Lee et al., 2024). Content-level biases are relatively straightforward to detect, appearing in word embeddings, completion patterns, classification outcomes, and sentiment asymmetries (Caliskan et al., 2017; Dhamala et al., 2021; Nadeem et al., 2021). They arise from structural inequalities present in the large-scale corpora used to train LLMs that may be unrepresentative, incomplete, or historically skewed. Consequently, mitigation strategies have focused on fairness-oriented interventions, including data balancing, embedding debiasing, toxicity filtering, and demographic-parity tuning (Da et al., 2024; Ferrara, 2024; Gallegos et al., 2023).

However, social biases do not manifest exclusively through propositional content. Beyond overt stereotypes, humans also exhibit linguistically mediated biases, which operate at the level of language structure and pragmatic choice rather than explicit semantic content. One example is the *negation bias* (Beukeboom et al., 2010, 2020) that refers to speakers' preferential use of negations when describing stereotype-inconsistent behaviors. For instance, descriptions such as *He is not smart* are more likely to occur when portraying a professor in stereotype-inconsistent scenarios (e.g., displaying low intelligence) than affirmative antonyms like *He is stupid*. Conversely, in stereotype-consistent scenarios, such as a soccer hooligan displaying low intelligence, speakers prefer affirmative descriptors like *He is stupid* over negations such as *He is not smart*. The relevance of the negation bias lies in its subtlety and structural embedding within language use. Unlike content-level biases, which involve explicit associations between social groups and attributes, linguistic biases operate at the level of form rather than content. By shaping how information is framed, they influence recipients' interpretations and preserve stereotypical beliefs even when the explicit content appears neutral or non-stereotypical. For instance, a speaker may describe an assertive woman as *She is not passive* rather than *She is assertive*. Although both statements convey a similar trait, the use of negation subtly marks the behavior as unexpected, presenting it as an exception to the prevailing stereotype and thus indirectly reinforcing the recipient's stereotypic expectations.

Negation is not only used to communicate stereotype-inconsistent behavior. From a pragmatic perspective, negation often functions to refute an implicit assumption held by the listener (Bonnefon & Villejoubert, 2007; De Villiers & Flusberg, 1975; Givón, 1978; L. Horn, 1989; Johnson-Laird & Tridgell, 1972; Wason, 1965). This pragmatic use is particularly salient in the context of plausible denial, in which the negated proposition corresponds to what is expected or assumed to be likely (Wason, 1965). For example, a statement such as *The train was not late this morning* is more naturally produced when the train is usually late rather than when it is typically on time, as the negation functions to correct the implicit and inaccurate assumption that the train was late as usual. Just as negation signals expectation violations in everyday events, in social interactions it serves to convey that a person's behavior is unexpected (Beukeboom et al., 2010, 2020; Wigboldus et al., 2006). Accordingly, the negation bias reflects the linguistic-pragmatic tendency to employ negated expressions precisely when the denied state of affairs runs

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

counter to an established expectation.

Listeners interpret such negated statements as deliberately informative, inferring that the behavior deviates from what is normally expected or socially typical (Bonnefon & Villejoubert, 2007; L. Horn, 1989; Nordmeyer & Frank, 2015). This systematic interpretive process is grounded in general principles of cooperative communication. According to Grice (1975)'s maxims, speakers are expected to provide informative utterances while avoiding unnecessary prolixity. When a speaker opts for a negated formulation rather than a more economical affirmation, listeners infer that this choice is deliberate and contextually meaningful: the negation signals that the described behavior deviates from what is normally expected (see L. R. Horn, 1984), thereby fostering stereotype-preserving inferences and contributing to the maintenance of underlying stereotypic beliefs. In other words, conversational pragmatics makes the expectation-violating inference systematic.

Importantly, Beukeboom et al. (2010) show that marking a behavior as exceptional has consequences beyond trait inference: it also affects how recipients explain the causes of that behavior. Because negated formulations pragmatically signal a violation of expectations, recipients are inclined to interpret the behavior as atypical and therefore less diagnostic of the target's underlying disposition. Rather than revising their beliefs about the target's character, recipients are more likely to attribute the behavior to situational factors that temporarily overrode the target's usual tendencies. For example, describing an assertive woman as *not passive* may lead listeners to infer that her behavior was prompted by unusual circumstances rather than reflecting a stable personality trait. In this way, negated descriptions systematically foster *situational attributions* over *dispositional* ones.

A pragmatic account of negation bias explains both production and comprehension at a functional level: speakers use negation to signal mild scalar deviations in stereotype-inconsistent contexts, and recipients reliably interpret it as such due to shared conventions. However, this account does not mechanistically unpack the real-time cognitive processes that sustain the stereotypic residue during interpretation. Psycholinguistic research suggests that negation is processed incrementally: comprehenders initially activate a representation of the negated state of affairs before integrating the negation operator (Hasson & Glucksberg, 2006; Kaup et al., 2005; Kaup et al., 2006, 2007). This initial activation may linger transiently during interpretation, rather than being immediately suppressed, yielding weakened or intermediate interpretations rather than categorical opposites. One line of work links this processing profile to a *mitigation effect*, which observes that negated scalar expressions are reliably understood as conveying moderate rather than extreme meanings (Fraenkel & Schul, 2008; Giora, 2007; Giora et al., 2004, 2005). For example, *the coffee is not hot* is typically interpreted as indicating a moderate temperature, rather than strictly cold. In social contexts, this processing dynamic might have important consequences: the stereotype-consistent concept could be activated early and only partially attenuated, so negated descriptions (e.g., *she is not passive*) may continue to evoke the stereotypic expectation they ostensibly deny. Thus, negation does not merely signal expectation violation pragmatically; it might also sustain underlying stereotypic representations cognitively.

These findings highlight the cognitive and pragmatic complexity of negation and its relevance for examining the subtle ways bias operates in language. Many pragmatic aspects of language, including complex uses of negation (e.g. to signal expectation violations), have been argued to rely on sensitivity to shared norms, expectations, and speakers' perspectives. Some accounts relate these capacities to broader cognitive systems that support reasoning about others' beliefs and intentions (Theory of Mind, e.g., Cuccio (2012), Rubio-Fernandez (2021), and Schindele et al. (2008); cf. Bosco et al. (2018)). While fine-tuning can encourage large language models to approximate such inferences behaviorally, these models do not explicitly represent communicative intentions or social expectations in the way proposed by human cognitive accounts (Mistral AI, 2026; Ouyang et al., 2022; Vaswani et al., 2017). Instead, pragmatic behavior, when it emerges, is induced indirectly through training objectives and exposure to patterns of use.

Recently, Wang et al. (2025) investigated negation bias in LLMs using a probability-based approach. They constructed stereotypical and anti-stereotypical sentences in affirmative and negated form and measured model surprisal across several architectures. Their findings show that negation bias can be detected in model-internal probability distributions, especially in autoregressive models. The present study complements this work by adopting a behavioral elicitation paradigm closely aligned with the original human experiments by Beukeboom et al. (2010). Crucially, whereas probability-based approaches establish whether negation affects model-internal expectations, they do not directly test whether these expectations are used to generate socially meaningful inferences such as attribution and expectation judgments. The present study targets this inferential level explicitly.

The primary objective of our study is to replicate the findings of Beukeboom et al. (2010) using an LLM, testing whether it draws analogous inferences about speakers' expectations from negated versus affirmative descriptions of stereotype-consistent and -inconsistent behaviors, and how it attributes those behaviors causally. If the LLM shows negation bias, we expect positive performances described in a negated way (e.g., *not stupid*) to yield more negative/less positive inferred expectations than affirmative alternatives (e.g., *smart*), while negative performances described in a negated way (e.g., *not fast*) to yield more positive/less negative expectations than affirmatives (e.g., *slow*). For attributions, we expect negated descriptions to elicit more situational explanations and fewer dispositional ones.

Method

Model Sessions

Following Brysbaert (2019) recommendation of at least 110 observations for .80 power to detect a medium effect, we prompted independent chat sessions of Mistral-7B-Instruct-v0.2 via Hugging Face until obtaining N = 110 valid runs. Seven sessions were excluded: four for nonsensical scale responses and three because the model stated that it lacked sufficient context. We focused on Mistral-7B-Instruct-v0.2 as an open-weight, instruction-tuned chat model. The choice was motivated by the model's accessibility and relevance as a deployed interaction format. We therefore treat the present study not as evidence about LLMs in general, but as a proof-of-concept that pragmatic bias can emerge in explicit judgment settings.

The study was preregistered. Data and analysis script are available on OSF.

Materials

We used the English translation of the items by Beukeboom et al. (2010) as prompts. All items had the same structure. First, there was a brief introduction of the scenario (e.g., *A teacher has a pupil called Hans who has completed an assignment with him. A colleague asks him how Hans did*). This was followed by a direct-speech statement in which the speaker described the target's behavior (e.g., *The teacher says about Hans: "He did well. He completed the assignment in a smart way"*). Three distinct scenarios were constructed, each featuring a different speaker–target pair: (1) a teacher describing a student's performance on an assignment, (2) a woman describing her brother's assistance in organizing an event, and (3) a sports coach describing an athlete's performance in a race. We manipulated the linguistic description of the target's behavior along two dimensions: behavioral valence and negation. The valence of the described behavior was either positive (e.g., *He completed the assignment in a smart way*) or negative (e.g., *He completed the assignment in a stupid way*). Independently, the description was phrased either in an affirmative form (e.g., *He completed the assignment in a smart way*) or in a negated form (e.g., *He did not complete the assignment in a stupid way*). Crossing these factors resulted in four sentence types within each scenario (positive-affirmative, positive-negated, negative-affirmative, negative-negated).

Procedure

At the beginning of each prompt, the model received an explicit instruction to constrain its responses ("Restate the questions in order and limit the answer to a single number on the indicated scales"). Each chat session was presented with the full set of prompts, comprising three distinct scenarios, each instantiated in four variants. Prompts were presented in blocks of three, with one prompt per scenario per block. The order of scenarios remained fixed throughout the procedure. Within each scenario, one of the four variants was randomly selected without replacement; that is, once a specific variant had been used, it was not presented again within the same chat session. This process continued until all 12 prompts had been presented, yielding a total of four blocks. After each prompt, the model was presented with five questions, which were administered in a fixed order across all sessions. To assess the inferred a priori expectations regarding the target's behavior, the model was asked two separate questions: "To what extent did [speaker] have an a priori negative expectancy about the performance of [target]?" and "To what extent did [speaker] have an a priori positive expectancy about the performance of [target]?" Both questions were answered on a 7-point Likert scale ranging from 1 = *not at all* to 7 = *very strongly* and were evaluated independently. To assess dispositional and situational attributions in relation to the target's behavior, three additional questions were administered. The first two were, "To what extent is the behavior due to what [target] is like as a person?" and "To what extent is this behavior due to circumstances in this specific situation?" These questions were again answered on a 7-point Likert scale ranging from 1 = *not at all* to 7 = *very strongly*. The third question provided an overall attribution judgment, asking the model to locate the described behavior on a continuum from situational to dispositional causation: "To what extent is this behavior due to the situation or the person?" Responses were given on a 7-point bipolar Likert scale ranging from -3 = *completely due to the situation* to 3 = *completely due to the person*. This fifth attribution item was included to match Beukeboom et al. (2010) attribution logic, providing a forced relative judgment between situational and dispositional causation.

Design

Our study employed a 2x2 factorial design with *description type* (affirmative vs. negated) and *valence* (positive vs. negative) as within-subjects factors, resulting in four experimental conditions. We investigated five dependent variables, each corresponding to one of the five post-prompt questions: (1) inferred a priori negative expectation, (2) inferred a priori positive expectation, (3) dispositional attribution, (4) situational attribution, and (5) overall attribution.

Results

Inferred Negative Expectations

We found a significant main effect of the description type. When the LLM encountered a negated description of a behavior, it inferred a more negative a priori expectation of the target (M = 3.13, SD = 1.65) than when it encountered affirmatively described behavior (M = 2.57, SD = 1.91), F(1, 109) = 61.62, $p < .001$, $\eta_p^2 = .36$. We also found a significant main effect of valence of the behavior on inferred a priori negative expectations. When the LLM encountered behavior with a negative valence, it inferred a more negative a priori expectation of the target (M = 3.87, SD = 1.62) than when it encountered positive behavior (M = 1.82, SD = 1.34), F(1, 109) = 256.60, $p < .001$, $\eta_p^2 = .70$. Lastly, we found a significant interaction effect between the description type and the valence of the be-

havior, F(1, 109) = 60.83, p < .001, η_p^2 = .36 (see Fig. 1). For positive behaviors, negated descriptions led to substantially higher inferred negative expectations (M = 2.45, SD = 1.52) compared to their affirmative counterparts (M = 1.20, SD = 0.74). In contrast, for the negative behaviors, negated descriptions (M = 3.80, SD = 1.50) yielded lower inferred negative expectations than their affirmative counterparts (M = 3.94, SD = 1.73), though this difference was notably smaller.

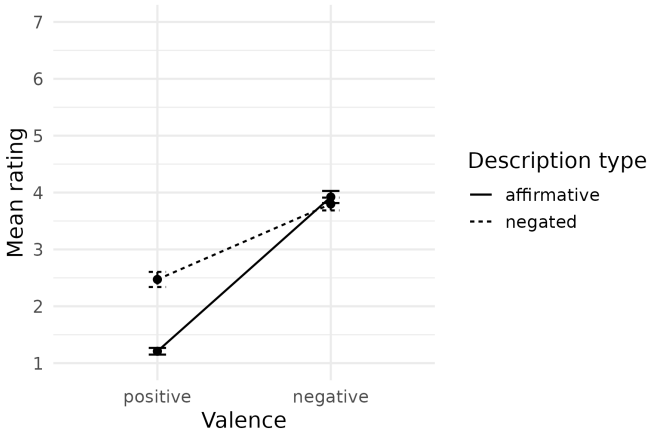

Figure 1: A priori negative expectation ratings.
Note. The error bars show the standard error.

Inferred Positive Expectations

We found a significant main effect of the description type. When the LLM encountered a negated description of a behavior, it inferred a less positive a priori expectation of the target (M = 2.40, SD = 1.48) than when it encountered an affirmatively described behavior (M = 3.35, SD = 2.08), F(1, 109) = 226.12, p < .001, η_p^2 = .67. We also found a significant main effect of the valence of the behavior on inferred a priori positive expectations. When the LLM encountered behavior of a negative valence, it inferred a less positive a priori expectation of the target (M = 2.25, SD = 1.60) than when it encountered positive behavior (M = 3.51, SD = 1.91), F(1, 109) = 142.18, p < .001, η_p^2 = .57. Lastly, we found a significant interaction effect between the description type and the valence of the behavior, F(1, 109) = 350.39, p < .001, η_p^2 = .76 (see Fig. 2). For positive behaviors, negated descriptions led to substantially lower inferred positive expectations (M = 2.41, SD = 1.31) compared to their affirmative counterparts (M = 4.58, SD = 1.79). In contrast, for the negative behaviors, negated descriptions (M = 2.39, SD = 1.63) yielded higher inferred positive expectations than their affirmative counterparts (M = 2.11, SD = 1.55), though this difference was notably smaller.

Dispositional Attribution Judgments

We found a significant main effect of the description type on the attribution judgments. When the LLM encountered a negated description of a behavior (M = 3.26, SD = 1.62), it was less likely to attribute it to the target's disposition than

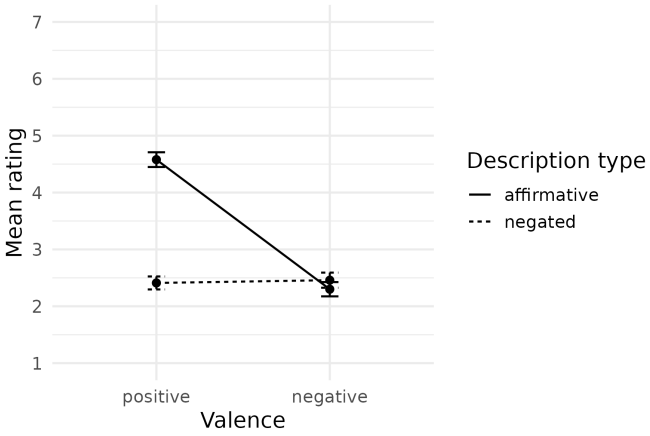

Figure 2: A priori positive expectation ratings.
Note. The error bars show the standard error.

when the description was affirmative (M = 4.46, SD = 1.93), F(1, 109) = 629.10, p < .001, η_p^2 = .85. We found a significant main effect of valence on the dispositional attribution judgments. When the LLM encountered a behavior with a negative valence (M = 3.92, SD = 1.78), it was more likely to attribute it to the person than when behavior had a positive valence (M = 3.81, SD = 2.00), F(1, 109) = 3.99, p < .05, η_p^2 = .04. We also found a significant interaction effect between description type and valence, F(1, 109) = 206.67, p < .001, η_p^2 = .65 (see Fig. 3). For positive behaviors, affirmative descriptions led to substantially higher dispositional attributions (M = 4.76, SD = 2.00) than negated descriptions (M = 2.83, SD = 1.43). For negative behaviors, affirmative descriptions (M = 4.15, SD = 1.84) also resulted in higher dispositional attributions than negated descriptions (M = 3.68, SD = 1.68), but this difference was notably smaller.

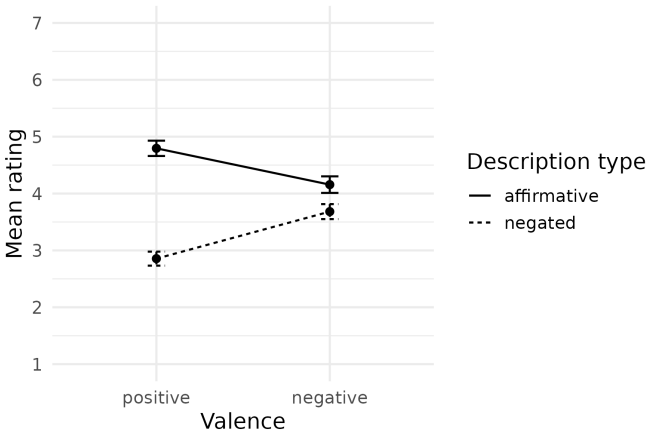

Figure 3: Dispositional attribution ratings.
Note. The error bars show the standard error.

Situational Attribution Judgments

We found a significant main effect of the description type on the attribution judgments. When the LLM encountered a negated description of a behavior (M = 3.12, SD = 1.38), it was more likely to attribute it to the circumstances than when the description was affirmative (M = 2.28, SD = 1.12), F(1, 109) = 173.95, p < .001, η_p^2 = .61. We also found a significant main effect of valence on the situational attribution judgments. When the LLM encountered a behavior with a negative valence (M = 2.88, SD = 1.31), it was more likely to attribute it to the circumstance than when behavior had a positive valence (M = 2.51, SD = 1.31), F(1, 109) = 25.19, p < .001, η_p^2 = .19. Finally, we again found a significant interaction effect between description type and valence, F(1, 109) = 47.86, p < .001, η_p^2 = .31 (see Fig. 4). For positive behaviors, negated descriptions led to substantially higher situational attributions (M = 3.11, SD = 1.48) than affirmative descriptions (M = 1.93, SD = 0.74). For negative behaviors, negated descriptions (M = 3.13, SD = 1.26) also resulted in higher situational attributions than affirmative descriptions M = 2.62, SD = 1.31), but this difference was notably smaller.

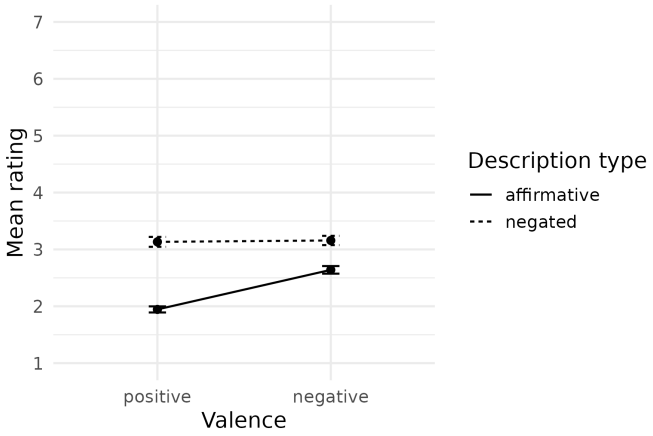

Figure 4: Situational attribution ratings.
Note. The error bars show the standard error.

Overall Attribution Judgments

Lastly, we found a significant main effect of description type on the overall attribution rating. When the LLM encountered a negated description of a behavior (M = 0.67, SD = 1.23), it attributed it more situationally than when the description was affirmative (M = 1.39, SD = 1.40), F(1, 109) = 157.30, p < .001, η_p^2 = .59. We also found a significant main effect of valence on the overall attribution rating. When the LLM encountered a behavior with a negative valence (M = 0.72, SD = 1.50), it attributed it more situationally than when it encountered a behavior with a positive valence (M = 1.34, SD = 1.12), F(1, 109) = 35.07, p < .001, η_p^2 = .24. We also found a significant interaction effect between description type and valence, F(1, 109) = 73.44, p < .001, η_p^2 = .40 (see Fig. 5). For positive behaviors, affirmative descriptions led to substantially higher dispositional attributions (M = 1.89, SD = 1.03) than negated descriptions (M = 0.79, SD = 0.93). For negative behaviors, affirmative descriptions (M = 0.88, SD = 1.54) also resulted in higher dispositional attributions than negated descriptions (M = 0.56, SD = 1.45), but this difference was notably smaller.

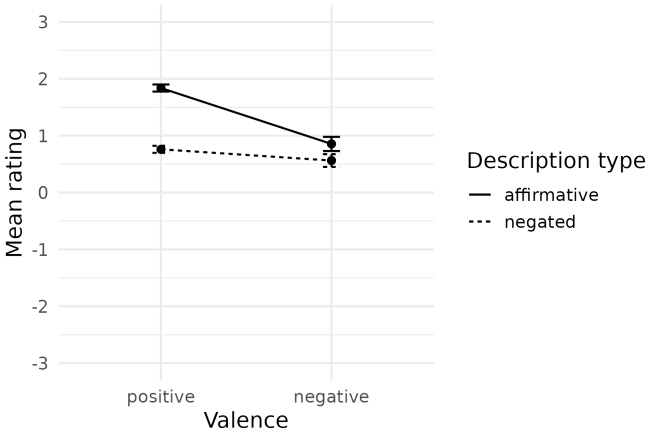

Figure 5: Overall attribution ratings.
Note. The error bars show the standard error.

Discussion

Whereas existing LLM bias research overwhelmingly targets semantic associations (Caliskan et al., 2017; Dhamala et al., 2021; Nadeem et al., 2021; Zhao et al., 2025), far less work explores whether models capture the pragmatic inferential processes that sustain stereotypes implicitly through linguistic form, even when propositional content appears neutral. We replicated Beukeboom et al. (2010) with an LLM and found the predicted interaction between description type and valence: when negation was used to describe a target's negative behavior, the LLM inferred that the speaker had a more positive a priori expectation. Inversely, when negation was used to describe a target's positive behavior, the LLM inferred that the speaker had a more negative a priori expectation. Additionally, we were able to replicate the finding that negated descriptions led to more situational attributions compared to affirmative descriptions. Thus, all hypotheses were supported.

Beyond the specific negation-bias pattern, the model's responses also resemble broader psycholinguistic regularities in negation processing. First, negated descriptions generally yielded interpretations that fell between the corresponding affirmative poles (e.g., *smart* vs. *stupid*), supporting the *mitigation hypothesis*, according to which negation weakens rather than fully reverses meaning (Fraenkel & Schul, 2008; Giora, 2007; Giora et al., 2005). However, this mitigation effect was not symmetric across valence. Crucially, for negative behaviors, negated descriptions (e.g., *not smart*) produced inferences that were numerically very close to their affirmative counterparts (e.g., *stupid*), suggesting that negation in this context approximates the opposite pole. In contrast,

for positive behaviors, negated descriptions (e.g., *not stupid*) remained substantially distinct from affirmative forms (e.g., *smart*), indicating a more attenuated, intermediate interpretation. This pattern departs from a purely symmetric weakening account and instead reveals a systematic valence asymmetry in how negation is interpreted. The near-equivalence observed for negative behaviors resembles *negative strengthening*, whereby negated adjectives are interpreted more strongly than their literal meaning (Gotzner & Mazzarella, 2021; L. R. Horn, 1984). A common explanation for this valence asymmetry appeals to politeness and face management: negating a positive evaluation (e.g., *not smart*) can function as a face-saving alternative to a blunt negative term (e.g., *stupid*), leading listeners to infer a stronger underlying negative evaluation. Conversely, negating a negative evaluation (e.g., *not stupid*) does not serve the same politeness function and is therefore less likely to trigger strong pragmatic enrichment, resulting in weaker shifts away from the literal meaning.

Importantly, interpreting these findings as "pragmatic-like" does not entail that the model engages in human-like social reasoning or possesses Theory of Mind. One plausible interpretation is that the model has acquired distributional regularities that approximate pragmatic cue sensitivity sufficiently well for its outputs to resemble communicative reasoning at the behavioral level. Prior work has shown that LLMs can succeed on tasks involving irony, indirect requests, belief tracking, and sensitivity to mental states sometimes matching human performance (e.g., Hu et al., 2023; Jones et al., 2024; Ni et al., 2025; Strachan et al., 2024). At the same time, evidence from larger models indicates persistent limitations in socially grounded reasoning, including humor and conversational norms (Hu et al., 2023). Given that models substantially larger than the one tested here still fall short of human performance on Theory-of-Mind-intensive tasks, it is likely that the present effects arise from surface-level pattern learning rather than explicit representation of speaker intentions or expectations. The present findings therefore speak primarily to the behavioral properties of LLM outputs under human-like task constraints, rather than providing new evidence about the cognitive mechanisms underlying human negation processing.

The present findings converge with those of Wang et al. (2025), who showed that negation bias can be detected in model-internal probability distributions using a surprisal-based approach. However, the two approaches operate at different levels of analysis. Wang et al. (2025) examine whether negated and affirmative forms differ in model expectancy, whereas the present study tests whether these differences give rise to the social-pragmatic inferences that define negation bias in human research, namely judgments about speaker expectations and causal attribution. This distinction is important because negation bias, as defined by Beukeboom et al. (2010), is not only a property of sentence likelihood but of interpretation. By eliciting explicit judgments in a task closely aligned with the original behavioral paradigm, the present study directly probes this inferential level. The two approaches also face different methodological constraints. Probability-based measures may be influenced by factors such as sentence length or word frequency, while behavioral prompting introduces sensitivities to prompt formulation, response scales, and context effects. In particular, Likert-scale responses from LLMs cannot be assumed to reflect stable internal representations in the same way as human judgments. In addition, LLMs may exhibit generic response tendencies (e.g., mid-scale preferences or defaulting to plausible-looking values), which complicates interpretation of absolute ratings. However, the structured interaction between description type and valence observed across five dependent variables argues against a simple scale-use artifact. Taken together, the convergence between probability-based and behavioral approaches suggests that negation bias is not confined to model-internal representations but also manifests in the explicit outputs through which users encounter instruction-tuned systems.

Several limitations should be noted. First, the study examines a single instruction-tuned model and therefore does not support generalization across architectures or training regimes. We treat this as a proof-of-concept demonstrating that pragmatic bias can emerge in explicit judgment settings rather than as evidence about LLMs in general. Second, the multi-turn prompting format may introduce context effects, as previous prompts and responses remain available in the model's context window. While human participants in repeated-measures designs may also show carryover effects, the persistence of prior turns in LLM context might differ in kind and should be investigated in future work. Finally, because the original Beukeboom et al. (2010) is publicly available, the possibility of training data contamination cannot be excluded. We therefore focus on qualitative pattern replication, not quantitative equivalence with human data.

In conclusion, this study demonstrates that LLMs can reproduce human-like negation bias effects, including expectation-based inferences and shifts in causal attribution. These findings highlight that bias in language models is not limited to explicit semantic associations, but can also emerge from structural and pragmatic features of linguistic form. Because such biases operate subtly and indirectly, they may evade standard content-based mitigation strategies, with implications for downstream applications such as dialogue systems and content generation. A comprehensive assessment of bias in LLMs therefore requires attention not only to what is said, but also to how it is expressed.

Acknowledgments

Barbara Kaup received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SFB 1629 "Negation in Language and Beyond" (509468465) [Projects C2/C3]. Francesca Capuano received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SFB1718 "Common Ground" (538184825) [Project S].

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Beukeboom, C. J., Burgers, C., Szabó, Z. P., Cvejic, S., Lönnqvist, J.-E. M., & Welbers, K. (2020). The negation bias in stereotype maintenance: A replication in five languages. *Journal of Language and Social Psychology*, *39*(2), 219–236. https://doi.org/10.1177/0261927x19869759

Beukeboom, C. J., Finkenauer, C., & Wigboldus, D. H. (2010). The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology*, *99*(6), 978–992. https://doi.org/10.1037/a0020861

Bonnefon, J.-F., & Villejoubert, G. (2007). Modus tollens, modus shmollens: Contrapositive reasoning and the pragmatics of negation. *Thinking & Reasoning*, *13*(2), 207–222. https://doi.org/10.1080/13546780601069488

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, *9*, 1453. https://doi.org/10.3389/fpsyg.2018.01453

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cuccio, V. (2012). Is embodiment all that we need? Insights from the acquisition of negation. *Biolinguistics*, *6*(3-4), 259–275. https://doi.org/10.5964/bioling.8919

Da, Y., Bossa, M. N., Berenguer, A. D., & Sahli, H. (2024). Reducing bias in sentiment analysis models through causal mediation analysis and targeted counterfactual training. *IEEE Access*, *12*, 10120–10134. https://doi.org/10.1109/ACCESS.2024.3353056

De Villiers, J. G., & Flusberg, H. B. T. (1975). Some facts one simply cannot deny. *Journal of Child Language*, *2*(2), 279–286. https://doi.org/10.1017/s0305000900001100

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.2196/preprints.48399

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. *Intercultural Pragmatics*, *5*, 517–540. https://doi.org/10.1515/iprg.2008.025

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. (2023). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*, 1097–1179. https://doi.org/10.1162/coli_a_00524

Giora, R. (2007). "A good Arab is not a dead Arab – a racist incitement": On the accessibility of negated concepts. In I. Kecskes & L. R. Horn (Eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (pp. 129–164). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110198843.2.129

Giora, R., Balaban, N., Fein, O., & Alkabets, I. (2004). Negation as positivity in disguise. In H. L. Colston & A. N. Katz (Eds.), *Figurative language comprehension: Social and cultural influences* (pp. 233–258). Lawrence Erlbaum Associates.

Giora, R., Fein, O., Ganzi, J., & Alkeslassy Levi, N. (2005). On negation as mitigation: The case of negative irony. *Discourse Processes*. https://doi.org/10.1207/s15326950dp3901_3

Givón, T. (1978). Negation in language: Pragmatics, function, ontology. In *Pragmatics* (pp. 69–112). Academic Press. https://doi.org/10.1163/9789004368873_005

Gotzner, N., & Mazzarella, D. (2021). Face management and negative strengthening: The role of power relations, social distance, and gender. *Frontiers in Psychology*, *12, Article 602977*. https://doi.org/10.3389/fpsyg.2021.602977

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. speech acts* (pp. 26–40). Academic Press. https://doi.org/10.1163/9789004368811_003

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation?: An examination of negated metaphors [Special Issue: Processes and Products of Negation]. *Journal of Pragmatics*, *38*(7), 1015–1032. https://doi.org/10.1016/j.pragma.2005.12.005

Horn, L. (1989). *A natural history of negation*. The University of Chicago Press. https://doi.org/10.1016/0024-3841(90)90064-R

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Georgetown University Press. https://hdl.handle.net/10822/555477

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4194–4213). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.230

Johnson-Laird, P. N., & Tridgell, J. (1972). When negation is easier than affirmation. *Quarterly Journal of Experimental Psychology*, *24*(1), 87–91. https://doi.org/10.1080/14640747208400271

Jones, C. R., Trott, S., & Bergen, B. (2024). Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (epitome). *Transactions of the Association for Computational Linguistics*, *12*, 803–819. https://doi.org/10.1162/tacl_a_00674

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2005). Effects of negation, truth value, and delay on picture recognition after reading affirmative and negative sentences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *27*. https://escholarship.org/uc/item/19s068vb

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? [Special Issue: Processes and Products of Negation]. *Journal of Pragmatics*, *38*(7), 1033–1050. https://doi.org/10.1016/j.pragma.2005.09.012

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). Experiential simulations of negated text information. *Quarterly Journal of Experimental Psychology*, *60*(7), 976–990. https://doi.org/10.1080/17470210600823512

Lee, M. H., Montgomery, J. M., & Lai, C. K. (2024). Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1321–1340. https://doi.org/10.1145/3630106.3658975

Mistral AI. (2026). Mistral AI [Large Language Model] [Accessed January 26, 2026].

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Ni, Q., Yu, Y., Ma, Y., Lin, X., Deng, C., Wei, T., & Xuan, M. (2025). The social cognition ability evaluation of LLMs: A dynamic gamified assessment and hierarchical social learning measurement approach. *ACM Transactions on Intelligent Systems and Technology*, *16*(6). https://doi.org/10.1145/3673238

Nordmeyer, A., & Frank, M. C. (2015). *Negation is only hard to process when it is pragmatically infelicitous* [Unpublished manuscript].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Rubio-Fernandez, P. (2021). Pragmatic markers: The missing link between language and theory of mind. *Synthese*, *199*(1), 1125–1158. https://doi.org/10.1007/s11229-020-02768-z

Schindele, R., Lüdtke, J., & Kaup, B. (2008). Comprehending negation: A study with adults diagnosed with high functioning autism or asperger's syndrome. *Intercultural Pragmatics*, *5*, 421–444. https://doi.org/10.1515/iprg.2008.021

Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*, 1–11. https://doi.org/10.1038/s41562-024-01882-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., Sommerauer, P., & Bloem, J. (2025). The negation bias in large language models: Investigating bias reflected in linguistic markers. *Second Conference on Language Modeling*.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, *4*(1), 7–11. https://doi.org/10.1016/s0022-5371(65)80060-3

Wigboldus, D. H., Semin, G. R., & Spears, R. (2006). Communicating expectancies about others. *European Journal of Social Psychology*, *36*(6), 815–824. https://doi.org/10.1002/ejsp.323

Zhao, Y., Wang, B., Wang, Y., Zhao, D., He, R., & Hou, Y. (2025). Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *Findings of the Association for Computational Linguistics: ACL 2025*, 1–12. https://arxiv.org/abs/2501.02295