Do LLMs Have Social Sense? Evaluating Alignment of Conversation Understanding and Reasoning

Zhaoqing Wu and Dan Goldwasser
Department of Computer Science
Purdue Universiy
West Lafayette, IN, USA
wu1828@purdue.edu, dgoldwas@purdue.edu

Abstract

Large language models (LLMs) demonstrate strong instruction-following and reasoning abilities, yet struggle with the socially nuanced understanding humans handle naturally. Humans rely on social sense to interpret intentions, emotions, and relationships in context. We investigate whether and to what extent LLMs exhibit similar abilities when reasoning about social interactions. We propose a structured approach that decomposes social interactions into three aspects and prompts LLMs to generate aspect-based summaries capturing complementary signals. We evaluate these summaries and resulting reasoning within a unified framework across three dimensions, analyzing both LLM–human alignment and the quality of LLM-generated social explanations. Our results show that LLMs achieve moderate agreement with humans and with one another as evaluators and capture some socially nuanced information across aspects, but struggle with interaction-level dynamics shaped by situational norms. Across four downstream CSS tasks, aspect summaries improve aggregate prediction performance across three model families, though effects vary by task. Complementary rationale analysis shows that summaries substantially alter expressed reasoning. Finally, we introduce a benchmark dataset with gold-standard annotations to support systematic study of social understanding and situated reasoning in LLMs.

1 Introduction

Humans possess a natural social sense that allows them to navigate social situations by understanding what to say and how to act given the context and the people involved. Prior work suggests that this ability is an innate human capacity (Kovács et al., 2010), enabling individuals to infer others' beliefs and respond appropriately. This broader social cognitive ability (Markus and Zajonc, 1985) shapes how people interpret others, explain behavior, and reconcile conflicting beliefs in social interactions (Kunda, 1999).

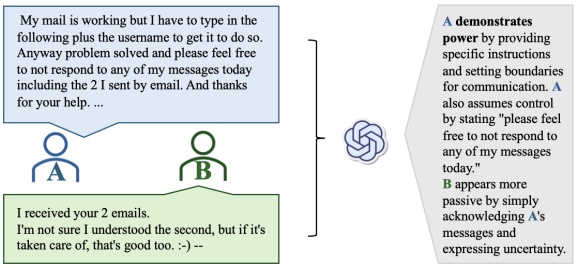


Figure 1: In this conversational exchange, humans can easily infer the social scenario and power dynamics: A asks B for help with a mail issue, implying B is in charge. However, LLMs consistently interpret the relationship in the opposite direction. Their reasoning suggests that they rely heavily on surface-level textual cues to infer social relations, rather than incorporating the broader contextual signals about the social scenario.

To form social judgments, people rely on subtle behavioral and conversational cues. The same behavior can carry different meanings across cultures (Saha et al., 2025), the same utterance can express different intentions depending on tone (Beckman and Venditti, 2010), and similar speaking styles may signal contrasting attitudes through rhetorical patterns (Haas and Flower, 1988). Interpreting such interactions therefore requires integrating multiple signals within their social context.

Given the rapid advancement and widespread use of LLMs in social scenarios, an important question is whether they can interpret these dynamics as humans do. Despite strong performance in instruction following (Zhou et al., 2023; Qin et al., 2024) and multi-step reasoning (Kojima et al., 2023; bench authors, 2023), LLMs still struggle with nuanced social and contextual understanding (Qiu et al., 2025; Choi et al., 2023). At the same time, their growing use in computational social science (CSS) highlights increasing interest in how

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5786–5808
November 6-10, 2026 ©2026 Association for Computational Linguistics

LLMs reason about complex social interactions (Park et al., 2022). Figure 1 shows one example.

In this work, we investigate **whether, and to what extent, LLMs can sense social information in ways similar to humans**. Our notion of *social sense* does not aim to embed formal social theories or psychological mechanisms into LLMs, nor to make them "act like humans". Instead, we compare how LLMs and humans perceive the same social dynamics and evaluate the extent to which their judgments align. Specifically, we ask three questions: (1) given the same information, can LLMs reach the same judgments as humans (Section 5.1)? (2) can LLMs summarize conversational social dynamics across different aspects and make socially informed decisions in a manner similar to human reasoning (Section 5.2)? and (3) do aspect-based representations support downstream CSS decisions and change models' reasoning behaviors when making those decisions (Section 6)?

We study conversations as real interactions between speakers in social settings, where understanding requires navigating complex social dynamics. To examine whether LLMs exhibit social sense, we propose a structured framework that decomposes conversations into distinct aspects capturing different layers of information in the textual exchange, and then integrates these signals to reason about the social scenario and conversational outcomes.

In this work, we use the term ***aspect*** to refer to the components of our conversation decomposition, ***situated reasoning*** to describe the process of reasoning about conversational scenarios, and ***evaluation dimensions*** to denote our framework for assessing explanation quality. We define three aspects: *style*, *content*, and *interaction*, and instruct LLMs to generate summaries that capture these different layers of social meaning. These aspect-level summaries are then integrated to support situated reasoning about the conversation's situation and outcome. Unlike prior work that often assumes LLMs naturally produce reasonable social explanations, we explicitly evaluate their quality and examine whether LLM judgments of explanation quality align with human judgments. To do this, we introduce three evaluation dimensions: *grounding*, *interpretation*, and *completeness*, and conduct human evaluations of aspect-based summaries generated by LLMs, while also comparing these judgments with those produced by LLM as evaluators.

Our findings show that LLMs behave differently from both humans and other LLMs when acting as evaluators. While they can capture some socially nuanced signals across multiple aspects and integrate them into coherent reasoning, they still struggle to interpret interaction-level dynamics beyond surface text and to connect textual descriptions with pragmatic, situation-aware understanding. When applied to downstream CSS tasks, aspect summaries improve aggregate prediction performance across models, but their effects vary substantially across tasks. They provide the clearest gains for empathy, while producing smaller or inconsistent effects for power, persuasion, and toxicity. A complementary rationale analysis shows that summaries change the content of generated explanations and slightly increase the reasoning theme alignment with gold reasons.

The remainder of the paper is organized as follows. Section 3 introduces the aspect decomposition and evaluation dimensions. Section 4 describes the data collection and annotation process for human and LLM social sense judgments. Section 5 presents the analysis of LLM social sense alignment. Section 6 applies our framework to downstream CSS tasks. We summarize our contributions as follows:

1. We introduce a unified framework for understanding social interactions through three distinct aspects and evaluate comprehension using a three-dimensional evaluation framework. We adopt this framework to assess LLMs' social sense and to guide situational aware reasoning.
2. We investigate the alignment between human and LLM perceptions of social information and provide detailed analyses.
3. We evaluate three model families under six prompting conditions on four downstream CSS tasks. We show that aspect summaries yield task-dependent performance gains and substantially change generated rationales.
4. Our human annotation results in a dataset of 290 entries spanning four CSS tasks. Each entry includes the original conversation text, three human-annotated aspect summaries, and one contextually grounded rationale, yielding a total of 1,160 human-annotated instances. This dataset enables systematic and qualitative analysis of social behavior understanding and situational reasoning across diverse interaction dynamics [1].

[1]We release the dataset at https://github.com/kandicew/situated_reasoning_data.

2 Related Work

LLM Reasoning: Extensive research has examined LLMs' ability to explain their decision-making processes (Wei et al., 2023; Li et al., 2023; Fu et al., 2023). One major challenge is that LLMs often struggle to maintain consistent reasoning and beliefs throughout the process Wang et al. (2023). Prompt-based (Amirizaniani et al., 2024), graph-based (Kassner et al., 2023) and generation-based (Singh et al., 2024) have been used to addressed such challenges. Our work integrates reasoning, verification, and explanation capabilities to encourage LLMs to consider multiple perspectives when analyzing complex social issues.

LLM Judges for CSS: LLMs have been increasingly adopted and evaluated across a broad range of CSS tasks (Ziems et al., 2024). Rathje et al. (2024) demonstrates that GPT models can accurately detect psychological constructs using simple prompts. Zheng et al. (2023); Liu et al. (2023) explore LLM's ability as evaluators. Our work examines this alignment specifically for socially grounded evaluations.

Social Reasoning: Social reasoning represents the ability to understand social dynamics and respond appropriately, which is closely related to the concept of Theory of Mind (ToM) (Premack and Woodruff, 1978) in LLMs. Sap et al. (2022) argues that LLMs lack robust theory of mind capabilities. Building on this, subsequent research has developed benchmarks (Le et al., 2019) to evaluate LLMs' strengths and limitations in social reasoning (Xu et al., 2024; Wu et al., 2023), as well as strategies to enhance these skills (Jung et al., 2024) (see (Chen et al., 2025) for an overview). Additionally, Zhou et al. (2024) explores how LLMs navigate social scenarios within simulated environments. Our work focuses on evaluating the social sense of LLMs in the context of real-world tasks, aiming to understand how well they interpret and reason about social interactions grounded in real social settings. Beyond evaluating final decisions, we also compare the themes expressed in model-generated rationales, treating this analysis as a complementary diagnostic of how structured social information changes the explanations accompanying those decisions.

3 Aspects and Evaluation Dimensions

In this section, we introduce the social aspects for decomposing conversations and the evaluation dimensions, showing how they build on and generalize insights from prior research and theoretical foundations.

3.1 Aspects

Style Research in conversation analysis has examined linguistic features ranging from basic spoken units such as utterances (Sacks et al., 1974) and sentence structures (Stolcke et al., 2000) to higher-level properties such as shifts in tone (Hua et al., 2024). Stylistic choices vary with social and individual factors and shape conversational dynamics through lexical features (Bramsen et al., 2011), discourse structure (Prabhakaran and Rambow, 2013), and conversational flow (Liu et al., 2020).

Content Conversations are fundamentally shaped by what is discussed. Early work shows how interactions evolve as participants exchange information (Berger and Calabrese, 1975), while group discussions often reveal biases in what information is shared or emphasized (Stasser and Titus, 1985). Recent work leverages LLMs to analyze conversational content at scale, capturing sentiment (Zhang et al., 2024), framing (Alonso del Barrio and Gatica-Perez, 2023), and related meanings to reveal social and ideological patterns.

Interaction Conversation unfolds through the structure of turns and responses. Conversation analysis highlights the role of turn-taking in coordinating interaction (Sacks et al., 1974) and reflects how talk can signal workplace hierarchies (Holmes and Stubbe, 2015). Linguistic coordination between speakers has also been linked to power relations (Danescu-Niculescu-Mizil et al., 2012), and recent work summarizes interaction trajectories to predict conversational outcomes (Hua et al., 2024).

Together, decomposing conversations into *style* (how people speak), *content* (what is discussed), and *interaction* (how participants engage) builds on classic social and linguistic theory while leveraging modern computational methods. These complementary dimensions capture signals that help reveal social context and conversational dynamics, and provide a structured foundation for reasoning about social situations. We provide detailed definitions of these aspects in Appendix A.

3.2 Evaluation Dimensions

Grounding Evaluating grounding (or hallucination) is essential because LLMs can produce content that appears plausible but is unsupported, leading to misinformation or unfaithful summaries

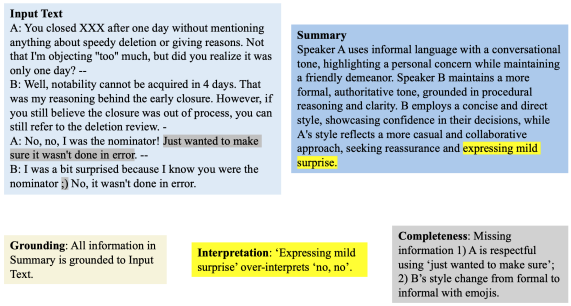
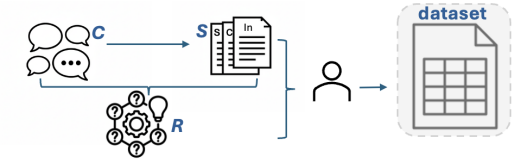
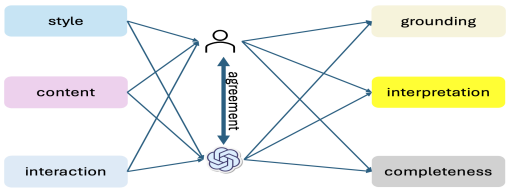


Figure 2: An example summary evaluation: the upper left is a input text, and the upper right is a summary of the style aspect for this input; on the bottom are the critiques based on the three evaluation dimensions.



(a) Human annotate conversation summaries and reasons.



(b) Our proposed two-step framework evaluates an LLM's understanding of social interactions. First, an input interaction is distilled into three *aspect*-based summaries (left). These summaries are then assessed by both LLM-as-a-judge and human annotators, using a unified set of evaluation dimensions (right).

Figure 3: Overview of data collection and evaluation.

(Maynez et al., 2020). Recent surveys and benchmarks focus on factuality and hallucination detection (Qi et al., 2025; Rahman et al., 2025). For social interaction analysis, grounded summaries ensure that observations about conversational aspects are supported by the original dialogue, providing a reliable basis for higher-level social inference.

Interpretation Evaluating interpretation examines whether summaries draw valid and well-supported conclusions from the text. LLMs may generate plausible but unsupported explanations or over-interpret subtle cues (Singh et al., 2024; Turpin et al., 2023). In social interactions, such errors often arise when models infer meaning from implicit behavioral signals, including cues related to social context (Kulkarni et al., 2021) and cultural background (Pawar et al., 2024).

Completeness Evaluating completeness assesses whether summaries capture the key information necessary to understand the interaction, including contextual cues that shape the social scenario. Prior work applies this criterion to evaluate LLM-generated content and design structured benchmarks that measure information coverage (Kassner et al., 2023; Lee et al., 2024; Min et al., 2025).

Together, evaluating *grounding* (factual accuracy), *interpretation* (sound reasoning), and *completeness* (information coverage) provides a structured framework for assessing LLM-generated summaries used to understand and analyze social interactions. An example evaluation is shown in Figure 2.

4 Methods and Data

We adopt a generalized two-step framework to obtain and evaluate LLMs' understanding of social interactions. We use real conversational exchanges from established CSS tasks: predicting power relations, persuasiveness, toxicity, and empathy, for which humans rely on social sense to navigate complex social dynamics (task details in Appendix B).

First, for each conversation, we generate three summaries corresponding to the defined aspects. Next, we integrate these summaries to produce situated reasoning that predicts the outcome of the conversation. In both steps, the generated summaries and reasoning are evaluated using the three evaluation dimensions. Figure 3 illustrates the overall workflow.

4.1 Aspect Summary Generation

Prior work (Hua et al., 2024) shows that topic-focused summarization can improve judgment quality for both humans and LLMs. Building on this idea, we generate summaries of each conversation based on the aspects, each highlighting a distinct perspective to provide a more comprehensive view. This decomposition encourages LLMs to consider multiple linguistic and contextual cues rather than relying on a single perspective that may introduce bias or misinterpretation. Formally, given a conversation C, and the definition of aspect A, we generate the aspect-specific summaries **S** from the

summary prompt P_S.

$$S_A = P_S(C, A), \quad A \in \{style, content, interaction\}$$

4.2 Situated Reasoning Generation

We use the aspect-based summaries from the previous step to provide complementary perspectives that support a more holistic understanding of the conversation and its outcome. By integrating these summaries, we generate situated reasoning that explains the dynamics underlying the correct social label. Given a conversation C, its set of three aspect summaries $\mathbf{S}_A = [S_{Style}, S_{Content}, S_{Interaction}]$ and the social outcome O (i.e, label of the CSS task), we generate the situated reasoning R using the reason prompt P_R.

$$R = P_R(C, \mathbf{S}_A | O)$$

4.3 Evaluation

Summary quality is evaluated on a 3-point Likert scale, widely used in natural language generation (Amidei et al., 2019), for each evaluation dimension. The score of a summary S or a reason R along dimension $d \in \{grounding, interpretation, completeness\}$ is computed via a function f applied to conversation C and the generated text S or R, conditioned on the Likert scale of the dimensions L^d.

$$Score^{d \in g,i,c}_{S \text{ or } R} = f(C, S \text{ or } R | L^d), \quad Score \in [1, 2, 3]$$

The Likert scales and prompts are detailed in Appendix D. The scoring function f may be implemented through either LLM-based evaluation or human annotation.

4.3.1 Human Annotation

Human annotators evaluate and refine summaries generated by a base language model (GPT-5). For each summary, annotators assign Likert-scale scores across the three evaluation dimensions for each aspect. These ratings are used to benchmark LLM evaluators and measure agreement with human judgments. When a summary does not receive a perfect score, annotators revise it to better align with the intended aspect, producing gold-standard aspect summaries.

This process yields three human-aligned summaries for each conversation, one for each aspect, that capture the corresponding signals in the conversation. In addition, annotators refine the situated reasoning to produce a situated gold rationale that integrates the aspect summaries and the original conversation to explain the social outcome (i.e., the label of the CSS task). Human annotate 290 conversations, resulting in 870 summaries and 290 situated reasoning. Annotation details are provided in the Appendix B.2.

5 Results: LLM Social Sense Alignment

We use our dataset to benchmark LLMs along two capabilities. Section 5.1 evaluates whether LLMs assess summaries and reasoning in ways similar to humans, measured by alignment with human judgment scores. Section 5.2 evaluates LLMs' ability to generate aspect summaries and situated reasoning, compared against human-refined gold summaries and gold rationales.

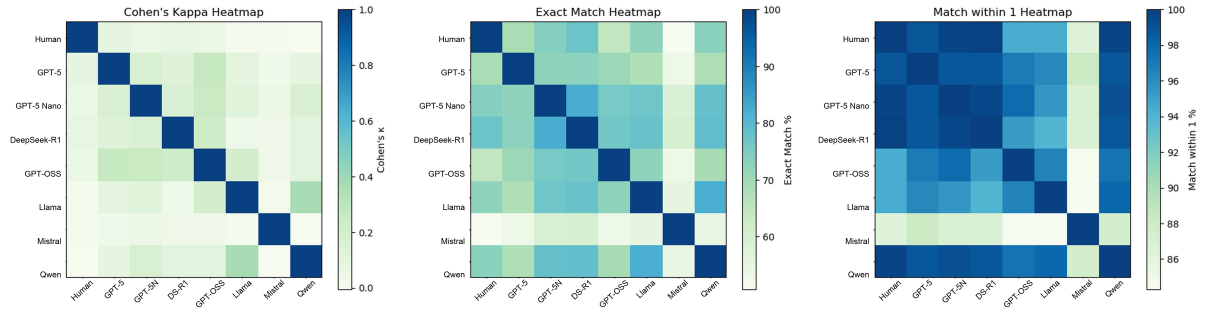
5.1 LLM as Judge

We evaluate agreement by asking multiple LLMs to score the same summaries using the same criteria, and then comparing their scores with those of human judges. We select seven LLMs of varying sizes, including both open-source and closed-source models, namely GPT-5 (OpenAI, 2025a), GPT-5 Nano (OpenAI, 2025a), DeepSeek-R1-70B (Guo et al., 2025), gpt-oss-120b (OpenAI, 2025b), Llama 4 (Meta AI, 2025), Mistral 7B (Jiang et al., 2023), Qwen2.5-72B (Qwen et al., 2025).

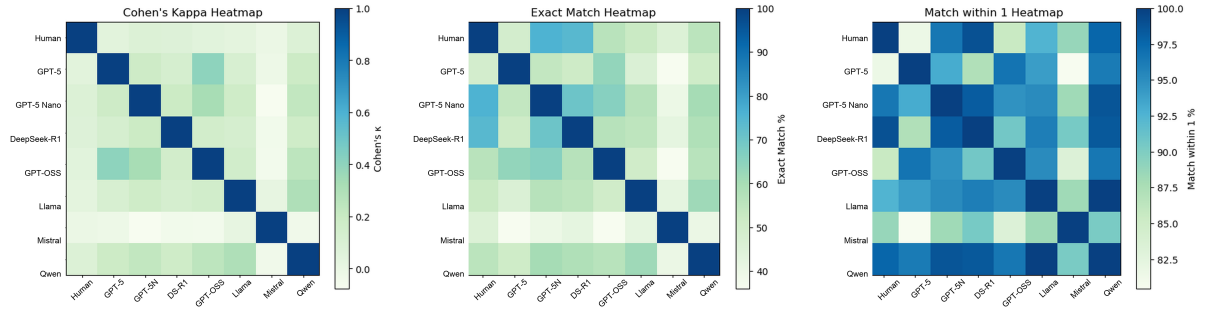
For each conversation, we obtain nine scores for the summaries: three evaluation dimensions for each of the three aspects, and three scores for the situated reasoning: one for each evaluation dimension. Using these scores, we measure evaluator agreement in three ways: **Cohen's** κ to assess overall alignment, **exact match** to count identical scores, and **within-1 match** to capture cases where evaluators (human or LLM) differ by at most one point, indicating non-contradictory judgments. Figure 4 reports these results.

Overall, models exhibit noticeably different behaviors under the same instructions. In general, they show moderate agreement with humans and with each other, achieving over 50% exact match and positive Cohen's κ values. We observe slightly higher agreement among models from the same family (e.g., gpt-oss, GPT-5, and GPT-5 Nano).

For aspect summary evaluation, most models assign the same scores as humans more than 60%



(a) Human-LLM agreement on summary scores.



(b) Human-LLM agreement on reason scores.

Figure 4: Human-LLM alignment: The first column shows the Cohen's κ. The second column reports the exact match rate, where two evaluators assign the same rating. The third column reports the within-1-point match rate, where ratings are either identical or differ by at most one point on the three-point Likert scale, indicating that the evaluators do not give opposite judgments.

of the time and produce non-contradictory judgments in about 92% of cases. Mistral appears to be an outlier, likely due to occasional failures to follow the evaluation instructions. We also find that larger models do not consistently show better alignment with human judgments than smaller models. Evaluation of situated reasoning follows a similar trend but with lower agreement overall. This suggests that judging social reasoning differs from judging summaries and remains more challenging for LLMs, highlighting a gap between model and human reasoning about social scenarios.

We further analyze how LLM-judge scores align with human judgments for aspect summaries. Figure 5 presents the detailed agreement patterns across evaluation scores.

Across all models, the content aspect achieves the highest exact match with human judgments. This is expected, as content primarily reflects topic-related information that can be identified directly from the text. In contrast, the largest disagreements occur in the grounding dimension of style aspect, which relies on pragmatic signals that vary by situations and can diverge based on lexical cues.

For grounding and completeness, LLMs tend to be stricter than human judges, assigning lower scores when information is sufficiently supported. For interpretation, particularly in the style and interaction aspects, LLMs consistently assign scores about one point higher than humans. This suggests that LLMs are more lenient and less likely to identify gaps in interpretation, whereas human annotators more often recognize missing or incorrect inferences. In other words, models tend to overlook issues in how language is used and how interactions between speakers are established. This pattern suggests that while models can verify textual support, they still overlook how language is used and how interactions are structured in social dynamics.

5.2 LLM as Reasoner

We evaluate six LLMs on their ability to generate aspect-based summaries and situated reasoning. Generated outputs are assessed against human-annotated gold standards using our evaluation dimensions. For this evaluation, we use Llama-4 (Meta AI, 2025) as an automated judge, treating

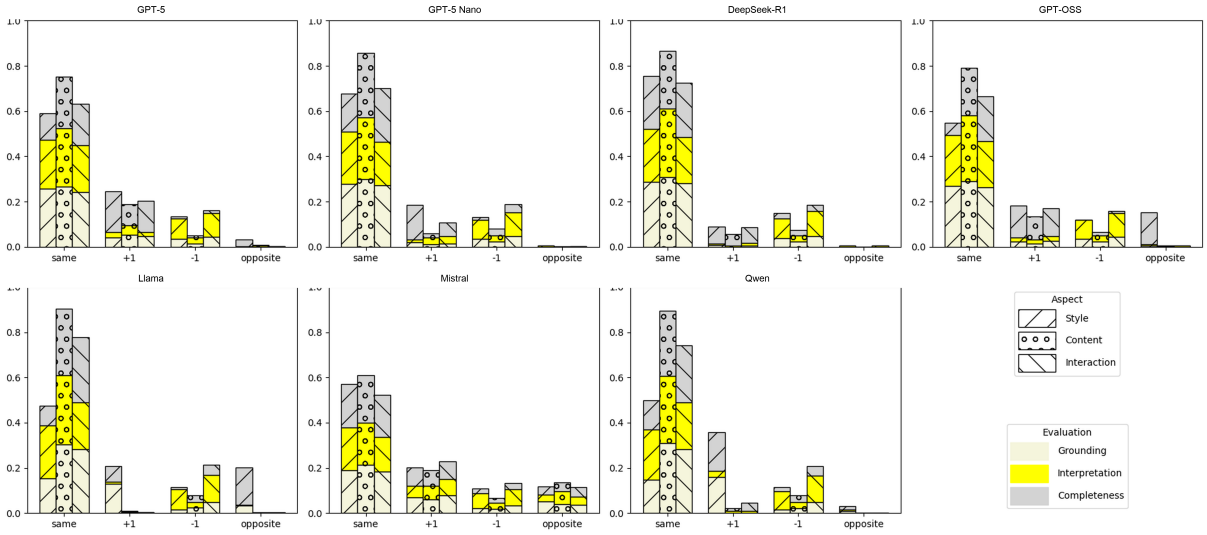


Figure 5: Human-LLM agreement for all models across aspects and evaluation dimensions: A value of +1 indicates that the human assigns a score one point higher than the LLM, suggesting that the LLM may be stricter, judging as incorrect what humans consider acceptable. A value of –1 captures the reverse pattern, where the LLM fails to detect errors that humans view as important. "Opposite" indicate a difference of 2 points given the 3-point scale, meaning the human and the LLM give inverted judgments (e.g., very good versus very poor).

the task as text matching with the gold annotations as reference. While using an LLM as a judge may produce different results than human evaluation, applying the same model consistently ensures internal comparability across all results. Llama-4 was selected because it was not part of the benchmarked models and demonstrated the highest agreement with human judgments among all tested models. We provide model and experiment details in Appendix E.1.

Figure 6 compares the quality of summaries generated by various LLMs against human annotations. Overall, GPT-5 achieves the highest scores, while other models perform moderately, around 60%, reflecting differences in model size and their ability to generate text and follow instructions.

Completeness is consistently lower for LLMs, indicating that their summaries often omit details relevant for human social understanding of conversations. Interpretation scores appear relatively high, but this does not fully align with human judgments, likely due to limitations of LLM used as the judge. Across all models, the content aspect consistently scores highest, whereas style and interaction scores are lower, reflecting the challenge of capturing subtle conversational dynamics.

We filter examples that receive a score of 1 on the interaction, worst-performing aspect, within the lowest-rated dimension, completeness. The most common issue is **content focus instead of social focus** (38.9%): LLM-generated summaries often describe the conversation content and become a narrative of what was said, whereas human summaries focus on the speakers' actions and how they shape the interaction. Another frequent issue is the **omission of interactive roles** (22.2%): LLM summaries tend to use neutral, descriptive verbs such as "mention" or "collaborate," while human summaries interpret speaker relationships and the function of utterances, using more interaction-oriented language such as "take initiative" or "encourage". Examples are provided in Appendix E.2.1

For evaluating situated reasoning, we find that scores are generally higher than those for summaries, suggesting that LLMs can produce high-quality reasoning when conditioned on aspect summaries and gold labels. This conditioning helps models identify the conversational situation and anchor their reasoning in the relevant evidence. Across models, completeness remains consistently lower than the other dimensions. The most common omission in LLM-generated reasoning is the lack of relevant social-context cues that contribute to the conversation outcome. Figure 8 presents the quality scores of LLM-generated reasoning compared to gold-standard human rationales.

6 Results: Downstream CSS Tasks

We evaluate whether aspect summaries support four downstream CSS decisions: power, persua-

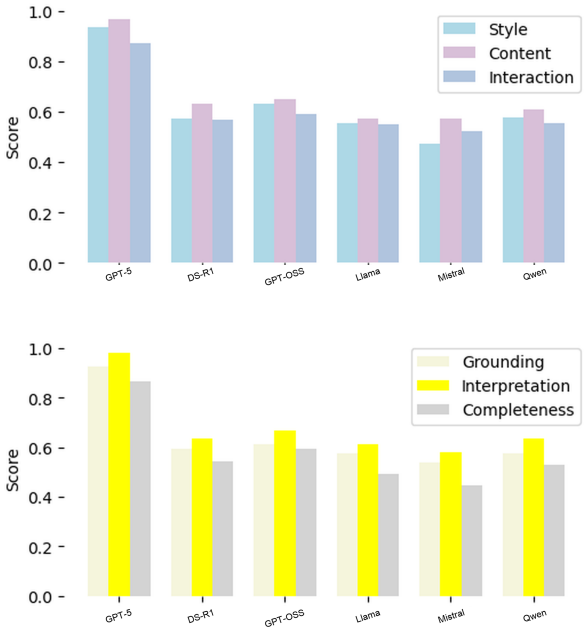


Figure 6: Summary quality of various LLMs, using human annotation as gold standard, and evaluated across aspects and dimensions.

sion, empathy, and toxicity. For each task, we sample 1,000 conversations and compare six prompting conditions: a direct prompt; a prompt augmented with the three aspect summaries; direct and summary-augmented chain-of-thought prompts; and situation-reasoning prompts that ask model to first identify situations of the conversation then make prediction using either the original text or the text together with the summaries. We evaluate GPT-4o Mini (OpenAI, 2024), Gemma3-27B (Team et al., 2025), and Qwen2.5-72B (Qwen et al., 2025). We report accuracy and macro F1; task-level results are provided in Appendix F.

6.1 Effects of Aspect Summaries

Table 1 shows that adding aspect summaries improves performance over the corresponding matched text-only condition for every model and reasoning family. This consistent improvement across models suggests that aspect summaries provide useful intermediate representations for downstream CSS prediction. However, the best-performing prompt varies across models, indicating that adding more reasoning steps does not necessarily yield better prediction performance. These results further suggest that LLMs continue to struggle to identify the socially relevant situational information needed to make accurate social predictions.

We further investigate whether aspect summaries improve the selection of few-shot demonstrations. For each query and task label, we use BM25 (Lin et al., 2021) to retrieve the most similar example from the human-annotated data, selecting one example per label. Retrieval is performed either over the original conversation text or over the concatenated aspect summaries. The retrieved examples are then used as few-shot demonstrations for in-context learning, and GPT-4o Mini predicts the final label. Table 2 shows that summary-based retrieval improves overall accuracy by 1.45% and macro F1 by 1.94%. The gains vary across tasks: Power shows the largest improvement in accuracy, followed by Toxicity and Empathy, whereas Persuasion shows a decline. Additional experimental settings are reported in Appendix Table 8.

6.2 Reasoning Quality

To examine whether and how summaries change model reasoning on CSS tasks, we identify common themes in generated rationales by clustering their embeddings and assigning each rationale to one or more discovered themes. The full experimental protocol is provided in Appendix F.3. This analysis is intended as a discovery and explanatory tool for comparing reasoning across methods, rather than as a human-validated annotation of gold reasoning. Accordingly, the resulting clusters should be interpreted as explanatory themes rather than definitive representations of accurate reasoning.

We separately construct a reference theme space from GPT-5 rationales generated using the gold task outcomes. As shown in Section 5.2, GPT-5 exhibits the highest alignment with human reasoning among the evaluated models. We therefore use GPT-5-generated rationales as silver-label references. These references are not human-validated proofs of correctness, nor do they provide evidence of a model's internal reasoning. Full details on clustering, theme naming and assignment, calibration, and sensitivity analyses are reported in Appendix F.3.

The generated-theme analysis indicates that summaries substantially redirect the content of model explanations: 87.3% of paired rationales generated from the same prompt using either text alone or text augmented with summaries are assigned different themes. Relative to the GPT-5 reference space, summaries increase the probability that a rationale shares at least one reference theme by 1.3%. De-

Method	GPT-4o Mini		Gemma3-27B		Qwen2.5-72B	
	Acc.	Macro F1	Acc.	Macro F1	Acc.	Macro F1
Direct prompt	62.14	59.18	64.92	64.58	61.58	59.26
Summary prompt	**64.42**	**62.10**	65.67	64.30	**64.12**	**61.17**
Direct CoT	61.89	58.28	63.49	61.95	61.08	58.15
Summary CoT	63.84	59.73	**66.62**	**66.59**	63.11	60.01
Text situation	61.91	59.23	63.26	61.81	60.63	57.67
Text + summary situation	64.29	61.98	65.35	65.27	64.07	60.72

Table 1: Aggregate downstream performance across tasks. Bold marks the best result for each model and metric.

Task	Text Acc.	Text F1	Summary Acc.	Summary F1	Δ Acc.	Δ F1
Power	59.80	57.34	64.40	64.25	+4.60	+6.91
Persuasion	55.60	52.69	54.30	48.37	-1.30	-4.32
Empathy	81.80	81.76	82.10	82.10	+0.30	+0.34
Toxicity	52.70	41.91	54.90	46.76	+2.20	+4.85
Overall	62.48	58.43	**63.93**	**60.37**	+1.45	+1.94

Table 2: GPT-4o Mini label-balanced BM25 few-shot results (%). One annotated example is retrieved per label using either original-text or aspect-summary similarity. "Overall" is the unweighted mean across four tasks.

tailed task-level results are provided in Appendix Table 10.

Together, these results show that summaries change what models discuss and modestly increase the likelihood that at least one reference theme appears, but do not consistently make the complete rationale more aligned with the reference. This pattern also varies across tasks. The divergence is particularly evident for Empathy, where prediction accuracy improves substantially while changes in rationale alignment are minimal. Thus, downstream performance gains should not be interpreted as evidence that summaries globally improve or correct model reasoning; predictive performance and expressed-rationale alignment capture related but distinct effects.

7 Conclusion and Future Work

In this work, we propose a general framework for probing nuanced social dynamics in LLMs using three social aspects and evaluating their outputs along three dimensions. We compare LLM judgments with human judgments, highlighting variation across aspects, dimensions, and models. On four downstream CSS tasks, aspect summaries improve aggregate prediction performance across three model families, with the clearest gains for empathy and mixed effects elsewhere. Our complementary rationale analysis further shows that summaries substantially redirect expressed reasoning and modestly increase overlap with GPT-5 reference theme, without uniformly improving full-rationale alignment. Finally, we present an annotated dataset that provides a foundation for future research on enabling LLMs to interpret and reason more effectively about nuanced social dynamics.

In future work, we aim to further investigate how different social situations shape the interpretation of behaviors and interactions, and how LLMs can be guided to model, simulate, and distinguish such contextual variations in social reasoning.

Limitations

We acknowledge several limitations of the current work:

Cultural scope: Our annotated data and downstream experiments are limited to English conversations in a U.S.-centric cultural context. While the proposed framework could be extended to other cultural settings, the findings reported here primarily apply to the domain studied.

Narrow focus on CSS tasks: Our analysis is based on a relatively small dataset and a limited set of computational social science tasks. These tasks cannot fully represent the underlying social concepts they aim to capture, nor do they cover the full range of real-world social reasoning challenges.

Aspect coverage: Social dynamics are inherently complex and multifaceted. Although the proposed aspects provide a useful decomposition, they may not capture all interacting behaviors and contextual nuances. Moreover, the relative importance of different aspects may vary across cultural settings, tasks, or social contexts. Exploring how aspect weighting should adapt across scenarios is beyond the scope of this work but represents an important direction for future research.

Statistical improvements: Improvements on

downstream tasks are not consistently statistically significant. This may stem from dataset variability as well as variation in human annotations that are not always strictly tied to explicit social guidelines. Our goal is therefore not to maximize performance gains but to identify gaps in LLM social reasoning and provide a structured framework for analyzing them.

Rationale and theme analysis: Generated rationales are post-hoc textual outputs and may not faithfully reflect a model's internal computation. The theme taxonomies and assignments are produced by an outcome-blind clustering and calibrated assignment pipeline, but they are not independently human-validated. Likewise, GPT-5 rationales conditioned on the gold outcome provide an operational reference rather than a definitive account of correct human reasoning. The rationale results should therefore be read as evidence about changes in expressed explanations and their alignment with this reference, not as causal evidence that summaries improve reasoning.

Finally, this framework is intended for research and analysis of LLM behavior, rather than immediate deployment. Given the remaining challenges in nuanced social reasoning, we do not recommend relying on LLMs for real-world applications requiring sensitive social judgment without further validation.

Ethics Statement

Human annotation for this study was conducted on a voluntary basis. The language data used in our experiment and analysis may contain biased viewpoints or harmful language. Such content reflects the nature of real-world interactions captured in the original datasets or LLM generation from these and does not represent the views of the authors.

We also acknowledge the risk that the capabilities of LLMs in performing or evaluating social reasoning may be overstated. While our work studies how LLMs interpret social dynamics, it does not imply that these systems possess genuine social understanding. Claims about AI's ability to reason about social situations should therefore be interpreted cautiously and evaluated with rigorous human oversight.

References

David Alonso del Barrio and Daniel Gatica-Perez. 2023. Framing the news: From human perception to large language model inferences. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 627–635, New York, NY, USA. Association for Computing Machinery.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 34–44, New York, NY, USA. Association for Computing Machinery.

Mary E. Beckman and Jennifer J. Venditti. 2010. Tone and intonation. In *The Handbook of Phonetic Sciences*, chapter 16, pages 603–652. John Wiley & Sons.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Charles R. Berger and Richard J. Calabrese. 1975. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, 1(2):99–112.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. *Preprint*, arXiv:2505.00026.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power:

Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. *Preprint*, arXiv:2210.00720.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Christina Haas and Linda Flower. 1988. Rhetorical reading strategies and the construction of meaning. *College Composition & Communication*, 39(2):167–183.

Janet Holmes and Maria Stubbe. 2015. *Power and Politeness in the Workplace: A Sociolinguistic Analysis of Talk at Work*, 2 edition. Routledge.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. How did we get here? summarizing conversation dynamics. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Ágnes Melinda Kovács, Ernő Téglás, and Ansgar Denis Endress. 2010. The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012):1830–1834.

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An approach for socially sensitive pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziva Kunda. 1999. *Social Cognition: Making Sense of People*. MIT Press, Cambridge, MA.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making large language models better reasoners with step-aware verifier. *Preprint*, arXiv:2206.02336.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Yafei Liu, Hongjin Qian, Hengpeng Xu, and Jinmao Wei. 2020. Speaker or listener? the role of a dialog agent. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4861–4869, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Hazel Markus and Robert B Zajonc. 1985. The cognitive perspective in social psychology. *Handbook of social psychology*, 1(1):137–230.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. 2025a. Introducing gpt-5 for developers. https://openai.com/index/introducing-gpt-5-for-developers/.

OpenAI. 2025b. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *Preprint*, arXiv:2411.00860.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 216–224, Nagoya, Japan. Asian Federation of Natural Language Processing.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Siya Qi, Rui Cao, Yulan He, and Zheng Yuan. 2025. Evaluating LLMs' assessment of mixed-context hallucination through the lens of summarization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16480–16503, Vienna, Austria. Association for Computational Linguistics.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Subhey Sadi Rahman, Md. Adnanul Islam, Md. Mahbub Alam, Musarrat Zeba, Md. Abdur Rahman, Sadia Sultana Chowa, Mohaimenul Azam Khan Raiaan, and Sami Azam. 2025. Hallucination to truth: A review of fact-checking and factuality evaluation in large language models. *Preprint*, arXiv:2508.03860.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025. Meta-cultural competence: Climbing the right hill of cultural awareness. In *Proceedings of the 2025 Conference of the Nations of the*

Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *Preprint*, arXiv:2402.01761.

Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Aspects

We define the generalized aspects as the following.
Style: Style refers to the patterned and distinctive ways individuals use linguistic and paralinguistic resources, including vocabulary choices (e.g., formal vs. informal, jargon, technical terms), sentence structure (e.g., simple vs. complex, passive vs. active voice, syntactic patterns), and tone (e.g., sarcasm, warmth, assertiveness), to express meaning, reflect manner and social beliefs, construct identity, and manage relationships in conversation.
Content: Content represents the substantive and expressive dimension of communication, encompassing the topics, themes, and informational elements exchanged between participants. It reflects both the cognitive (what is being discussed or conveyed) and affective (how it is expressed and interpreted) layers of meaning. Content involves the subject matter, evidence, framing, and contextual cues that shape interpretation—such as sentiment, emphasis, and perspective—and includes the emotional and exploratory processes through which speakers articulate experiences, uncover underlying meanings, and co-construct shared understanding.
Interaction: Interaction is the dynamic, collaborative process through which participants co-construct meaning, manage social relationships, and influence each other's understanding. It encompasses alignment of language, mutual engagement, and both conscious and unconscious adaptations to each other's communicative behaviors. Interaction involves cognitive, emotional, and social dimensions, including active listening, questioning, empathetic responses, use of politeness strategies (e.g., gratitude, indirectness, deference), framing techniques, rhetorical devices (e.g., analogies, counterarguments), and modulation of tone or emotional stance. Through these processes, participants negotiate meaning, explore perspectives, and co-create shared understanding, while responding to and shaping the relational and situational context.

B Dataset Details

B.1 Task

In this study, we experiment with four well-established tasks in computational social science: *power*, *persuasion*, *toxicity*, and *empathy*. Each of these tasks requires reasoning over diverse forms of explicit and implicit information, including linguistic cues, interpersonal dynamics, and broader social context, which have been challenging for LLMs. We frame each task as a binary classification problem, ensuring a balanced distribution across both labels. We define the tasks and list the corresponding datasets in Table 3.

B.1.1 Power

Identifying power dynamics in conversations remains a challenge for LLMs. Prior work by Ziems et al. (2024) demonstrates that LLMs still have significant room for improvement on this task, using a dataset of online Wikipedia discussions between administrators and users originally introduced by Danescu-Niculescu-Mizil et al. (2012). Interactions in such contexts exhibit a wide range of conversational behaviors both within individual utterances as well as across the exchange between participants. These variations give rise to multiple expressions of power dynamics in textual exchanges. Therefore, a comprehensive analysis from multiple perspectives is necessary to form an accurate overall judgment.

B.1.2 Persuasion

Determining persuasiveness has been a challenge for LLMs, as persuasive scenarios often involve conflicting viewpoints and subjective judgments about what is acceptable. Language plays a critical role in persuasion, shaping both the conveyed message and the audience's overall impression, which ultimately influences whether they are convinced.

For this part of the experiment, we adopt the persuasion task introduced in Althoff et al. (2014). Although persuasiveness is inherently subjective, there are commonly shared indicators of what makes a message persuasive.

B.1.3 Toxicity

We adopt the task introduced by Zhang et al. (2018), which involves predicting the likelihood of toxicity in future turns of a conversation based on its initial exchanges. The core challenge lies in detecting subtle, non-explicit information embedded in both the text and the interaction dynamics that may

Task	Description	Dataset
Power	Predict who has more power in a two-speaker interaction	Wikipedia Talk Pages Corpus (Danescu-Niculescu-Mizil et al., 2012)
Persuasion	Determine whether a post is persuasive	Random Acts of Pizza (Althoff et al., 2014)
Toxicity	Predict whether a discussion details in future turns	Conversations Gone Awry Dataset (Zhang et al., 2018; Hua et al., 2024)
Empathy	Determine whether the response shows empathy	TalkLife Dataset (Sharma et al., 2020)

Table 3: Task descriptions and datasets used in this study.

signal a shift toward toxicity. This task is particularly difficult for LLMs, as it requires a nuanced understanding of the conversational atmosphere especially cues that suggest potential conflict or give rise to tension. While humans can infer attitudes and interpret tone to anticipate such shifts, LLMs often struggle to situate themselves within these contexts using only textual input.

B.1.4 Empathy

Detecting empathy is a challenge for LLMs because, unlike humans who build connections through shared experiences and memories, they lack such personal experiences and contextual grounding to simulate such connections. Sharma et al. (2020) introduces a framework for categorizing different expressions of empathy in text, along with a corpus of post and response to mental health seekers. While their work focuses on identifying common strategies used to convey empathy, we simplify the task in our study to predict whether a given text expresses empathy.

B.2 Data Collect and Annotation

We construct our annotation dataset using the tasks listed in Table 3. From each dataset, we sample a subset of conversations for manual annotation. To ensure balanced evaluation, the annotated conversations for each task are selected such that the label distribution in the annotation set is balanced across task outcomes.

Our annotators are Computer Science PhDs students and faculty who are experienced in NLP research. The annotation process is conducted in two phases. In the first phase, we annotate 35 conversations for each task. During this stage, each instance is annotated by two annotators independently. We achieve an inter-annotator agreement of 82%. Based on annotator feedback and manual inspection of the initial annotations, we identify that the persuasion task is substantially more difficult to annotate than the other tasks. Annotators report that persuasion outcomes are often ambiguous and appear more random, making it significantly harder to identify clear reasoning patterns compared to the other tasks. Due to these challenges, we exclude the persuasion task from the subsequent phase of annotation.

	Ground	Interpete	Complete
Style	0.962	0.902	0.962
Content	0.976	0.973	0.964
Interaction	0.951	0.876	0.956
Reason	0.991	0.969	0.962

Table 4: Scores for GPT-5 generated summaries (upper) and reasons (lower), calculated by human assigned scores.

In the second phase, incorporating the refined guidelines and a new annotation interface 7, we first reannotate the examples in phase one, then we focus on the remaining three tasks and annotate an additional 50 conversations for each task. In this phase, each instance is annotated by a single annotator, as the annotation guidelines have been validated during the first phase.

Overall, we annotate 290 organic social interaction conversations. Each conversation is summarized from three aspects, evaluated and corrected across three dimensions, producing a total of 870 high-quality summaries. In addition, for each conversation we annotate one reasoning explanation that leads to the correct prediction of the corresponding CSS task outcome, resulting in 290 high-quality social reasoning annotations.

B.2.1 Human Annotation Analysis

Table 4 presents human scores, across aspects and evaluation dimensions, for GPT-5-generated summaries (upper) and reasoning (lower) as a reference. Human scores indicate that GPT-5 summaries are generally strong, with content rated highest and interaction lower, likely because capturing speaker interactions requires reasoning over conversation dynamics and context beyond the text. Interpreta-

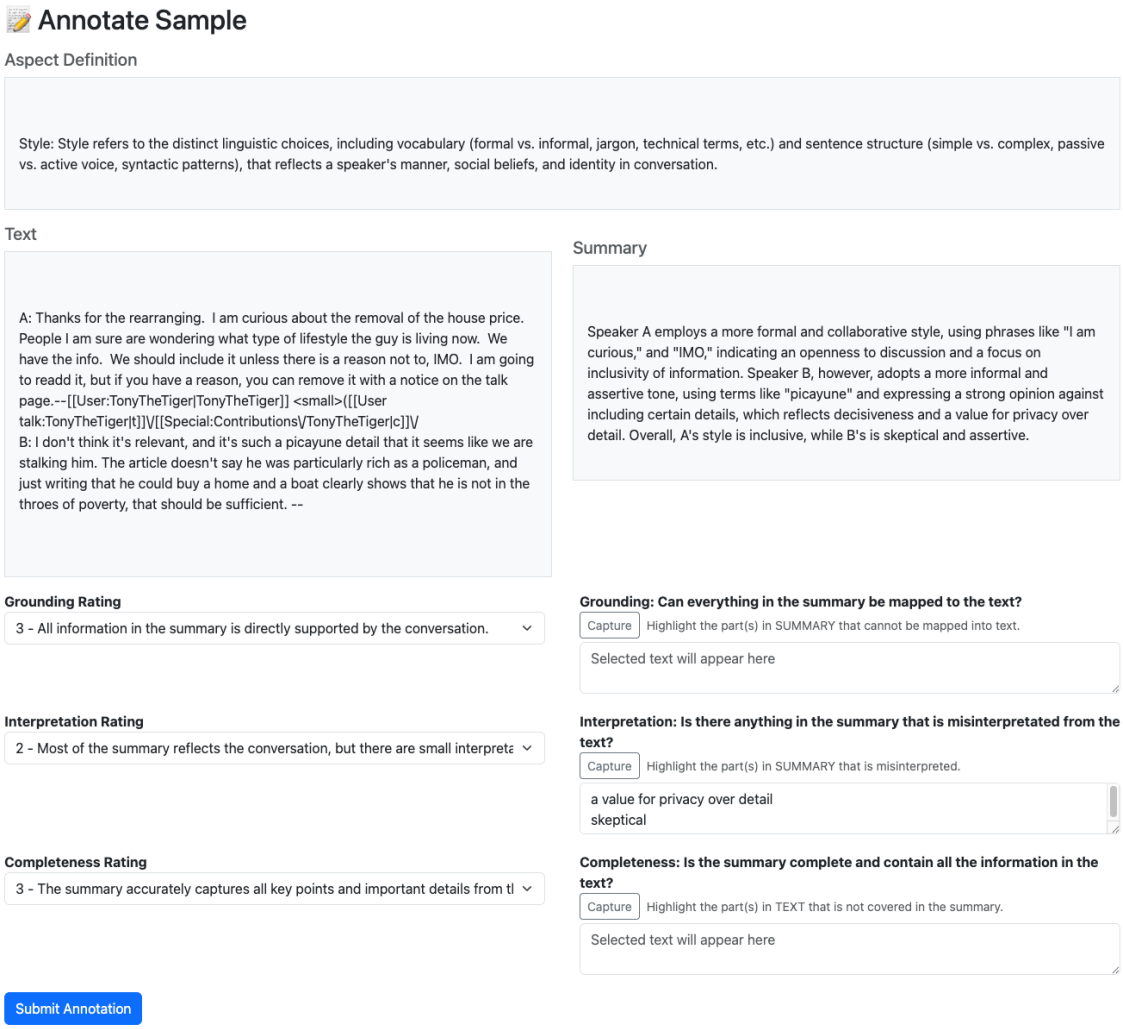


Figure 7: User interface for human annotation.

tion also scores lower, as it involves scenario-level reasoning. The scores for reasons show that LLMs can generate human-aligned reasoning when provided with aspect information and conversation outcomes.

Since there is disagreement between models and humans, we further investigate if higher-quality summaries consistently lead to correct conversation outcome predictions (Table 5). We use three different LLMs (Touvron et al., 2023; OpenAI, 2025b; Jiang et al., 2023) to make predictions on the downstream task using the annotated dataset. Including the gold summaries as part of the input provides a constant signal for the models to reason toward the correct prediction (upper table). For each model, we separate the summaries associated with correctly predicted instances from those associated with incorrectly predicted instances when summaries are included in the input. We find that, across all models, the average human rating of summaries for correctly predicted instances is consistently higher than that of incorrectly predicted ones.

	Llama13b	gptoss120b	Mistral7b
w/o summary	0.567	0.627	0.513
w/ summary	0.653	0.687	0.593
Correct	0.946	0.946	0.954
Incorrect	0.942	0.943	0.935

Table 5: CSS tasks prediction accuracy with and without summary as part of input (upper); average human scores for summaries for correctly and incorrectly predicted instances (lower).

C Prompts

C.1 Summary Generation

```
<aspect definition>
Write a summary of the <aspect> of the
following text. Limit the summary to 80
```

words.
Text:
<text>

C.2 LLM as Judge

The <aspect> of a textual communication is defined as <aspect definition>.
Rate the overall correctness of the <aspect> SUMMARY of the TEXT based on the following likert scale. Return your response in the following strict JSON format: {'score': <integer>}.
<likert scale D>
TEXT:
<text>
<aspect> SUMMARY:

C.3 LLM as Reasoner

You are given a summary and a gold standard summary of a conversation. Please compare the two summaries based on the following criteria using a scale from 1 to 3, where 1 indicates poor alignment and 3 indicates excellent alignment:
Grounding: Can everything in the summary be mapped to the gold summary?
Interpretation: Is there anything in the summary that is shows misinterpretation based on the gold summary?
Completeness: Is the summary complete and contain all the information in the gold summary?
Please provide your evaluation in the following format, only output the scores without any additional text:
Grounding: <score>
Interpretation: <score>
Completeness: <score>

C.4 Downstream Prompts

C.4.1 Power

Direct prompt.
You are a careful text classifier. Use only the supplied evidence.
Question: Which speaker has more power in this conversation?
Labels: A means Speaker A; B means Speaker B.
Return exactly one label (A or B) and no explanation, using:
<prediction>LABEL</prediction>
<evidence>
Answer:

CoT prompt.
You are a careful text classifier. Analyze only the supplied evidence.
Question: Which speaker has more power in this conversation?
Labels: A means Speaker A; B means Speaker B.
First provide concise reasoning grounded in the evidence, then give one final label. Return exactly:
<reasoning>concise reasoning</reasoning>
<prediction>LABEL</prediction>, where LABEL is A or B.
<evidence>
Answer:

Situation prompt.
You are a careful text classifier. Use only the supplied evidence.
Question: Which speaker has more power in this conversation?
Labels: A means Speaker A; B means Speaker B.
First state the concrete situation represented by the evidence. Then give a concise rationale connecting that situation to one final label. Return exactly:
<situation>concise situation description</situation>
<rationale>concise evidence-based rationale</rationale>
<prediction>LABEL</prediction>, where LABEL is A or B.
<evidence>
Answer:

Few-shot prompt.
Read the conversation and predict which speaker has more power. Output exactly one label: A or B. Do not explain your answer.
<text>[Retrieved example for A]</text>
<label>A</label>
<text>[Retrieved example for B]</text>
<label>B</label>
<text>[Conversation]</text>
<label>

C.4.2 Persuasion

Direct prompt.

```
You are a careful text classifier. Use only the supplied evidence.
Question: Is this post persuasive?
Labels: Y means yes, persuasive; N means no, not persuasive.
Return exactly one label (Y or N) and no explanation, using:
<prediction>LABEL</prediction>
<evidence>
Answer:
```

CoT prompt.

```
You are a careful text classifier. Analyze only the supplied evidence.
Question: Is this post persuasive?
Labels: Y means yes, persuasive; N means no, not persuasive.
First provide concise reasoning grounded in the evidence, then give one final label. Return exactly:
<reasoning>concise reasoning</reasoning>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:
```

Situation prompt.

```
You are a careful text classifier. Use only the supplied evidence.
Question: Is this post persuasive?
Labels: Y means yes, persuasive; N means no, not persuasive.
First state the concrete situation represented by the evidence. Then give a concise rationale connecting that situation to one final label. Return exactly:
<situation>concise situation description</situation>
<rationale>concise evidence-based rationale</rationale>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:
```

Few-shot prompt.

```
Read the post and predict whether the post is persuasive. Output exactly one label: Y or N. Do not explain your answer.
<text>[Retrieved example for Y]</text>
<label>Y</label>
<text>[Retrieved example for N]</text>
<label>N</label>
<text>[Post]</text>
<label>
```

C.4.3 Empathy

Direct prompt.

```
You are a careful text classifier. Use only the supplied evidence.
Question: Does the response show empathy toward the seeker?
Labels: Y means yes, empathetic; N means no, not empathetic.
Return exactly one label (Y or N) and no explanation, using:
<prediction>LABEL</prediction>
<evidence>
Answer:
```

CoT prompt.

```
You are a careful text classifier. Analyze only the supplied evidence.
Question: Does the response show empathy toward the seeker?
Labels: Y means yes, empathetic; N means no, not empathetic.
First provide concise reasoning grounded in the evidence, then give one final label. Return exactly:
<reasoning>concise reasoning</reasoning>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:
```

Situation prompt.

```
You are a careful text classifier. Use only the supplied evidence.
Question: Does the response show empathy toward the seeker?
Labels: Y means yes, empathetic; N means no, not empathetic.
First state the concrete situation represented by the evidence. Then give a concise rationale connecting that situation to one final label. Return exactly:
<situation>concise situation description</situation>
<rationale>concise evidence-based
```

rationale</rationale>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:

Few-shot prompt.
Read the seeker post and response and predict whether the response shows empathy. Output exactly one label: Y or N. Do not explain your answer.
<text>[Retrieved example for Y]</text>
<label>Y</label>
<text>[Retrieved example for N]</text>
<label>N</label>
<text>[Conversation]</text>
<label>

C.4.4 Toxicity

Direct prompt.
You are a careful text classifier. Use only the supplied evidence.
Question: Will future turns of this conversation become toxic?
Labels: Y means yes, likely toxic; N means no, unlikely toxic.
Return exactly one label (Y or N) and no explanation, using:
<prediction>LABEL</prediction>
<evidence>
Answer:

CoT prompt.
You are a careful text classifier. Analyze only the supplied evidence.
Question: Will future turns of this conversation become toxic?
Labels: Y means yes, likely toxic; N means no, unlikely toxic.
First provide concise reasoning grounded in the evidence, then give one final label. Return exactly:
<reasoning>concise reasoning</reasoning>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:

Situation prompt.
You are a careful text classifier. Use only the supplied evidence.
Question: Will future turns of this conversation become toxic?
Labels: Y means yes, likely toxic; N means no, unlikely toxic.
First state the concrete situation represented by the evidence. Then give a concise rationale connecting that situation to one final label. Return exactly:
<situation>concise situation description</situation>
<rationale>concise evidence-based rationale</rationale>
<prediction>LABEL</prediction>, where LABEL is Y or N.
<evidence>
Answer:

Few-shot prompt.
Read the conversation and predict whether future turns of the conversation will become toxic. Output exactly one label: Y or N. Do not explain your answer.
<text>[Retrieved example for Y]</text>
<label>Y</label>
<text>[Retrieved example for N]</text>
<label>N</label>
<text>[Conversation]</text>
<label>

D Likert Scales

D.1 Grounding

3- All information in the summary is directly supported by the conversation.
2- Most of the summary reflects the conversation, but there are some details that are not clearly supported or cannot be directly mapped to the conversation.
1- The summary contains multiple hallucinations or misrepresentations that could alter the meaning or intent of the original conversation.

D.2 Interpretation

3- The summary contains no misinterpretations or over-interpretations.
2- The summary mostly reflects the original conversation but includes small interpretation errors that do not significantly alter the meaning.
1- The summary contains significant interpretation errors that misrepresent the original conversation.

D.3 Completeness

3- The summary accurately captures all key points and important details from the conversation.
2- The summary includes most main ideas but omits some relevant details.
1- The summary misses important points, resulting in a lack of clarity or misrepresentation of the conversation.

E Experiment Details

E.1 LLM Details

The LLMs used for generating aspect summaries and reasoning in Section 5.2 are GPT-5 (OpenAI, 2025a), DeepSeek-R1-70B (Guo et al., 2025), gpt-oss-120b (OpenAI, 2025b), Llama-70b (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Qwen2.5-72B (Qwen et al., 2025).
The LLMs used for downstream CSS tasks prediction are GPT-4o Mini (OpenAI, 2024), Gemma3-27B (Team et al., 2025), and Qwen2.5-72B (Qwen et al., 2025).
We access GPT models through the OpenAI API, while the remaining models are accessed via an internal API provided by the authors' affiliation. All LLM calls are performed in a zero-shot setting using a single call unless stated otherwise.

E.2 LLM Aspect Summary Quality

E.2.1 Common Omission in Interaction

Content focus instead of social focus
Human: A shares an engaging summary of a TV episode. B aligns politely, notes how they watched, and probes with questions.
LLM: A shares an enthusiastic, detailed recap of a Channel4 documentary about a girl with severe verrucas caused by a T-helper cell deficiency, mentioning her upcoming bone-marrow transplant and expressing mild annoyance at the commentator's misuse of "antibodies." The tone is informal, slightly sarcastic, and appreciative of the program's educational value. B replies politely, noting they only saw it on4oD, and asks a focused question about why the girl's impaired lymphocytes didn't predispose her to common infections like colds or diarrhoea, seeking clarification on HPV-related immunity.
Omission of interactive roles
Human: B respectfully concedes expertise to A, aligning with Wikipedia norms and encouraging A to take the lead.
LLM: The interaction involves B acknowledging A's expertise on the topic, encouraging A to take bold actions like rewriting or renaming the article. B shows deference by conceding their own lack of knowledge and offering support for decisions such as nominating the article for deletion, while committing to abstain from voting. The exchange reflects collaborative alignment, mutual respect, and a willingness to empower A's judgment, fostering a cooperative relationship through clear communication and polite concession.

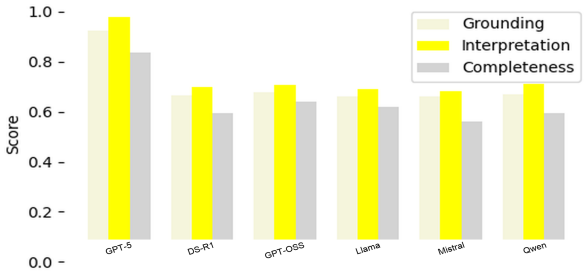


Figure 8: Reason quality of various LLMs, using human annotated reasons as comparison, evaluated using evaluation dimensions.

E.3 LLM Reason Quality

Figure 8 presents the quality scores of LLM-generated reasoning compared to gold-standard human rationales.

Table 6 shows the evaluation scores are higher for correctly predicted instances than incorrectly predicted ones among the annotated instances.

F Downstream CSS Task

For each task, we sample 1,000 instances with balanced labels and no overlap with the human annotated examples. To ensure paired comparisons, the main analysis is conducted on the intersection of 3,988 instances for which every model and method produces a valid prediction. Some models fail to return valid predictions for certain instances due to token limitations, as we set the maximum number of generated tokens to 1 for direct prediction prompts and 256 for prompts containing reasoning components, with the temperature fixed at 0.

We explicitly use GPT-4o Mini rather than newer models because it does not include a reasoning feature and can produce a direct one-token prediction without consuming hidden tokens for internal reasoning.

Task	Grounding		Interpretation		Completeness	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Power	0.759	0.758	0.812	0.814	0.771	0.774
Persuasion	0.828	0.820	0.830	0.812	0.723	0.721
Empathy	0.839	0.815	0.868	0.820	0.775	0.770
Toxicity	0.863	0.865	0.922	0.916	0.845	0.848

Table 6: Evaluation scores of correct vs incorrect prediction of CSS task instances in annotated dataset.

F.1 Task-Level Prediction Results

Table 7 reports task-level accuracy for the common set of valid predictions. Each summary-bearing condition should be compared with the corresponding text-only condition within the same reasoning family.

F.2 BM25 Retrieval Results

We also evaluate unconstrained top-k BM25 retrieval, in which the selected examples are not required to cover every label and may therefore be dominated by the locally most similar label. We conduct this experiment for Persuasion, Toxicity, and Empathy, whose label spaces are Yes/No. We exclude Power because its labels correspond to Speaker A or Speaker B, making retrieval particularly sensitive to speaker roles and potential label flips. Table 8 reports the unweighted mean performance across these three tasks.

F.3 Reasoning Theme Analysis

F.3.1 Theme Discovery

We analyze 47,991 non-empty rationales generated by three models across four tasks and four reasoning conditions: chain-of-thought (CoT) and situation prompts, each with either text-only or text+summary inputs. Themes are discovered separately for each task using rationales from all models and methods, ensuring that theme identities are shared across comparisons within each task. We embed the generated rationales using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and cluster them with UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017). Llama 4 Maverick (Meta AI, 2025) is used for theme naming and for assigning rationales to themes. One rationale can have multiple themes assigned. Model identity, prompt condition, predictions, and correctness are hidden during parameter selection and theme naming to reduce potential bias. For each task, we additionally construct a separate reference taxonomy from 1,000 GPT-5 rationales generated while conditioning on the gold outcome. Generated rationales are then assigned to the reference themes using the same procedure.

For each matched text-only/text+summary pair, we measure: (1) the cosine similarity between the generated rationale and its paired GPT-5 rationale; (2) whether their assigned theme sets share at least one reference theme; and (3) the Jaccard overlap between their complete reference-theme sets. Results using the primary combined assignment are reported in Table 9.

These quantities characterize alignment with GPT-5's gold-conditioned explanations, not independently human-validated correctness. Theme labels are semantic summaries of clusters, and multi-label assignments are calibrated computational annotations rather than human gold labels. Accordingly, model-level theme frequencies and theme–correctness associations are diagnostic descriptions and are not used to make causal claims about why a prediction changes.

F.4 Power Prediction Example

Figure 9 summarizes these error categories and provides representative example instances.

Figure 11 shows an example using aspect summary and situated reasoning to make prediction.

We observed that LLM performance improves more substantially when the model captures the situational context underlying the textual exchange. For instance, in the task of power prediction, recognizing that a conversation occurs in an online Wikipedia discussion provides a critical clue for reasoning about the speakers' roles. By combining this situational information with behavioral cues extracted from the summaries, the model is better able to make the correct power prediction. We therefore hypothesize that while LLMs can effectively capture behaviors and interaction dynamics, they must situate these behaviors within the appropriate context to make socially coherent judgments. Figure 10 shows the qualitative analysis, illustrating how the same aspect-based behaviors, when placed in different social contexts, can lead to distinct social reasoning and decisions.

Task	Reasoning family	GPT-4o Mini		Gemma3-27B		Qwen2.5-72B	
		Text	+Summ	Text	+Summ	Text	+Summ
Power	Direct	64.60	64.70	61.80	65.60	64.80	64.60
	CoT	62.80	63.00	63.00	65.40	64.50	63.30
	Situation	59.90	61.90	63.90	64.40	65.50	65.10
Persuasion	Direct	56.77	55.47	55.57	53.66	55.57	55.37
	CoT	52.56	50.25	53.56	57.37	57.57	53.56
	Situation	59.78	55.47	51.86	53.36	56.57	54.66
Empathy	Direct	75.08	81.28	81.58	81.08	70.07	80.48
	CoT	71.07	80.28	77.98	80.68	67.07	78.58
	Situation	68.57	79.08	77.58	80.18	66.57	79.88
Toxicity	Direct	52.02	56.15	60.69	62.30	55.85	55.95
	CoT	61.09	61.79	59.38	63.00	55.14	56.96
	Situation	59.38	60.69	59.68	63.41	53.83	56.55

Table 7: Task-level accuracy.

k	Text retrieval		Summary retrieval		Δ	
	Acc.	F1	Acc.	F1	Acc.	F1
1	52.97	51.57	59.50	59.30	+6.53	+7.73
3	53.23	50.90	60.67	60.33	+7.43	+9.43
5	53.07	49.93	60.33	59.90	+7.27	+9.97

Table 8: Unconstrained BM25 few-shot retrieval results for persuasion, toxicity, and empathy tasks.

Metric	Text only	+Summaries	Δ
Cosine similarity	.571	.573	+.002
Theme overlap	.425	.438	+.013
Theme-set Jaccard	.223	.221	-.002

Table 9: Overall alignment with the GPT-5 reference reasoning.

Task	Cosine Δ	Overlap Δ	Jaccard Δ
Power	+.005	+.021	-.004
Persuasion	+.017	+.014	+.004
Empathy	-.007	-.009	-.011
Toxicity	-.008	+.026	+.004
Overall	+.002	+.013	-.002

Table 10: Change from text-only to text-summary rationales relative to the GPT-5 reasoning reference.

Action -> power
A: I think you should close the first one too, as I gave it as an example of a bad FA a few times ”’
B: Will do. When I closed Dorset it wasn't exactly 14 days for those. I'll get to them tomorrow morning (it's getting to be bedtime here!) Have you had a chance to check out the two I pinged you about?
Reasoning and Decision:
Speaker B exhibits power by suggesting a delayed action in closing the cases and mentioning their personal schedule.

Imperatives tone -> power
A: Hello Dana, thanks for contribute in this ..., but i need tell you that..., if you think that im not right please tell me for not fixing the edit at last. Enjoy your day ;)–
B: My apologies, I forgot to include the disambiguator. Please check now - it should be right. Thanks for letting me know.
Reasoning and Decision:
Speaker A establishes authority by correcting Speaker B's contribution, setting the expectation for compliance.

Figure 9: examples

Aspect Summary:
Speaker A is formal and technical. Speaker B is humorous and casual.
Situation 1:
In a technical meeting.
Reasoning and Decision:
A's knowledge suggests expertise and familiarity with the procedure. So **A** holds more power.
Situation 2:
In a collaborative discussion.
Reasoning and Decision:
B's style allows others to speak freely without hierarchical concerns. So **B** holds more power.

Aspect Summary:
Speaker expresses reciprocation when asking for a favor on a social platform.
Situation 1:
By sending pictures of puppies.
Reasoning and Decision:
People will feel the warmth and emotionally connected. **Successful** persuasion for favor.
Situation 2:
By sending pictures of themselves.
Reasoning and Decision:
People may find this oversharing and awkward. **Unsuccessful** persuasion for favor.
Situation 3:
By sending music albums.
Reasoning and Decision:
People may not resonate if they do not like music. **Unsuccessful** persuasion for favor.

Figure 10: Examples illustrating how similar aspect-based behaviors can lead to different social decisions depending on the social context.

Input Conversation:
A: I think you should close the first one too, as I gave it as an example of a bad FA a few times
B: Will do. When I closed Dorset it wasn't exactly 14 days for those. I'll get to them tomorrow morning (it's getting to be bedtime here!) Have you had a chance to check out the two I pinged you about?
Aspect Summaries:
Style: In the conversation, Speaker A employs a formal tone with concise vocabulary, reflecting a professional demeanor. Speaker B uses a casual style, incorporating informal expressions like "Will do" and "it's getting to be bedtime here," indicating a friendly and approachable identity. The sentence structures differ, with A favoring straightforward commands and B blending simple and conversational syntax, showcasing a more relaxed attitude toward the discussion.
Content: In the conversation, Speaker A advises Speaker B to close a specific project considered a poor example. Speaker B acknowledges the suggestion and mentions that the closure wasn't exactly 14 days for another project. They express their intention to address it the following morning due to the late hour. Additionally, Speaker B inquires if Speaker A has reviewed two projects they previously discussed.
Interaction: In the conversation, both speakers demonstrate good engagement. Speaker A provides a suggestion and offers context, while Speaker B acknowledges the advice, communicates their timeline, and promptly asks about additional topics they previously discussed. This exchange reflects active listening and a willingness to collaborate, enhancing their overall interaction.
Reason:
Supporting A: Based on Style, A "employes a formal tone" to "advise" B, showing A's influence on B; based on Content and Interaction, A is describing actions they took, demonstrating their leadership position in the collaboration.
Supporting B: Based on Style, B uses casual expression like "will do" and "it's getting to be bedtime here"; based on Content and Style, B acknowledges A's suggestion, but communicates timeline ("will get to it tomorrow") indicating that B is the one to make decision and guide the collaboration forward.
Scenario:
A and B are working on editing and reviewing content. A drives the collaboration and make suggestions for B. B is asking A to review content and indicates they are waiting for a decision from A about it.
Overall Judgement:
A

Figure 11: Power prediction example using aspect summaries and situated reasoning.