Rethinking Social Cognition Benchmarks: Why Context Matters for Reliable Evaluation

Anonymous ACL submission

Abstract

Social cognition—the ability to understand others' mental states—relies heavily on contextual subtleties. However, current NLP evaluations of social cognition predominantly focus on static evaluation using decontextualized statements with majority-vote labels, overlooking the nuanced interpretations crucial to such interactions. This raises concerns about the reliability of these evaluations in assessing social cognitive capabilities. In this work, we conduct extensive human experiments across six social cognition benchmarks to quantify how static evaluations lead to inconsistent interpretations and demonstrate that incorporating contextual information significantly improves human agreement and performance. Building on these insights, we propose a novel framework that uses persona-based simulations to systematically identify ambiguous items before benchmark deployment. Finally, we evaluate large language models (LLMs) under both static and contextualized conditions, revealing that model rankings shift substantially when context is provided for ambiguous items—highlighting that current evaluation approaches may not accurately reflect true social cognitive capabilities. Our findings underscore the importance of considering context when designing and deploying social cognition benchmarks.

1 Introduction

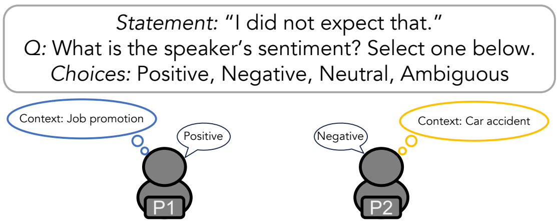

Statement: "I did not expect that."
Q: What is the speaker's sentiment? Select one below.
Choices: Positive, Negative, Neutral, Ambiguous
Context: Job promotion
Positive
Negative
Context: Car accident
P1
P2

Figure 1: An example of how different context can lead to different interpretations of the same statement in a typical annotation task.

Understanding others' mental states—beliefs, intentions, and emotions—is a cornerstone of human social interaction. This ability, often referred to as social cognition, is deeply contextual, relying on the subtleties of dialogue and individual experiences (Pons et al., 2003; Apperly, 2012). Individual variation in interpreting mental states comes from differences in cultural backgrounds (Perez-Zapata et al., 2016; Shahaeian et al., 2011), personal experiences (Dweck, 2017), and cognitive styles (Wellman et al., 2001). For instance, Figure 1 shows the ambiguous statement "I did not expect that" could express delight when heard after a promotion or dismay following an accident. Context is essential to resolve such ambiguities and enable accurate social reasoning (Kuhlen and Brennan, 2013).

Despite this, current NLP evaluations of social cognition are *static*, relying on decontextualized statements with majority-vote labels, assuming a single correct interpretation (Creanga and Dinu, 2024; Li et al., 2024). These evaluations fail to account for the inherent variability and ambiguity of human language, oversimplifying the complexity of mental state inference (Demszky et al., 2020; Srinivasan and Choi, 2022; Farha et al., 2022; Nakov et al., 2019; Mohammad et al., 2016). As a result, such benchmarks often produce unreliable assessments of both human and LLMs capabilities (Li et al., 2024). Recent efforts in evaluating LLMs have seen a move towards interactive settings such as Chatbot Arena (Chiang et al., 2024), where users engage in open-ended conversations with paired LLMs and provide preferences feedback based on interactions. While these evaluations offer insights into overall conversational ability, it is difficult to isolate and measure social cognitive abilities in this type of evaluations due to the potential influence of surface-level factors, such as writing styles (Cao, 2024) and engagement (Peng et al., 2024). This signifies the need for domain-specific, well-designed social cognition benchmarks that can provide isolated measure of LLM's ability to accurately infer people's mental states including intent, belief and emotion.

However, current social cognition benchmarks face significant limitations of their own. In this work, we conduct extensive human experiments to 1) reveal that static evaluations frequently result in significant inconsistencies, with ambiguous items contributing to low agreement, 2) quantify how adding context significantly improves human agreement and performance by comparing static to contextualized evaluations. We then introduce a novel framework using persona-based simulation to systematically identify potentially ambiguous items in social cognition benchmarks before deployment. Finally, we evaluate large language models under both static and contextualized conditions, revealing that model rankings shift substantially when context is provided — highlighting that current static evaluations may not accurately reflect models' true social cognitive capabilities.

2 Related Work

2.1 Social Cognition Evaluations

Current social cognition datasets consist primarily of decontextualized statements from internet sources, annotated for intent (Farha et al., 2022; Srinivasan and Choi, 2022), belief (Nakov et al., 2019; Davydova and Tutubalina, 2022), and emotion (Demszky et al., 2020; Mohammad et al., 2016). There have been increasing calls to develop context-enhanced approaches for NLP evaluation (Li et al., 2024). Recent advances in emotional intelligence evaluation have incorporated human-designed situational contexts (Wang et al., 2023; Sabour et al., 2024), while LLM-generated content has been used for emotion understanding (Paech, 2023; Gandhi et al., 2024b) and belief inference (Gandhi et al., 2024a; Xu et al., 2024a). Fully interactive frameworks, where LLMs are evaluated in agentic settings, have also emerged (Zhou et al., 2023; Wang et al., 2024). While these approaches represent important progress, they lack systematic analysis of how context impacts human agreement and model evaluation. Our work provides the first comprehensive study quantifying the effects of contextual information on both human and model performance across multiple social cognition benchmarks.

2.2 Label Variations and Ambiguity

Variations in annotations are prevalent across various NLP tasks, including natural language inference (NLI) (Beck et al., 2020; Huang and Yang, 2023), sentiment analysis (Jiang and Marneffe, 2022), text classification (Yuan et al., 2024). These variations stem from factors such as inherent ambiguity in language (Xu et al., 2024b), the subjective nature of interpretation (Deng et al., 2023) and variations in annotator background and understanding (Wan et al., 2023). For example, in sentiment analysis, annotators may disagree on the sentiment expressed in a text due to differing interpretations of nuanced expressions or vague statements (Xu et al., 2024b). Similarly, in text classification tasks, variations arise from challenges in assigning labels to documents with multiple applicable categories or when dealing with ambiguous or overlapping categories (Yuan et al., 2024). These challenges highlight the critical need to account for and address label variation. Approaches such as entropy-based methods for identifying ambiguous instances (Baumler et al., 2023), leveraging annotator certainty with multiple annotations (Andresen et al., 2020), and incorporating contextual cues to mitigate ambiguity (Beck et al., 2020) have proven effective across NLI tasks. (Deng et al., 2023) show that models can learn significantly better by explicitly leveraging annotator disagreements across a wide range of NLP tasks, indicating that disagreement is not merely noise but a rich signal that can improve performance when appropriately modeled. Our work extends these disagreement-aware approaches by introducing persona-backed annotations and entropy-based methods to model ambiguity in social cognition tasks, while showing how context can resolve different interpretations.

3 Contextual Effects on Human Annotations

We assess how context affects human understanding of *people's mental states* through our human annotation experiments, and provide quantitative evidence demonstrate the critical need for context in social cognition benchmarks and reveal limitations of current static evaluation approaches.

3.1 Dataset selection

To comprehensively evaluate the impact of context on social cognition, we selected *six* benchmarks spanning the three core domains identified in the Social AI taxonomy by (Li et al., 2024): intent recognition, belief understanding, and sentiment recognition. Following the criteria outlined in (Li et al., 2024), we prioritized widely-used and highly cited datasets that are representative of each domain, open-sourced, and do not require any crawling with the X API [1]. Performing human experi-

[1] https://developer.x.com/en/products/x-api

ments on the entire datasets would incur significant costs. Therefore, we sampled 120 data points from each dataset, using stratified sampling to ensure that the class distribution in our sample is representative of the original dataset. For intent recognition, we selected TyDiP (Srinivasan and Choi, 2022) for politeness detection and iSarcasm (Farha et al., 2022) for sarcasm detection. For belief understanding, we used the COVID-19 Vaccine Stance dataset (Davydova and Tutubalina, 2022) and Abortion Stance dataset (Mohammad et al., 2016) to capture different belief stances on societal issues. For sentiment recognition, we used the sentiment labels from GoEmotions (Demszky et al., 2020) and the Tweet Sentiment dataset (Nakov et al., 2019).

3.2 Context Augmentation

We augmented the datasets with dialogue-based contexts, as conversations are a primary medium through which social cognition naturally occurs in human communication.

Context Generation We first filter the six datasets to include only sentences written from a first-person perspective using a DeBERTa V3 zero-shot model (He et al., 2021) to ensure relevance in conversational contexts. This step aimed to select statements that were more likely to occur within natural conversations, as many social interactions involve expressing personal thoughts, feelings, and beliefs. We then employ a three-step process using GPT-4o. The process begins with scenario creation, where we generate a realistic conversation setting where the statement might naturally occur, carefully considering the target mental state. For instance, when dealing with a statement labeled as "sarcastic," we might generate a scenario involving friends discussing an obviously unsuccessful event. Following scenario creation, we generate a conversation between two participants (Person A and B) without including the original statement, allowing the dialogue to develop organically. The final stage involves identifying the most natural position to insert the original statement within the dialogue, ensuring it aligns with the intended mental state while maintaining conversational coherence.

Automatic Quality Check We implement automated verification using GPT-4o to maintain dataset quality. The verification process examines whether the original statement's intended mental state label aligns with the generated context (*Label Alignment*), confirms that the dialogue progression has a natural flow (*Natural Flow*), and verifies that the inserted statement adds value without redundancy (*Redundant Content*). This systematic approach ensures that our contextualized versions preserve the original labels while providing natural conversational settings that can help disambiguate mental state interpretations.

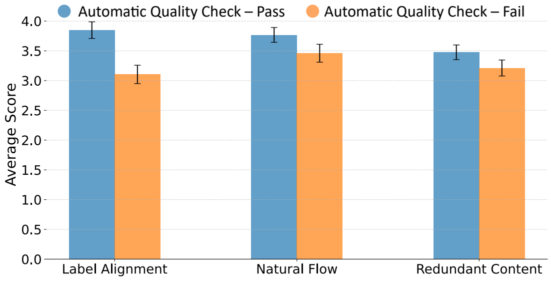


Figure 2: Results on human verification. Automatic Quality Check - Pass (AQC-Pass) and Automatic Quality Check - Fail (AQC-Fail) show average human ratings on a 5-point Likert scale (1: strongly disagree, 3: neutral, 5: strongly agree) across three evaluation criteria.

Human Verification To validate our automatic quality check process, we recruited 10 annotators to evaluate 50 randomly sampled context-statement pairs form six datasets, consisting of 25 pairs that passed (Automatic Quality Check - Pass) and 25 that failed (Automatic Quality Check - Fail) our automated verification. For each pair, annotators rate three aspects on a 5-point Likert scale: label alignment (whether the statement's intended mental state remains consistent), natural flow (whether the dialogue flows naturally), and redundant content (whether the conversation contains redundancy). Figure 2 shows our quality check effectively identifies contexts that maintain label consistency - pairs that passed received agreement on label alignment (3.8) while failed pairs averaged near neutral (3.1). While both conditions received above-neutral ratings for natural flow and redundant content, passed pairs still showed better quality (3.8 vs 3.4 for flow, 3.5 vs 3.2 for redundancy). These results validate our automatic quality check's ability to filter for label consistency, though they also suggest room for improving the naturalness of generated conversations. Based on all of the above, we obtain high-quality final contextualized datasets, selecting *120 data points per dataset for subsequent experiments*.

3.3 Annotation Collection

We collect annotations from participants recruited through Prolific[2]. Since all tasks are conducted in English, we recruit participants located in the US and UK who have English as their first language, an approval rate above 95%, and completed

[2] https://www.prolific.com/

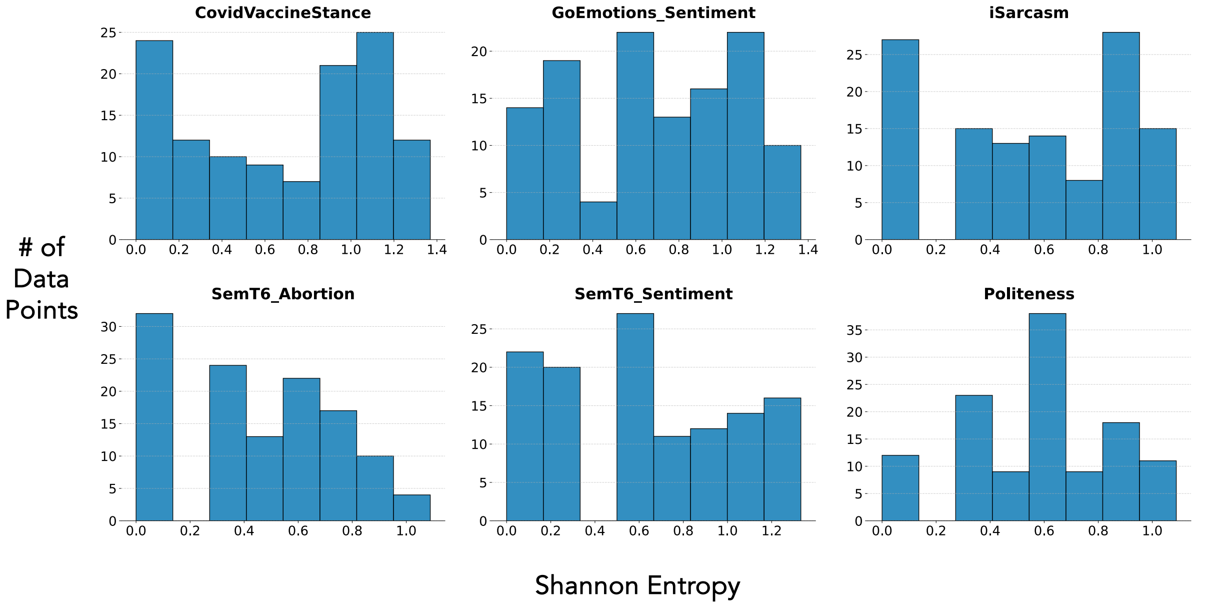


Figure 3: Entropy of the distribution of human annotations across six benchmarks. There is considerable amount of variation consistently in all datasets, suggesting that human annotations really vary on these static, out-of-context statements.

at least 1,000 submissions to ensure high-quality annotations. We design two experimental conditions for comparison, each containing the same 120 statement-label pairs. In the **No Context** condition, participants are presented with standalone statements and asked to select the appropriate label from all possible options. In the **Full Context** condition, participants view a conversation between two speakers (Person A and Person B) containing the target statement. They then label the target statement within the context of the dialogue, selecting from the same set of label options. The label choices include the original set of dataset labels, along with an **Ambiguous** option to allow participants to indicate uncertainty in their interpretation. To maintain reasonable session lengths, we divide each dataset into six batches of 20 items per condition, with 10 participants annotating each batch. We incorporate random attention checks throughout the study and exclude data from participants who fail these checks. Full annotation templates are provided in Section A.2.

3.4 Results

We examine human annotations to reveal two critical insights about social cognition evaluation. First, we demonstrate that static evaluations result in significant inconsistencies in human interpretation. Second, we quantify the extent to which providing contextual information improves both human agreement and performance through systematic comparison of static and contextualized evaluation conditions.

Annotation Inconsistency To analyze variability in human interpretations, we compute the Shannon entropy (Shannon, 1948) of the label distribution for each statement, where higher entropy indicates greater disagreement among annotators. As shown in Figure 3, we observe substantial variation in entropy values across all datasets, ranging from 0.0 to 1.4 with many items have entropy higher than 0.7. The multi-modal distributions, particularly evident in GoEmotions-Sentiment and SemT6-Sentiment, suggest systematic differences in how annotators interpret the same statements. This pattern underscores the inherent ambiguity in many static statements, which often lack sufficient context to constrain interpretation. For example, Figure 5 demonstrates how the same statement can elicit a wide range of interpretations between different annotators, further challenging the validity of static benchmarks as reliable measures of social cognition. While uncertainty could potentially stem from unclear task design, our subsequent analysis of contextual effects provides strong evidence that the primary source of these variations is insufficient contextual information.

Effect of Context To systematically examine how context affects evaluation quality, we analyze changes in ambiguity, agreement, and performance metrics. We quantify explicit ambiguity through an *ambiguity score*, defined as the proportion of times annotators selected the 'Ambiguous' option. Figure 4[a] shows that providing context leads to significant reductions in ambiguity scores across

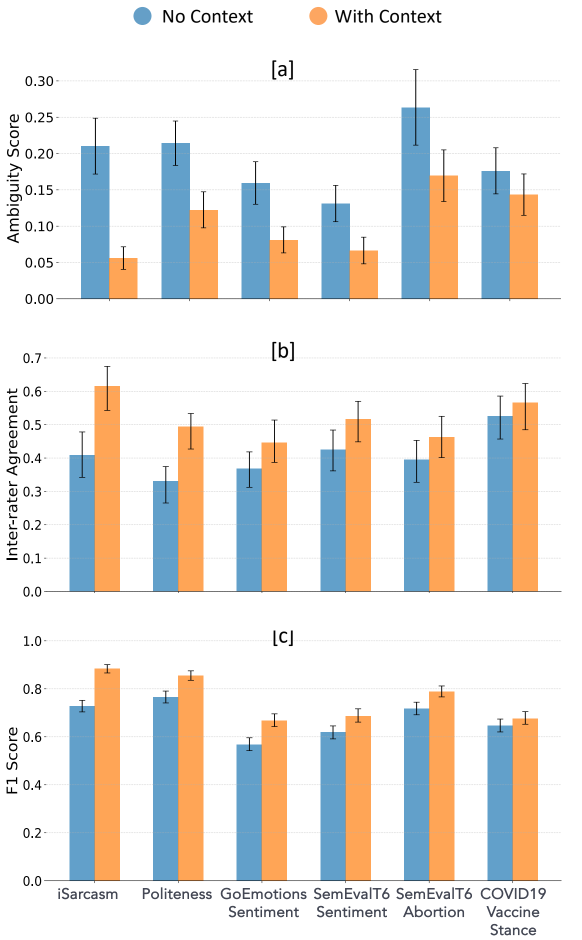
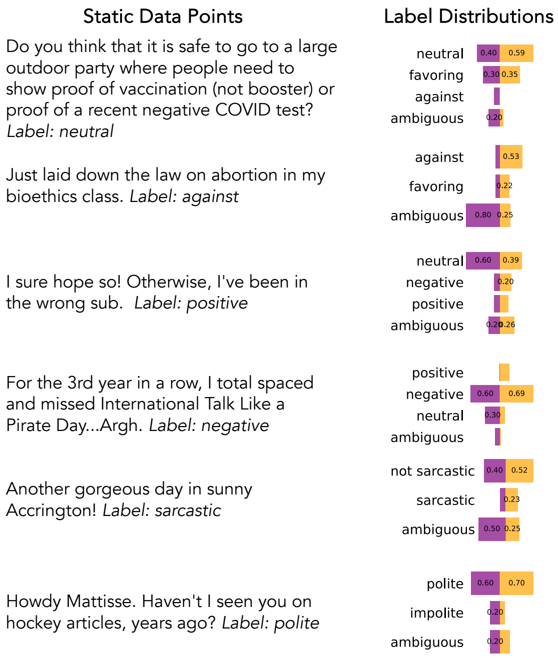


Figure 4: Comparison between dataset with no context and context for all items. [a] **Ambiguity score** is the percentage of ‘ambiguous‘ option selected by people across all items in a dataset. [b] **Inter-subject agreement**, computed using Kripendroff’s alpha, across 10 humans for all items in a dataset. [c] **F1 score** on the original labels. The error bars indicate bootstrapped 95% confidence intervals.

all datasets, with the most dramatic improvements in tasks requiring nuanced social interpretation: ambiguity decreased by 15% for iSarcasm and 9% for Politeness.e assess annotation consistency using *Kripendroff’s alpha*, which measures inter-annotator agreement. As shown in Figure 4[b], context consistently improves agreement scores across all tasks, particularly with iSarcasm increasing from 0.42 to 0.61. To evaluate accuracy, we compute F1 scores by comparing individual annotator labels against the original labels. The results in Figure 4[c] show substantial improvements with context, with average F1 scores increasing by 8-22% across tasks. Notably, the COVID19 Vaccine Stance task shows only minor improvements, suggesting that context’s impact varies by task type. Nevertheless, the consistent improvements across multiple metrics and most datasets provide strong evidence that the initial disagreements stem from insufficient contextual information rather than fundamental task design issues.



Do you think that it is safe to go to a large outdoor party where people need to show proof of vaccination (not booster) or proof of a recent negative COVID test? *Label: neutral*

Just laid down the law on abortion in my bioethics class. *Label: against*

I sure hope so! Otherwise, I've been in the wrong sub. *Label: positive*

For the 3rd year in a row, I total spaced and missed International Talk Like a Pirate Day...Argh. *Label: negative*

Another gorgeous day in sunny Accrington! *Label: sarcastic*

Howdy Mattisse. Haven't I seen you on hockey articles, years ago? *Label: polite*

Figure 5: Examples of label distribution from **Simulated** and **Real** human participants, showing entropy and label distribution for specific cases.

4 Ambiguous Data Points Detection

Our analysis shows that providing conversational context improves the reliability of social cognition evaluations across multiple dimensions: reducing explicit ambiguity in annotations, increasing inter-annotator agreement, and improving alignment with ground truth labels. However, this raises a practical challenge: How can benchmark creators identify which items actually need contextual augmentation? While our human experiments revealed patterns of ambiguity, conducting such extensive human studies for every new benchmark or dataset would be prohibitively resource-intensive. We need an efficient, automated method to identify potentially ambiguous items before deployment. To address this challenge, we propose a novel persona-based simulation framework that leverages large language models to simulate how different individuals would interpret social statements based on their relevant experiences. Our key insight is that by simulating interpretations from a set of personas, who can ground their judgment in relevant personal experience, we can identify statements that consistently yield multiple valid interpretations. This approach enables benchmark creators to proactively identify and address ambiguous items before deployment, improving benchmark quality while minimizing the need for extensive human validation.

4.1 Selective Persona-based Method

Our method leverages a pool of 40 personas from SOTOPIA (Zhou et al., 2023), each with detailed profiles including demographics, personalities, occupations, and background stories. To ensure meaningful interpretations rather than forced judgments (e.g., a lighthouse keeper evaluating internet memes), we implement a two-stage process demonstrated in Figure 6. First, we filter personas based on relevance — each LLM-simulated persona must recall a specific occasion when they encountered the target statement in conversation. Only personas who can ground their interpretation in relevant personal experience proceed to the labeling stage. Second, qualified personas provide labels through chain-of-thought reasoning using Llama 3.1-70B (Dubey et al., 2024). We generate 20 responses per relevant statement-persona pair (temperature=1) to capture natural interpretative variability. From these label distributions (examples in Figure **??**, we identify ambiguous items using an entropy-based threshold optimized for each dataset (detailed in Section A.3) inspired by (Baumler et al., 2023).

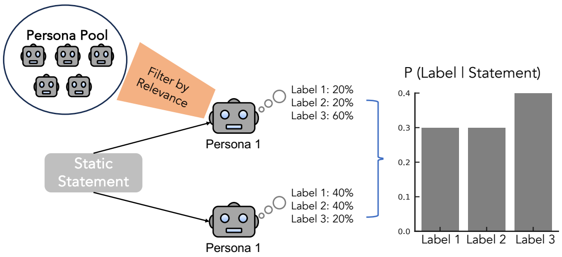


Figure 6: Examples of label distribution from **Simulated** and **Real** human participants, showing entropy and label distribution for specific cases.

Validation Metrics Building on our findings in Section 3 that showed how context significantly reduces annotation ambiguity and improves human performance, we evaluate our method using two complementary metrics derived from the same pool of annotators described in Section 3.3: the ambiguity score, defined as the proportion of 'Ambiguous' selections by human annotators (Section 2.2), and the human F1 score after selectively augmenting identified ambiguous items with context. These metrics directly measure our method's ability to identify items where contextual information is most crucial for resolving interpretative uncertainty, as demonstrated by our earlier human experiments.

4.2 Comparison with alternative approaches

To validate our method (Selective Persona - CoT), we compare against several baseline approaches that simulate label distributions. These include zero-shot prompting with logprobs for label distribution (No Persona - Direct), Chain-of-Thought prompting with 20 responses (No Persona - CoT), zero-shot prompting with unfiltered personas (Unfiltered Persona - Direct), and Chain-of-Thought prompting with unfiltered personas (Unfiltered Persona - CoT).

Selective Persona-CoT Achieves Best Performance-Efficiency Trade-off Our experimental results, averaged across all six datasets, demonstrate that all methods improve upon the static baseline (F1: 0.67, ambiguity score: 0.19). The Selective Persona-CoT approach achieves marginally better performance with the highest F1 score of 0.75 and lowest ambiguity score of 0.11. While Unfiltered Persona-Direct shows strong performance (F1: 0.74, ambiguity score: 0.13), it requires extensive hyperparameter tuning across different persona counts, introducing significant computational overhead (see Section A.4. The consistent improvements compared to different approaches suggest that both chain-of-thought reasoning and persona-based methods contribute to better ambiguity detection. Detailed performance breakdown by dataset is available in Section A.4.

Selective Persona CoT Outperforms the Alternatives Based on Table 2, the first observation we make is that all method is better than static method when it comes to reducing ambiguity score and improving F1, with Selective Persona - CoT. The Selective Persona - CoT approach has the highest F1 score and the lowest ambiguity score, although the difference there is less pronounced. The Unfiltered Persona - Direct method performs very similarly to our approach, but this approach requires a sweep through different numbers of randomly sampled personas in order and the results here are the best ones from the sweep which incur additional computation cost. This validation suggests our approach provides a robust framework for improving context augmentation in social cognition tasks through automatic persona relevance determination.

Context Most Benefits Tasks Requiring Subtle Social Interpretation Our method's

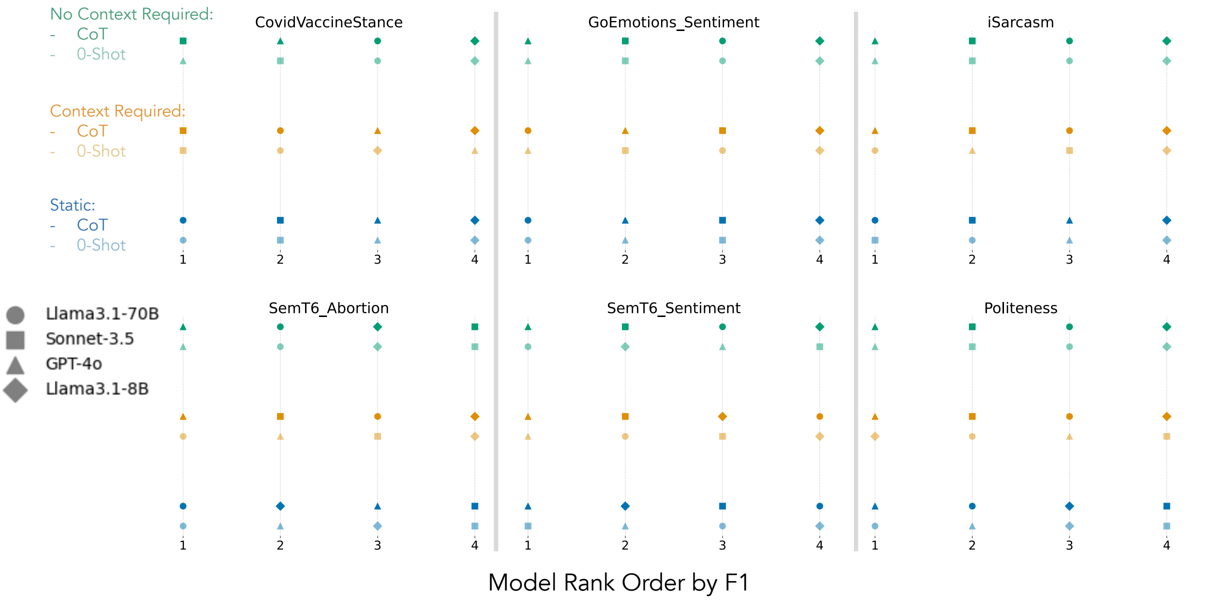


Figure 7: A detailed break down of model order by dataset, evaluation method (CoT vs 0-Shot) and different conditions. There is no single stable rank order of models across any axis of the variations.

Model	iSarcasm		Politeness		GoEmotions		SemEvalT6		Covid Vaccine		SemEvalT6 - Abortion	
	None	All	None	All	None	All	None	All	None	All	None	All
Human (Average)	0.73	0.89	0.78	0.86	0.59	0.69	0.62	0.69	0.65	0.70	0.73	0.80
Human (Best)	0.82	**0.99**	**0.94**	0.98	**0.79**	**0.87**	**0.80**	**0.86**	**0.81**	**0.84**	0.83	0.85
GPT-4o (0-shot)	0.55	0.91	0.91	0.95	0.73	0.77	0.70	0.83	0.61	0.61	0.73	0.87
GPT-4o (COT)	0.71	0.96	0.93	**0.99**	0.73	0.77	0.75	0.83	0.65	0.72	0.73	0.87
Llama3.1-70B (0-shot)	0.71	0.91	0.92	0.96	0.74	0.77	0.66	0.72	0.70	0.71	**0.86**	**0.89**
Llama3.1-70B (COT)	**0.84**	0.96	0.87	0.98	0.74	0.77	0.64	0.68	0.73	0.78	0.84	0.86
Llama3.1-8B (0-shot)	0.48	0.90	0.88	0.97	0.70	0.66	0.63	0.63	0.56	0.60	0.71	0.84
Llama3.1-8B (COT)	0.50	0.89	0.86	0.77	0.70	0.66	0.66	0.68	0.56	0.64	0.78	0.86
Sonnet-3.5 (0-shot)	0.76	0.91	0.78	0.93	0.67	0.75	0.70	0.63	0.64	0.80	0.61	0.86
Sonnet-3.5 (COT)	0.76	0.97	0.77	**0.99**	0.70	0.75	0.70	0.74	0.67	**0.84**	0.64	0.86

Table 1: F1 score for None (no context provided for any item) and All (context is provided for all items). Model results are averages over three runs with temperature at 0.2.

Method	F1	Ambiguity Score
Static	0.67	0.19
No Persona - Direct	0.72	0.14
No Persona - CoT	0.73	0.14
Unfiltered Persona - Direct	0.74	0.13
Unfiltered Persona - CoT	0.73	0.14
Selective Persona - CoT	**0.75**	**0.11**

Table 2: Comparison between our method and alternative methods by taking the average performance across 6 datasets. Our method has the lowerst ambiguity score and the highest F1.

Dataset	Ambiguity Score		F1	
	With Context	Static	With Context	Static
CovidVaccineStance	0.15	0.18	0.66	0.64
GoEmotions Sentiment	0.10	0.16	0.65	0.57
iSarcasm	0.05	0.21	0.87	0.71
Politeness	0.12	0.21	0.86	0.77
SemT6 Abortion	0.18	0.26	0.77	0.71
SemT6 Sentiment	0.08	0.13	0.67	0.62

Table 3: Difference in human ambiguity score and performance on the original labels between adding context to identified ambiguous items (With Context) and No Context.

effectiveness varies across different social cognition tasks, with particularly strong improvements on datasets requiring nuanced interpretation shown in Table 3. The most substantial gains appear in tasks involving subtle social cues: iSarcasm shows the largest reduction in ambiguity score (from 0.21 to 0.05) and the highest improvement in F1 score (from 0.71 to 0.87), while Politeness exhibits similar strong improvements (ambiguity reduction from 0.21 to 0.12, F1 increase from 0.77 to 0.86). For stance detection tasks, the improvements are more moderate - CovidVaccineStance shows minimal changes in both metrics (ambiguity: 0.18 to 0.15, F1: 0.64 to 0.66), suggesting that stance interpretation may rely less on immediate conversational context. Sentiment analysis tasks show consistent but moderate improvements, with GoEmotions Sentiment experiencing notable ambiguity reduction and F1 improvement, while SemT6 Sentiment shows more modest gains.

5 Evaluation with the Selective Persona Pipeline

Prior sections established that static evaluations lead to inconsistent human interpretations of social interactions. Here we investigate whether these inconsistencies affect how models are ordered by their F1 scores. Specifically, we examine if the relative ordering of models based on their F1 scores changes when context is added to the evaluation. We evaluate four state-of-the-art models (GPT-4o, Llama3.1-70B, Llama3.1-8B, and Claude3-Sonnet-3.5) using both zero-shot and chain-of-thought prompting strategies.

Static Evaluation Rankings Don't Transfer to Context-Dependent Cases Our rank correlation analysis in Figure 8 reveals that model rankings shift substantially when comparing Static versus Context Required conditions across multiple tasks. While zero-shot evaluations show low rank correlations (0.4) in (iSarcasm, Politeness, and SemT6 Sentiment), CoT prompting leads to more dramatic shifts - from perfect correlation in GoEmotions Sentiment (1.0) to near-zero in Politeness and negative correlation in SemT6 Abortion (-0.6). A detailed breakdown in Figure 7 further illustrates these inconsistencies: models' relative performance changes dramatically across evaluation conditions. For instance, while Llama3.1-70B leads in the Static setting with GPT-4o ranking second, their ordering reverses in the Context Required condition, with GPT-4o taking the lead and Llama3.1-70B dropping significantly. These substantial variations persist across all dimensions of evaluation - whether comparing different context conditions or prompting strategies - indicating that current evaluation frameworks cannot provide a stable assessment of models' relative capabilities.

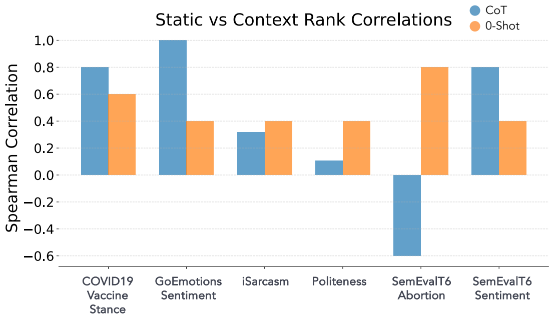


Figure 8: Rank correlation for order of models based on F1 score between the original *Static* setup and the subset where *Context is needed*.

Models Match Human Performance with Task-Specific Exceptions Given that we investigate how evaluation settings affect model orderings, human performance provides a consistent reference point for validating whether performance variations across settings reflect meaningful differences rather than evaluation artifacts. Using identical evaluation conditions as our human study in Section 3.3, we compare model and human performance across six datasets, each containing 120 statements evaluated both with and without context. We use zero-shot and CoT prompting and sample responses from each model at *temperature=0.2* three times and compute the average. We find that top models perform on par with human experts in most settings, achieving comparable F1 scores in tasks like Politeness (0.99 vs 0.98) and iSarcasm (0.96 vs 0.99), as shown in Table 1. However, models still lag behind in specific scenarios - notably on static COVID-Vaccine stance detection (0.73 vs 0.81) and contextualized GoEmotions Sentiment (0.77 vs 0.87). These results show minimal performance gaps between models and humans across most settings, with differences emerging only in specific cases like COVID-Vaccine stance detection and contextualized GoEmotions Sentiment. This suggests that current evaluation frameworks may be insufficient to meaningfully distinguish between human and model capabilities in social cognition tasks.

6 Conclusion

Our work demonstrates the importance of context in evaluating social cognition capabilities through two key findings. Through systematic human experiments across six social cognition benchmarks, we show that static evaluation setups lead to inconsistent interpretations, while adding context significantly reduces annotation ambiguity and improves human agreement and performance. We propose a selective persona-based framework that provides a practical method for identifying ambiguous items requiring contextual augmentation before benchmark deployment. Evaluation of state-of-the-art LLMs reveals that current static benchmarks do not reliably capture social cognitive capabilities, as model rankings shift substantially between static and contextualized evaluations and vary across prompting strategies. Our findings indicate that static benchmark performance does not predict ability on context-dependent cases, suggesting that future social cognition benchmarks should systematically validate items using methods like our persona-based simulation to ensure reliable evaluation.

7 Limitations

Our study has several important limitations. First, while our method effectively identifies ambiguous statements, the context generation process relies on GPT-4o, potentially introducing biases or artifacts specific to this model. Future work could explore more diverse sources of contextual information or methods for validating generated contexts. Second, our evaluation focuses on classification-based tasks with predefined label sets. This structure, while practical for large-scale evaluation, may not fully capture the open-ended nature of human social reasoning. More complex tasks involving free-form responses or multi-turn interactions could provide additional insights. Third, while we demonstrate improved performance with context, our approach still relies on majority-vote labels for evaluation. This may not fully capture the nuanced ways humans navigate ambiguous social situations, particularly in cases where multiple interpretations are equally valid. Finally, our study's scope is limited to English-language datasets from primarily Western sources. Social cognition norms and interpretations can vary significantly across cultures, and future work should examine how these findings generalize to other cultural and linguistic contexts. Despite these limitations, our work provides a foundation for developing more nuanced and reliable evaluations of social cognition capabilities in language models. Future research could explore alternative methods for context generation, investigate more complex social reasoning tasks, and examine cross-cultural aspects of social cognition evaluation.

Acknowledgments

References

Melanie Andresen, Michael Vauth, and Heike Zinsmeister. 2020. Modeling ambiguity with many annotators and self-assessments of annotator certainty. In *proceedings of the 14th Linguistic Annotation Workshop*, pages 48–59.

Ian A Apperly. 2012. What is "theory of mind"? concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology*, 65(5):825–839.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73.

Hongliu Cao. 2024. Writing style matters: An examination of bias and fairness in information retrieval systems. *arXiv preprint arXiv:2411.13173*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Claudiu Creanga and Liviu P Dinu. 2024. Designing nlp systems that adapt to diverse worldviews. *arXiv preprint arXiv:2405.11197*.

Vera Davydova and Elena Tutubalina. 2022. Smm4h 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to covid-19. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 216–220.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carol S Dweck. 2017. From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological review*, 124(6):689.

Ibrahim Abu Farha, Silviu Oprea, Steve Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *The 16th International Workshop on Semantic Evaluation 2022*, pages 802–814. Association for Computational Linguistics.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024a. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C Ong, and Noah D Goodman. 2024b.

Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Anna K Kuhlen and Susan E Brennan. 2013. Language in dialogue: When confederates might be hazardous to your data. *Psychonomic bulletin & review*, 20:54–72.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024. Social intelligence data infrastructure: Structuring the present and navigating the future. *arXiv preprint arXiv:2403.14659*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2019. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*.

Daniel Perez-Zapata, Virginia Slaughter, and Julie D Henry. 2016. Cultural effects on mindreading. *Cognition*, 146:410–414.

Francisco Pons, Joanne Lawson, Paul L Harris, and Marc De Rosnay. 2003. Individual differences in children's emotion understanding: Effects of age and language. *Scandinavian journal of psychology*, 44(4):347–353.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Ameneh Shahaeian, Candida C Peterson, Virginia Slaughter, and Henry M Wellman. 2011. Culture and the sequence of steps in theory of mind development. *Developmental psychology*, 47(5):1239.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. *arXiv preprint arXiv:2211.16496*.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024. Towards objectively benchmarking social intelligence for language agents at action level. *arXiv preprint arXiv:2404.05337*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024a. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

Jin Xu, Mariët Theune, and Daniel Braun. 2024b. Leveraging annotator disagreement for text classification. *arXiv preprint arXiv:2409.17577*.

Ling Yuan, Xinyi Xu, Ping Sun, Hai ping Yu, Yin Zhen Wei, and Jun jie Zhou. 2024. Research of multi-label text classification based on label attention and correlation networks. *PloS one*, 19(9):e0311305.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Appendix

A.1 Prompts

COVID Vaccine Stance Detection

Interactive Evaluation Prompt:
Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on the COVID-19 vaccine?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined without more context
Constraint: Please only the number of the choice (A, B, C or D).
RESPONSE FORMAT: Answer: ...

Interactive Few Shot Prompt Ambiguous:
Example {n} Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on the COVID-19 vaccine?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined without more context
Answer: {answer}

Static Few Shot Prompt Ambiguous:
Example {n} Statement: "{statement}"
Question: Based on this statement alone, what is the speaker's stance on the COVID-19 vaccine?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined from the context
Answer: {answer}

Interactive Few Shot Prompt No Ambiguous:
Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on the COVID-19 vaccine?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against
Answer: {answer}

COVID Vaccine Stance Detection DEI

DEI Evaluation Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
Question: From the perspective of the persona you're embodying, what is a person's stance on COVID19 vaccine when they said "{statement}"?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined without more context
Constraint: Please only the number of the choice (A, B, C or D).
PLEASE PROVIDE YOUR RESPONSE IN THE FOLLOWING FORMAT: Answer: ...

DEI Evaluation CoT Prompt:
ROLE-PLAY INSTRUCTIONS:
You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
Question: From the perspective of the persona you're embodying, what is a person's stance on COVID19 vaccine when they said "{statement}"? Think step by step.
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined without more context
Constraint: Please only the number of the choice (A, B, C or D).
PLEASE PROVIDE YOUR RESPONSE IN THE FOLLOWING FORMAT: Rationale: ... Answer: ...

DEI Evaluation Selective Prompt: Persona: {persona}
Task: Analyze the following statement from your persona's perspective. Remember to stay in character as this persona throughout your response.
Statement: {statement}
Question: From your persona's perspective, what is the speaker's stance on the COVID19 vaccine?
Choices: A. against: against B. favoring: in favor of C. neutral: neither favor nor against D. ambiguous: cannot be determined without more context

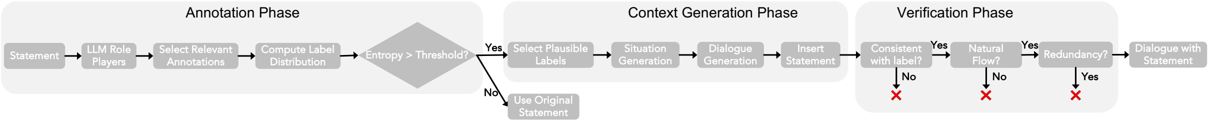


Figure 9: The process begins in the Annotation Phase, where LLM role players provide relevance and labels. Relevant annotations are selected, and entropy is computed from the label distribution. If the entropy exceeds a set threshold, the pipeline proceeds to the Context Generation Phase, where context is generated and the statement is embedded in a conversation. Finally, in the Verification Phase, the conversation is evaluated for natural flow, label consistency, and redundancy. Only items passing all verification checks are used in the final context-augmented output, while those failing are discarded.

COVID Vaccine Stance Detection DEI Continued

Provide your response in the following format: Relevance Assessment: [On a scale of 1-10, how relevant is this statement to your specific life experiences, knowledge, and values? Explain in detail, referencing at least two aspects of your persona that make this relevant or irrelevant. If the relevance is below 5, explain how this lack of relevance might affect your ability to make a judgment.]
Scenario: [If relevance is 5 or higher, recall a specific scenario where you encountered this statement, ensuring it aligns with your persona's characteristics. If relevance is below 5, write "No relevant scenario".]
[If a scenario was provided, continue with:] Rationale: [Your thought process before arriving at a label, explicitly referencing your traits, values, or experiences that influence this judgment]
Answer: [Provide your final choice here. Only provide the letter corresponding to your choice (A, B, C or D)]
Confidence: [On a scale of 1-10, how confident are you in this rating? Only provide a confidence score if you have provided a scenario.]

Abortion Stance

Interactive Evaluation Prompt:
Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on abortion?
Choices: A. against B. favoring C. ambiguous
Constraint: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Answer: ...

Interactive Few Shot Prompt Ambiguous:
Example {n} Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on abortion?
Choices: A. against B. favoring C. ambiguous
Answer: {answer}

Static Few Shot Prompt Ambiguous:
Example {n} Statement: "{statement}"
Question: Based on this statement alone, what is the speaker's stance on abortion?
Choices: A. against B. favoring C. ambiguous
Answer: {answer}

Interactive Few Shot Prompt No Ambiguous:
Conversation: {conversation}
Question: Based on the statement "{statement}" in the given conversation, what is the speaker's stance on abortion?
Choices: A. against B. favoring
Answer: answer

Abortion Stance DEI

DEI Evaluation Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona’s perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
Question: From the perspective of the persona you’re embodying, what is a person’s stance on abortion when they said "{statement}"?
Choices: A. against B. favoring C. ambiguous
Constraint: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Answer: ...

DEI Evaluation CoT Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona’s perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
Question: From the perspective of the persona you’re embodying, what is a person’s stance on abortion when they said "statement"? Think step by step.
Choices: A. against B. favoring C. ambiguous
Constraint: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Rationale: ... Answer: ...

DEI Evaluation Selective Prompt:
Persona: {persona}
Task: Analyze the following statement from your persona’s perspective. Remember to stay in character as this persona throughout your response.
Statement: {statement}
Question: From the perspective of the persona you’re embodying, what is the speaker’s stance on abortion?
Choices: A. against B. favoring C. ambiguous

Abortion Stance DEI Continued

Provide your response in the following format: Initial Reaction: [Briefly describe your immediate thoughts about this statement, considering your personality traits, values, and experiences.]
Relevance Assessment: [On a scale of 1-10, where 1 is "completely irrelevant" and 10 is "highly relevant", how relevant is this statement to your life and experiences? Explain why, referencing specific aspects of your persona.]
Scenario: [If relevance is 5 or higher, recall a specific scenario where you encountered this statement, ensuring it aligns with your persona’s characteristics. If relevance is below 5, write "No relevant scenario".]
[If a scenario was provided, continue with:] Rationale: [Your thought process before arriving at a label, explicitly referencing your traits, values, or experiences that influence this judgment]
Answer: [Provide your final choice here. Only provide the letter corresponding to your choice (A, B, or C)]
Confidence: [On a scale of 1-10, how confident are you in this rating? Only provide a confidence score if you have provided a scenario.]

iSarcasm

Interactive Evaluation Prompt:
Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be sarcastic when they said "{statement}"?
Choices: A. no, the speaker did not intend to be sarcastic B. yes, the speaker intended to be sarcastic C. ambiguous, it is unclear if the speaker intended to be sarcastic without more context
Constraint: Please only the number of the choice (A, B, or C).
RESPONSE FORMAT: Answer: ...

Interactive Few Shot Prompt Ambiguous:
Example {n} Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be sarcastic when they said "{statement}"?
Choices: A. no, the speaker did not intend to be sarcastic B. yes, the speaker intended to be sarcastic C. ambiguous, it is unclear if the speaker intended to be sarcastic without more context
Answer: {answer}

Static Few Shot Prompt Ambiguous:
Example n Statement: "{statement}"
Question: Does this statement intend to be sarcastic?
Choices: A. no, the statement is not intended to be sarcastic B. Yes, the statement is intended to be sarcastic C. ambiguous, it is unclear if the statement is intended to be sarcastic without more context
Answer: {answer}

Interactive Few Shot Prompt No Ambiguous:
Example n Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be sarcastic when they said "{statement}"?
Choices: A. no, the speaker did not intend to be sarcastic B. yes, the speaker intended to be sarcastic
Answer: {answer}

iSarcasm DEI

DEI Evaluation Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
QUESTION: From the perspective of the persona you're embodying, does this statement intend to be sarcastic?
CHOICES: A. No, the statement is not intended to be sarcastic B. Yes, the statement is intended to be sarcastic C. ambiguous, it is unclear if the statement is intended to be sarcastic without more context
CONSTRAINT: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Answer: ...

DEI Evaluation CoT Prompt:
OLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
QUESTION: From the perspective of the persona you're embodying, does this statement intend to be sarcastic? Think step by step.
CHOICES: A. No, the statement is not intended to be sarcastic B. Yes, the statement is intended to be sarcastic C. ambiguous, it is unclear if the statement is intended to be sarcastic without more context
INSTRUCTIONS: 1. Fully immerse yourself in the provided persona. 2. Carefully consider the statement from the perspective of the person you are embodying. 3. Choose the option that best represents how the person you're embodying would interpret the statement.
CONSTRAINT: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Rationale: ... Answer: ...

iSarcasm DEI Continued

DEI Evaluation Selective Prompt:
Persona: {persona}
Task: Analyze the following statement from your persona’s perspective. Remember to stay in character as this persona throughout your response.
Statement: {statement}
Question: From the perspective of the persona you’re embodying, does this statement intend to be sarcastic?
Choices: A. No, the statement is not intended to be sarcastic B. Yes, the statement is intended to be sarcastic C. ambiguous, it is unclear if the statement is intended to be sarcastic without more context
Provide your response in the following format: Initial Reaction: [Briefly describe your immediate thoughts about this statement, considering your personality traits, values, and experiences.]
Relevance Assessment: [On a scale of 1-10, where 1 is ”completely irrelevant” and 10 is ”highly relevant”, how relevant is this statement to your life and experiences? Explain why, referencing specific aspects of your persona.]
Scenario: [If relevance is 5 or higher, recall a specific scenario where you encountered this statement, ensuring it aligns with your persona’s characteristics. If relevance is below 5, write ”No relevant scenario”.]
[If a scenario was provided, continue with:] Rationale: [Your thought process before arriving at a label, explicitly referencing your traits, values, or experiences that influence this judgment]
Answer: [Provide your final choice here. Only provide the letter corresponding to your choice (A, B, or C)]
Confidence: [On a scale of 1-10, how confident are you in this rating? Only provide a confidence score if you have provided a scenario.]

GoEmotions and Tweet Sentiment

Interactive Evaluation Prompt:
Conversation: {conversation}
Question: Based on the statement ”{statement}” in the given conversation, what is the speaker’s sentiment?
Choices: A. positive B. negative C. neutral D. ambiguous
Constraint: Even if you are uncertain, you must choose one of A, B, C, or D, and ONLY output A, B, C or D as your answer.
RESPONSE FORMAT: Answer: ...

Interactive Few Shot Prompt Ambiguous:
Example {n} Conversation: {conversation}
Question: Based on the statement ”{statement}” in the given conversation, what is the speaker’s sentiment?
Choices: A. positive B. negative C. neutral D. ambiguous
Answer: {answer}

Static Few Shot Prompt Ambiguous:
Example {n} Statement: ”{statement}”
Question: What is the sentiment of the statement?
Choices: A. positive B. negative C. neutral D. ambiguous
Answer: {answer}

Interactive Few Shot Prompt No Ambiguous:
Conversation: {conversation}
Question: Based on the statement ”statement” in the given conversation, what is the speaker’s sentiment?
Choices: A. positive B. negative C. neutral
Answer: {answer}

GoEmotions and Tweet Sentiment DEI

DEI Evaluation Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona’s perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: ”{statement}”
QUESTION: From the perspective of the persona you’re embodying, what is the sentiment of the statement?
Choices: A. positive B. negative C. neutral D. ambiguous
Constraint: Please only the number of the choice (A, B, C or D).
RESPONSE FORMAT: Answer: ...

DEI Evaluation CoT Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona’s perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: ”{statement}”
QUESTION: From the perspective of the persona you’re embodying, what is the sentiment of the statement? Think step by step.
Choices: A. positive B. negative C. neutral D. ambiguous
Constraint: Please only the number of the choice (A, B, C or D).
RESPONSE FORMAT: Rationale: ... Answer: ...

DEI Evaluation Selective Prompt:
Persona: {persona}
Task: Analyze the following statement from your persona’s perspective. Remember to stay in character as this persona throughout your response.
Statement: {statement}
Question: From the perspective of the persona you’re embodying, what is the sentiment of the statement?
Choices: A. positive B. negative C. neutral D. ambiguous

GoEmotions and Tweet Sentiment DEI Continued

Provide your response in the following format: Initial Reaction: [Briefly describe your immediate thoughts about this statement, considering your personality traits, values, and experiences.]
Relevance Assessment: [On a scale of 1-10, where 1 is ”completely irrelevant” and 10 is ”highly relevant”, how relevant is this statement to your life and experiences? Explain why, referencing specific aspects of your persona.]
Scenario: [If relevance is 5 or higher, recall a specific scenario where you encountered this statement, ensuring it aligns with your persona’s characteristics. If relevance is below 5, write ”No relevant scenario”.]
[If a scenario was provided, continue with:] Rationale: [Your thought process before arriving at a label, explicitly referencing your traits, values, or experiences that influence this judgment]
Answer: [Provide your final choice here. Only provide the letter corresponding to your choice (A, B, C or D)]
Confidence: [On a scale of 1-10, how confident are you in this rating? Only provide a confidence score if you have provided a scenario.]

TyDiP

Interactive Evaluation Prompt:
Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be polite when they said "statement"?
Choices: A. No, the person did not intended to be polite B. Yes, the person intended to be polite C. Ambiguous, it is unclear if person intended to be polite without more context
RESPONSE FORMAT: Answer: ...

Interactive Few Shot Prompt Ambiguous:
Example {n} Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be polite when they said "{statement}"?
Choices: A. No, the person did not intended to be polite B. Yes, the person intended to be polite C. Ambiguous, it is unclear if person intended to be polite without more context
Answer: {answer}

Static Few Shot Prompt Ambiguous:
Example {n} Statement: "{statement}"
Question: Does this statement intend to be polite?
Choices: A. no, this statement did not intend to be polite B. yes, the statement intended to be polite C. ambiguous, it is unclear if the statement intended to be polite without more context
Answer: {answer}

Interactive Few Shot Prompt No Ambiguous:
Conversation: {conversation}
Question: In the given conversation, did the speaker intend to be polite when they said "statement"?
Choices: A. No, the person did not intended to be polite B. Yes, the person intended to be polite
Answer: answer

TyDiP DEI

DEI Evaluation Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
QUESTION: From the perspective of the persona you're embodying, does this statement intend to be polite?
Choices: A. no, this statement did not intend to be polite B. yes, the statement intended to be polite C. ambiguous, it is unclear if the statement intended to be polite without more context
Constraint: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Answer: ...

DEI Evaluation CoT Prompt:
ROLE-PLAY INSTRUCTIONS: You are to fully embody the persona described below. Analyze the given statement from this persona's perspective, considering their unique background, experiences, and mindset.
PERSONA INFORMATION: {persona}
TASK: As the person described above, analyze the following statement:
STATEMENT TO EVALUATE: "{statement}"
QUESTION: From the perspective of the persona you're embodying, does this statement intend to be polite? Think step by step.
Choices: A. no, this statement did not intend to be polite B. yes, the statement intended to be polite C. ambiguous, it is unclear if the statement intended to be polite without more context
Constraint: Please only the number of the choice (A, B or C).
RESPONSE FORMAT: Rationale: ... Answer: ...

DEI Evaluation Selective Prompt:
Persona: {persona}
Task: Analyze the following statement from your persona's perspective. Remember to stay in character as this persona throughout your response.
Statement: {statement}
Question: From the perspective of the persona you're embodying, what is the sentiment of the statement?
Choices: A. no, this statement did not intend to be polite B. yes, the statement intended to be polite C. ambiguous, it is unclear if the statement intended to be polite without more context

TyDiP DEI Continued

Provide your response in the following format: Initial Reaction: [Briefly describe your immediate thoughts about this statement, considering your personality traits, values, and experiences.]
Relevance Assessment: [On a scale of 1-10, where 1 is "completely irrelevant" and 10 is "highly relevant", how relevant is this statement to your life and experiences? Explain why, referencing specific aspects of your persona.]
Scenario: [If relevance is 5 or higher, recall a specific scenario where you encountered this statement, ensuring it aligns with your persona's characteristics. If relevance is below 5, write "No relevant scenario".]
[If a scenario was provided, continue with:] Rationale: [Your thought process before arriving at a label, explicitly referencing your traits, values, or experiences that influence this judgment]
Answer: [Provide your final choice here. Only provide the letter corresponding to your choice (A, B or C)]
Confidence: [On a scale of 1-10, how confident are you in this rating? Only provide a confidence score if you have provided a scenario.]

A.2 Human experiment details

We recruit participants via Prolific, filtering for people located in the United States whose primary language is English. Throughout the study, attention check questions are randomly interleaved with the actual items. Only data from participants who correctly answer these attention checks are included in our final analysis, ensuring a high level of data quality.

Introduction of an Ambiguous label Including an **"ambiguous"** label in addition to the existing labels for each dataset in the tasks is critical for capturing the inherent complexity of language and improving data quality. As (cite Andresen 2020) argue, ambiguity is an intrinsic property of natural language, and forcing annotators to choose between labels in unclear cases can lead to unreliable data. The "ambiguous" option allows annotators to explicitly mark cases where multiple interpretations are possible, preserving valuable information that would otherwise be lost. Moreover, as highlighted by (cite inter-vs-intra reliability), this approach helps distinguish between truly ambiguous cases and those where annotators have different but stable subjective interpretations. This not only provides a more nuanced view of human judgment in these tasks but also helps identify instances where additional context may help clarify social cognition. By including this option, we aim to capture a more realistic representation of human decision-making in social cognition tasks, acknowledging that not all situations yield clear-cut interpretations.

A.3 Threshold Selection for All Datasets

The reduction in human ambiguity scores is the criterion for selecting this threshold, as this directly addresses our goal of reducing ambiguity. Figure 10 demonstrates this process for the GoEmotions-Sentiment dataset. We observed an inverse relationship between ambiguity scores and F1 scores as we varied the threshold. This relationship suggests that as we provide context for more items (lowering ambiguity), the overall performance (F1 score) improves. In this example, the optimal threshold resulted in selecting approximately 90 out of 120 items for context augmentation. This balance represents a trade-off between reducing ambiguity and maintaining a manageable number of items for context generation.

The inverse relationship between ambiguity and F1 scores underscores the importance of context in improving task performance. By reducing ambiguity through context provision, we enable more accurate and consistent annotations, leading to better overall results. Moreover, the variability in the number of relevant personas across datasets highlights the importance of our selective approach. It allows us to capture task-specific nuances and ensure that only personas with relevant experiences contribute to the label distributions. This validation process confirms the effectiveness of our selective methods in identifying items that benefit most from context augmentation, thereby improving the quality and relevance of our augmented datasets for social cognition tasks.

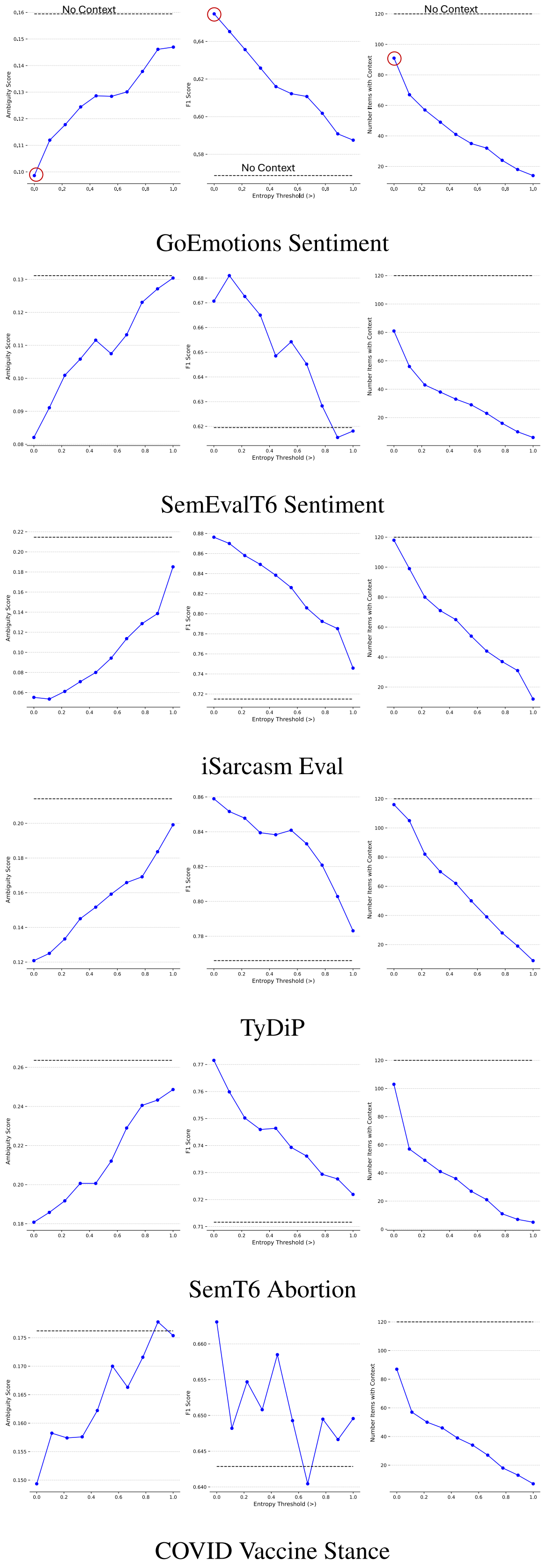
Persona Method Successfully Models Human Interpretation Patterns Our method generates label distributions that align well with human annotations on static datasets, as evidenced by both divergence and correlation metrics. Table 4 shows that the Jensen-Shannon Divergence between simulated and human distributions remains consistently low across all tasks (JSD range: 0.212-0.287), with Politeness showing the closest alignment (0.212) and GoEmotions-Sentiment showing slightly higher divergence (0.287). The strong positive Spearman's ρ across all tasks (0.635-0.738) further validate this alignment, with Politeness achieving the highest correlation (0.738) and iSarcasm showing relatively lower but still significant correlation (0.635). These results demonstrate that our selective persona method effectively captures the natural variation in human interpretations of social interactions.

Dataset	JSD Mean	Spearman's ρ
CovidVaccineStance	0.275	0.678
GoEmotions_Sentiment	0.287	0.648
iSarcasm	0.222	0.635
Politeness	0.212	0.738
SemT6_Abortion	0.236	0.698
SemT6_Sentiment	0.276	0.698

Table 4: Jessen-Shannon Divergence on the No Context condition between human and simulated persona label distributions and the ranking correlations. The simulated distribution comes from the Selective CoT Persona Ensemble method with Llama3.1-70B model.

A.4 More Analysis on Selective Persona - CoT

Additionaly Cost of Finding Optimal Number of Personas for Alternative Methods Unlike Selective Persona - CoT, the alternative



GoEmotions Sentiment

SemEvalT6 Sentiment

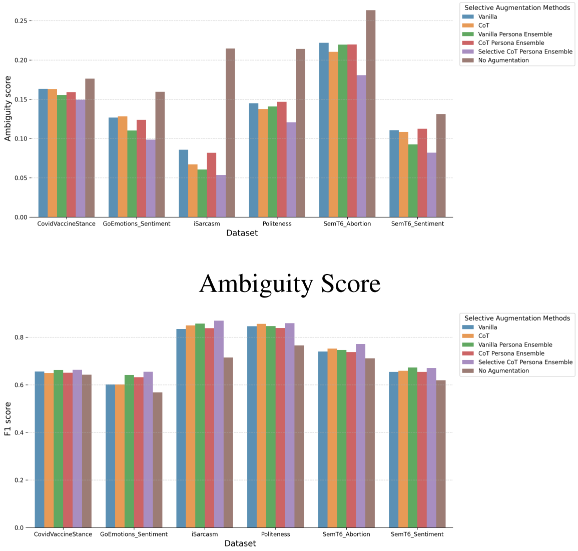
iSarcasm Eval

TyDiP

SemT6 Abortion

COVID Vaccine Stance

Figure 10: A demonstration of selecting the entropy threshold for selecting items that need more context.



Ambiguity Score

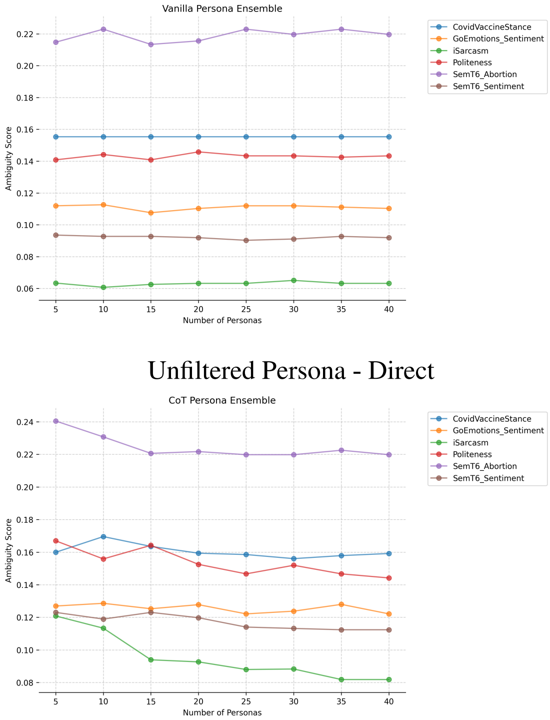
Figure 11: F1 scores

persona-based approaches require tuning the number of personas (M) from 5 to 40 as shown in Figure 12. These results demonstrate three key advantages of our method: (1) effective identification of statements needing context, (2) improved human agreement with original labels, and (3) strong performance without requiring persona count hyperparameter tuning.

A.5 Extra evaluation details

Few-shot evaluations Table 6 shows performance across different numbers of few-shot examples. For most models, there's a slight improvement from 0-shot to 3-shot or 5-shot, but the gains often plateau or even decrease slightly at 10-shot. The impact of few-shot learning varies across tasks and models, with some showing more consistent improvements than others. GPT-4o shows strong and consistent performance across tasks, often improving with few-shot examples. These results highlight the interplay between context, prompting strategies, and model capabilities in social cognition tasks, emphasizing the need for nuanced evaluation approaches in this domain.

Context Changes Model Performance Order Our rank correlation (RC) analysis in Table 5 reveals that the ordering of models by F1 scores shifts significantly when context is added, particularly with chain-of-thought prompting. For instance, in GoEmotions-Sentiment, while the model ordering under zero-shot remains relatively stable with the lowest 0.54 as the lowest RC,



Unfiltered Persona - Direct

Unfiltered Persona - CoT

Figure 12: The ambiguity score for both approaches across six datasets. The optimal number of persona does not stay the same thus these methods requiring specifc tuning for this hyperparameter.

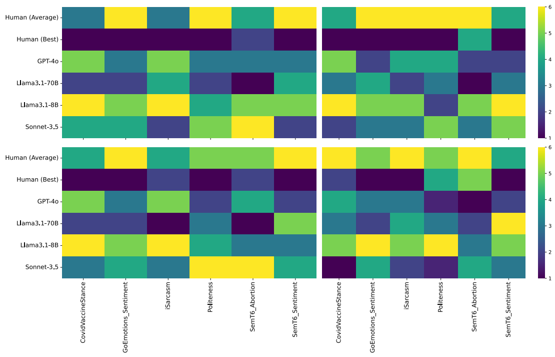


Figure 13: Performance ranking for each dataset. Top: Zero-shot evaluation vs humans. Bottom: CoT evaluation vs humans.

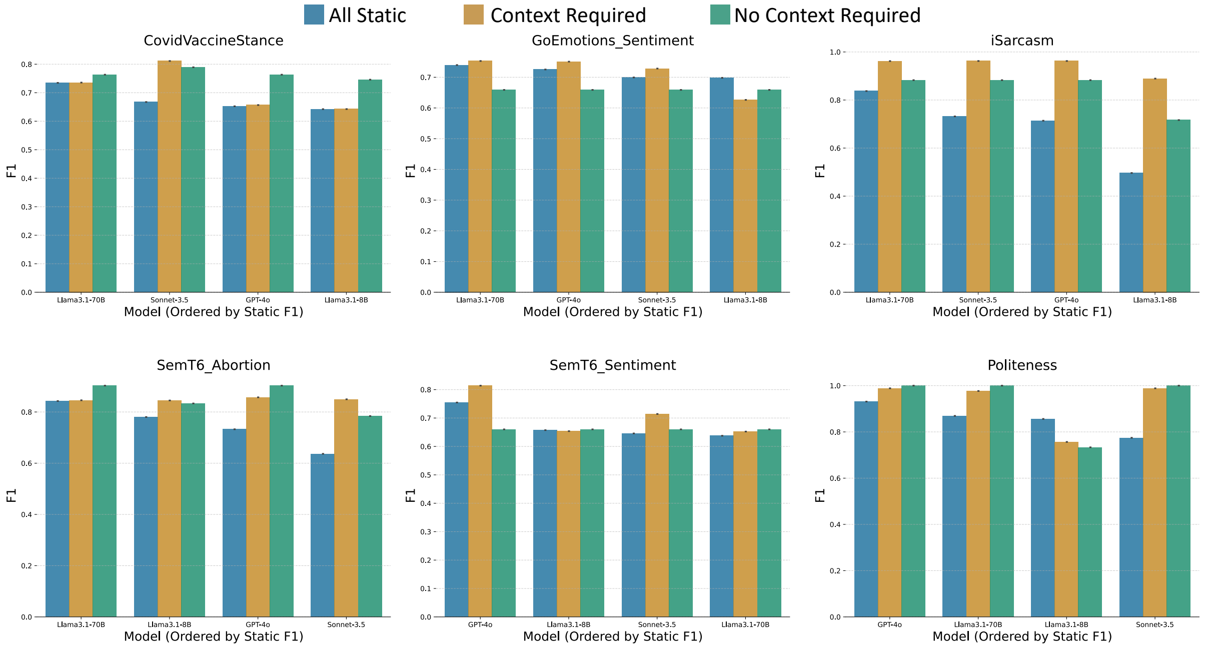


Figure 14: Model performance on different splits based on whether context is needed. **All Static** is the case where no statement has context. **Context Required** includes statements where context is needed. **No Context Required** consists of cases where context is not necessary.

chain-of-thought (CoT) prompting leads to substantial reordering in the case of the Polite. These findings indicate that static evaluations may not reliably capture models' true relative performance in social cognition tasks.

Dataset	**0-shot**	**CoT**
iSarcasm	0.54	0.43
Politeness	0.83	0.03
GoEmo-Sent	0.83	0.94
SemT6-Sent	0.54	0.71
CovidVacc	0.83	0.66
SemT6-Abor	0.49	0.31

Table 5: Rank correlation (**RC**) of performance between **No Context** and **Full Context** settings for Zero-Shot vs CoT performance.

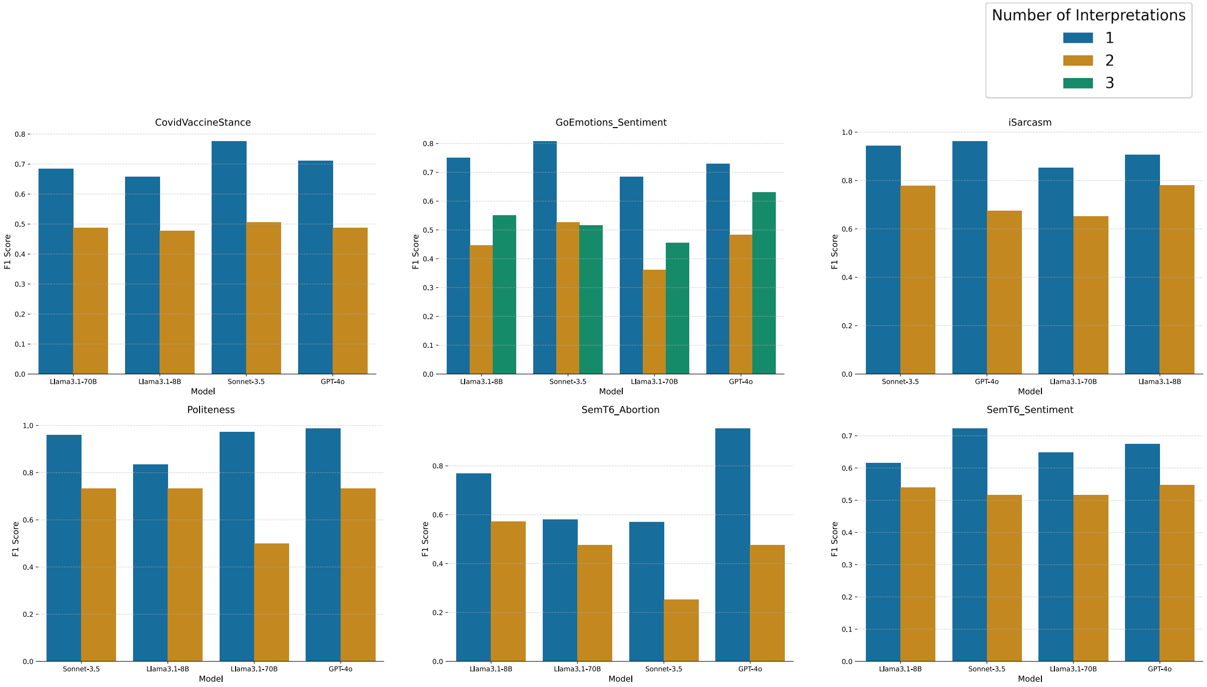


Figure 15: Evaluation with Zero-shot prompting

Dataset	**0-shot**		**3-shot**		**5-shot**		**10-shot**	
	Rank Corr	**p-value**	**Rank Corr**	**p-value**	**Rank Corr**	**p-value**	**Rank Corr**	**p-value**
iSarcasm	0.54	0.266	0.54	0.266	0.66	0.156	0.66	0.156
Politeness	0.83	0.042	0.83	0.042	0.89	0.019	0.83	0.042
GoEmotions - Sentiment	0.83	0.042	0.31	0.544	0.49	0.329	0.60	0.208
SemEvalT6 - Sentiment	0.54	0.266	0.83	0.042	0.77	0.072	0.77	0.072
Covid Vaccine	0.83	0.042	0.83	0.042	0.83	0.042	0.94	0.005
SemEvalT6 - Abortion	0.49	0.329	0.26	0.623	0.37	0.468	0.49	0.329

Table 6: Model ranks ordered by 0-shot