Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions

Maryam Amirizaniani
University of Washington
Seattle, USA
amaryam@uw.edu

Elias Martin
University of Washington - Bothell
Bothell, USA
eamart34@uw.edu

Maryna Sivachenko
University of Washington - Bothell
Bothell, USA
msiva@uw.edu

Afra Mashhadi
University of Washington - Bothell
Bothell, USA
mashhadi@uw.edu

Chirag Shah
University of Washington
Seattle, USA
chirags@uw.edu

ABSTRACT

Theory of mind (ToM) reasoning involves understanding that others have intentions, emotions, and thoughts, which is crucial for regulating one's reasoning. Although large language models (LLMs) excel in tasks such as summarization, question answering, and translation, they still face challenges with ToM reasoning, especially in open-ended questions. Despite advancements, the extent to which LLMs truly understand ToM reasoning and how closely it aligns with human ToM reasoning remains inadequately explored in open-ended scenarios. Motivated by this gap, we assess the abilities of LLMs to perceive and integrate human intentions and emotions into their ToM reasoning processes within open-ended questions. Our study utilizes posts from Reddit's ChangeMyView platform, which demands nuanced social reasoning to craft persuasive responses. Our analysis, comparing semantic similarity and lexical overlap metrics between responses generated by humans and LLMs, reveals clear disparities in ToM reasoning capabilities in open-ended questions, with even the most advanced models showing notable limitations. To enhance LLM capabilities, we implement a prompt tuning method that incorporates human intentions and emotions, resulting in improvements in ToM reasoning performance. However, despite these improvements, the enhancement still falls short of fully achieving human-like reasoning. This research highlights the deficiencies in LLMs' social reasoning and demonstrates how integrating human intentions and emotions can boost their effectiveness.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Reasoning in Large Language Models; Theory of Mind

ACM Reference Format:
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), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679832

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679832

1 INTRODUCTION

In the realm of artificial intelligence, the ability of machines to understand and respond to human-like social cues—like theory of mind (ToM) reasoning—remains a pivotal challenge. ToM entails the capacity to attribute mental states, such as intention, emotion, and belief, to oneself and others, and to understand that these states can be different from one's own [21]. This cognitive ability is fundamental in social human interaction and crucial for effective communication. Large language models (LLMs), which have achieved impressive success in various natural language processing tasks [12, 44, 60], are now being pushed to the frontier of social reasoning to see if they can mimic this quintessentially human trait, especially in interacting with a human.

Despite LLMs' prowess in handling structured tasks such as summarizing [3, 17, 48], question answering [1, 47], and language translation [15, 29], they have shown limitations when tasked with zero-shot ToM reasoning that requires nuanced understanding and integration of human mental state [30, 45, 46, 65]. Many studies have shown this limitation of LLMs via utilizing multiple choice and short answer questions [14, 18, 57], but not on open-ended questions. We aim to bridge this gap by rigorously evaluating the ability of LLMs to engage in zero-shot ToM reasoning tasks within open-ended scenarios, assessing how closely their performance aligns with human capabilities in ToM reasoning tasks. In particular, we aim to answer the following research questions:

- **RQ1**: To what degree are LLMs capable of zero-shot reasoning in open-ended questions?
- **RQ2**: To what extent are human and LLM social reasoning capabilities aligned in addressing open-ended questions?
- **RQ3**: How does considering human mental state affect the performance of LLMs in ToM reasoning of open-ended questions?

To address the posed research questions, we utilized data from Reddit's ChangeMyView subreddit—a platform noted for its intense social interactions and open-ended questions requiring robust reasoning to shift the viewpoints of posters. Our study examines the capability of LLMs to provide reasoning and respond effectively to

these queries, specifically focusing on how integrating individuals' mental states can enhance this ability.

Through comparative analyses assessing semantic similarity and lexical overlap scores between human and LLM responses, we observed significant disparities in reasoning capabilities within open-ended scenarios, similar to those in non-open-ended questions. These findings reveal considerable limitations in even the most advanced LLMs as mentioned in [50] and Huggingface leader board,[1] such as Zephyr-7B [50], Llama2-Chat-13B [49], and GPT-4 [2]. Our research underscores the effectiveness of incorporating mental states such as human intentions and emotions into LLM reasoning via prompt tuning. The results emphasize the need for LLMs to understand human-like mental states to effectively engage in social reasoning in open-ended questions, indicating a key area for development.

The **motivation** stems from assessing how well LLMs generate responses to open-ended questions and how their reasoning aligns with human reasoning. This analysis is crucial for understanding LLMs' limitations in replicating human ToM reasoning. The **novelty** of this research lies in its focus on comparing LLM reasoning with human reasoning in open-ended questions, using Reddit posts as a rich data source. Our **contributions** to this field are outlined as follows:

- We provide a comparative analysis of responses from humans and LLMs, emphasizing the disparities in ToM reasoning capabilities within open-ended scenarios.
- We rigorously assess LLMs' reasoning capabilities to integrate human intentions and emotions within open-ended questions, utilizing data from Reddit's ChangeMyView.
- We utilize a prompt tuning method that substantially improves the ToM reasoning performance of LLMs in open-ended questions by incorporating human intentions and emotions.

The rest of this paper is organized as follows: Section §2 reviews related work in the field. Section §3 details the experimental setup and methodology employed in this study. Section §4 is devoted to discussing the analysis and results of the experiment. Section §5 presents a discussion of the findings relative to the experiment. Finally, Sections §6 and §7 concludes the paper by summarizing the key findings and explores the limitations of this study, highlighting areas for future research.

2 RELATED WORK

2.1 Theory of Mind (ToM) Reasoning

ToM in reasoning refers to the ability to understand and attribute mental states, such as beliefs and intentions, to oneself and others to predict and interpret behavior in social interactions [21]. In exploring the developmental trajectory ToM abilities, van Duijn et al. [52] examined the performance of 11 state-of-the-art language models alongside children aged 7-10 on advanced ToM tests. Notably, instruction-tuned LLMs from the GPT family demonstrated superior performance, occasionally surpassing that of children. However, basic LLMs faced notable challenges in solving ToM tasks [52]. In this context, researchers have utilized a variety of cognitive science tests to probe the emergence of ToM reasoning within LLMs [8, 33, 36]. These studies aim to determine how well LLMs can emulate the human ability to understand.

Findings from Sap et al. [40] using the ToMi dataset [20] indicate that GPT-3, exhibits ToM capabilities that are significantly inferior to those of humans. This gap underscores a critical limitation in current LLMs with some traditional ToM tasks, they continue to struggle with reliably showing ToM capabilities [51, 59]. These studies suggest that the consistency with which LLMs demonstrate ToM abilities remains questionable, often defaulting to surface-level reasoning strategies rather than engaging in deep, robust ToM reasoning [43]. Moreover, Kim et al. [14] introduced FANToM benchmark to rigorously assess ToM abilities within conversational contexts. This benchmark has revealed significant challenges facing state-of-the-art LLMs, such as GPT-4, Llama 2, Falcon, and Mistral, particularly in maintaining performance in ToM reasoning tasks in comparison to humans, even with chain-of-thought reasoning or fine-tuning.

Wu et al. [57] highlights a significant decline in performance across various LLMs, such as GPT 4, GPT 3.5, Claude, and Guanaco, when tasked with higher-order ToM challenges. This trend suggests that as the complexity of ToM tasks increases, the ability of these models to accurately interpret and respond to nuanced mental states diminishes notably. On the other hand, these findings contradict the claims by Kosinski [18] and reiterated by Bubeck et al. [5], which posited that modern LLMs reached high ToM scores.

2.2 Mental States in ToM

Recent research in the application of ToM within LLMs has provided notable insights into the effects of understanding mental states on LLM reasoning. For instance, the Foresee and Reflect (FaR) framework offers a reasoning structure that encourages LLMs to anticipate future challenges and reason about potential actions. Analysis reveals the effectiveness of incorporating mental states into reasoning [64]. Additionally, Ma et al. [30] develops a comprehensive taxonomy for ToM in LLMs, known as Abilities in Theory of Mind Space (ATOMS), which categorizes crucial components such as Intentions, Percepts, Beliefs, Emotions, Knowledge, Desires, and Non-literal Communication. This framework aims to provide a structured approach to assess and systematically enhance ToM capabilities. These developments highlight the effectiveness of recent model iterations in approximating human mental states.

Concurrently, the BigToM benchmark has been developed to specifically assess LLMs' social reasoning capabilities, focusing on aspects such as beliefs, percepts, desires, and user actions [9]. Another tool that has emerged mental state in ToM is SymbolicToM, which enhances ToM capabilities in reading comprehension tasks by effectively representing entities' beliefs and facilitating higher-order reasoning. This approach has shown promise in providing a deeper understanding of belief states and their implications for ToM [41].

Moreover, different methods are utilized in ToM reasoning understanding. For instance, Li et al. [22] focused on ToM into dialogue models through reinforcement learning and demonstrated significant improvements in the quality of guidance these systems provide. This integration underscores the importance of accurately

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

interpreting and responding to intents within conversational contexts [22]. Additionally, Lim et al. [23] explores the integration of the "Bayesian Theory of Mind" with optimal-planning agents. This approach illustrates how explicitly representing others' intentions can enhance performance in ToM reasoning.

In contrast to earlier research, which typically relied on multiple-choice formats or short-answer questions to evaluate the ToM capabilities of LLMs [57], our study adopts a more nuanced approach by utilizing open-ended questions. This method allows for a broader range of responses, providing deeper insights into how LLMs interpret and respond to complex scenarios. By shifting from structured to more exploratory questioning, we aim to uncover subtler aspects of ToM reasoning in LLMs.

3 EXPERIMENTAL SETUP

3.1 Large Language Models

For the work reported here, we utilize three of the most popular LLMs known for their exceptional reasoning capabilities, as documented in [50]. Specifically, we have selected Zephyr-7B [50], Llama2-Chat-13B [49], and GPT-4 [2], setting the temperature parameter to 0.5.

3.2 Dataset

For our experiments, we leverage the Reddit r/ChangeMyView (CMV) community[2] as a source for prompts, capitalizing on its reputation for open-ended discussions. This platform is ideal as it enables questioners to share opinions and invite others to change these views with reasoning grounded in the questioners' beliefs, thoughts, intent, and other mental states, thus providing a rich dataset for this study.

The data for this study is comprised of three parts. The first part consists of Reddit posts, which were crawled from the aforementioned Reddit community. We initially collected the 1,000 most recent posts using the PRAW Reddit API,[3] with the oldest one being published in February 2024. It is worth noting that these posts could not have been incorporated in the training corpora of the LLMs examined in this study, as their most recent training data does not encompass data from February 2024 and more recent. The other two parts include human-written and LLM-generated responses to those posts, which will be discussed in more detail subsequently.

For the human-written responses, we extracted the top five responses based on community upvotes, utilizing Reddit's internal voting mechanism to rank comments. On Reddit, the 'Vote Number' of a post or comment represents its net score, calculated by subtracting the number of downvotes from the number of upvotes. To ensure reliability in our analysis, we selected only those responses with entirely positive scores. By selecting the top five responses, we ensure a diverse and well-supported dataset of argumentative responses, providing a comprehensive view of human reasoning. After filtering out data with fewer than five responses and further cleaning and preprocessing to ensure all columns were non-null and formatted as strings, we retained 845 distinct Reddit posts, each with five Reddit user responses.

[2]https://www.reddit.com/r/changemyview/
[3]https://www.reddit.com/r/PRAW/

Simultaneously, to create a dataset of LLM-generated responses, we input the same 845 posts into the above-mentioned LLMs using a specially crafted prompt template and requested it to generate five responses for each.

To develop this prompt template, we conducted several iterations. Initially, we utilized a generic prompt template, "`Generate five reasoning answers for 'Question'.`", which failed to produce satisfactory results. The ineffectiveness of this prompt template is attributed to its lack of specificity and inherent ambiguity, leading LLMs to generate generic and irrelevant responses rather than precise and relevant ones.

To address this challenge, we were inspired by the Chain of Thought (CoT) approach [56], which promotes a step-by-step reasoning process in LLMs, resulting in higher-quality answers [14, 27, 37, 54]. Additionally, following [16, 32], we enhanced our approach by clearly defining the tasks and setting explicit expectations within the prompts. This adjustment has been shown to significantly improve the quality of responses as supported by Ye et al. [58] and Zeng et al. [61], thus ensuring that our prompts are well-crafted to elicit detailed and contextually appropriate answers from the models. These comprehensive enhancements led to the development of the final prompt template, which not only elicited straightforward answers but encouraged reasoning that challenges and potentially shifts questioner viewpoints, as shown in Figure 1. This approach mirrors the nature of the conversation on the Reddit CMV platform.

Prompt Template

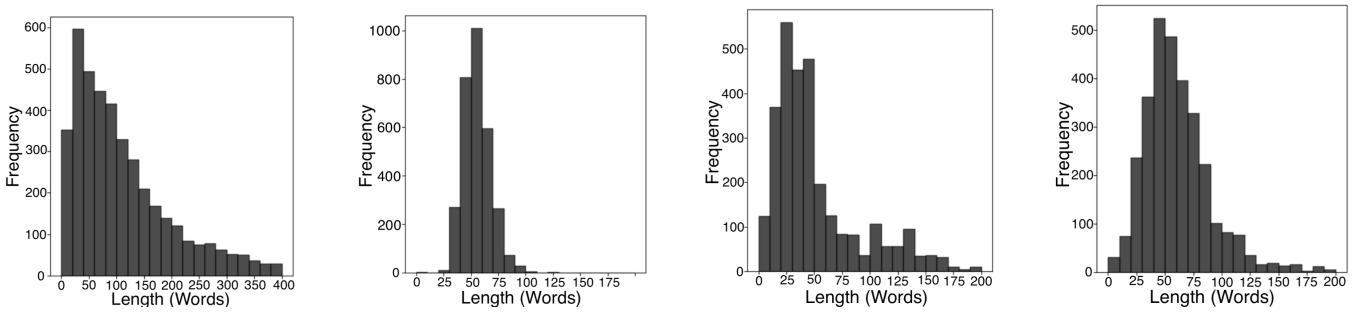
If you were a user on Reddit exploring the 'Change My View' subreddit, a platform dedicated to facilitating reasoned responses aimed at changing the questioner's perspective, imagine encountering a post titled "**{title}**" with the following content: "**{Body}**". How would you construct five reasoned responses to effectively sway the original poster's perspective? Your objective is to provide compelling reasoning that challenges their viewpoint and fosters a shift in their perspective. Approach this task step by step.

Figure 1: Initial prompt template for generating reasoning answers.

3.3 Data Description

Drawing on the dataset mentioned above, Figure 2 illustrates the distribution of average lengths for comments generated by both humans and LLMs. This visualization effectively highlights the variability in response lengths from each source, with each figure presenting a distribution that approximates a normal curve. Notably, human responses range up to 400 words but typically they are around 50 words. Also, responses from GPT-4.0 tend to cluster around an average of 50 words, demonstrating a narrower range in length variability. Conversely, Llama2-Chat-13B and Zephyr-7B display a broader spectrum in the length of their responses, suggesting a richer diversity in response detail and complexity.

Despite the numerical differences in length, the overall patterns observed across the distributions underscore an underlying consistency in how different models and humans generate response lengths.



(a) Human comments **(b) GPT-4 comments** **(c) Llama2-Chat-13B comments** **(d) Zephyr-7B comments**

Figure 2: Overview of length distribution of reasoning answers generated by humans and LLMs.

4 ANALYSIS AND RESULT

4.1 Human-based Evaluation

We utilize the Human-in-the-Loop (HIL) method to evaluate the reasoning capabilities of LLMs. To ensure a comprehensive assessment, trained assessors, who are graduate students in computer science, were asked to annotate the responses generated by these models.

The assessment process is divided into two rounds to comprehensively analyze the reasoning abilities of LLMs. In the initial round, we evaluate the LLMs' reasoning performance using the previously mentioned prompt template, aiming to gauge each model's basic reasoning capabilities. The second round intensifies the scrutiny by incorporating details of mental states, such as questioners' intentions, emotions, and sentiments embedded within the questions. The LLMs are tasked with generating reasoning responses that consider these aspects of mental states. Our annotators then evaluate these responses to determine the quality of reasoning, focusing on how the integration of mental state details influences the reasoning process.

In this phase, we randomly selected 100 posts from our dataset. Each post generated five reasoning answers, resulting in a total of 500 answers per LLM. With three LLMs under evaluation, this led to an aggregate of 1,500 reasoning answers for human assessment. To facilitate the labeling process, we recruited and trained ten assessors, ensuring they were well-versed in the goal and the task, while not having any conflict that could jeopardize this work. Each reasoning answer produced by the LLMs was independently evaluated twice by different assessors. The same set of questions was employed for labeling in both the first and second rounds. However, we ensured that assessors did not encounter the same questions in the second round to mitigate any familiarity bias.

For assessment, it is essential to establish specific criteria for assessors to evaluate the reasoning in responses across these two rounds of assessment. Drawing from previous research on the 'Reasoning' task, we focus on the criterion of "Reasoning Correctness", which mandates that the reasoning in responses should be valid, supporting the conclusions or explanations logically and accurately [26, 35]. The criteria for 'Reasoning Correctness' are detailed as follows:

- **Reasoning Correctness**: Does the reasoning in the response adequately address the question? Is the reasoning relevant to the question asked? Does the reasoning accurately represent the information provided and the conclusions drawn from it? Is the reasoning logically sound, without any fallacies or errors in logic? Does the reasoning lead to a clear and well-supported conclusion?

To ensure consistency in our evaluation process, we calculate the Agreement Rate among assessors to quantify the level of consensus on evaluation results. We adopt established statistical methods to determine inter-assessor agreement, primarily using Cohen's kappa [4], considered acceptable within the range of 0.61 to 0.80 [4]. Additionally, we measure the overlap rate among the responses provided by assessors, with an acceptable benchmark set at 80%. These methods provide a rigorous and quantifiable measure of inter-assessor agreement, thereby enhancing the reliability of our evaluations. The subsequent sections will detail the methodologies employed in the two rounds of assessments and discuss the results.

4.1.1 ***Human annotation - Round 1.*** This section outlines the first round of human annotation used to assess the reasoning capabilities of LLMs. We initially implemented a 5-point Likert scale to evaluate responses based on the "Reasoning Correctness" criteria but encountered significant challenges. The Cohen Kappa scores, as shown in Table 1, indicated that this method did not fulfill the acceptance criteria for consistent inter-annotator agreement. As a result, we transitioned to a 3-point Likert scale, which yielded a Cohen's kappa score of 0.4751 and an overlap rate of 61.73%. Although these scores represent an improvement, they still fail to achieve satisfactory agreement. This may have been due to the diversity of reasoning approaches among the annotators. In response, further iterations of our study introduced a dichotomous scale, which enhanced Cohen's kappa score to 0.6618 and the overlap rate to 83.66%, thereby meeting the agreement standards. This adjustment effectively reduced subjectivity among assessors in evaluating the reasoning answers.

Our findings reveal variation in the reasoning quality of responses to open-ended questions: 23.52% from Zephyr-7B, 28.19% from Llama2-Chat-13B, and 35.69% from GPT-4 were rated as demonstrating adequate reasoning, according to the consensus among assessors. These results are detailed in Table 2. This analysis indicates that from a human perspective, LLMs struggle to generate reasonable and contextually appropriate responses to open-ended questions.

Table 1: Agreement rate among annotators for "Reasoning Correctness" in the first and second rounds.

Scale	Cohen's kappa score	Overlap rate
Round 1		
5-point Likert scale	0.3527	45.10%
3-point Likert scale	0.4751	61.73%
Dichotomous scale	0.6618	83.66%
Round 2		
Dichotomous scale	0.7204	86.24%

Table 2: Assessor 'Yes' responses to the 'Reasoning Correctness' criterion across the first and second rounds, and delta change of these two rounds.

Metric	Reasoning Quality = "Yes"
Round 1	
Zephyr-7B	23.52%
Llama2-Chat-13B	28.19%
GPT-4	35.69%
Round 2	
Zephyr-7B	29.43%
Llama2-Chat-13B	35.09%
GPT-4	43.21%
Delta improvement	
Zephyr-7B	↑ 5.91%
Llama2-Chat-13B	↑ 6.9%
GPT-4	↑ 7.52%

4.1.2 ***Human annotation - Round 2.*** In this evaluation round, which includes human annotation, we investigate how detailed mental states—such as emotions, intentions, and question sentiment—affect the reasoning performance of LLMs. Following the findings of Kim et al. [14], which highlight the effectiveness of prompt tuning for integrating mental states over other methods, we have incorporated these elements into our analysis by embedding them into prompts. This prompt-tuning approach subsequently leads LLMs to generate reasoning responses based on these enriched prompts. The process of this integration is depicted in Figure 3. We initiate this phase by extracting the mental state information from the dataset. Subsequent sections will provide a detailed exposition of the methodologies utilized for the extraction of mental states.

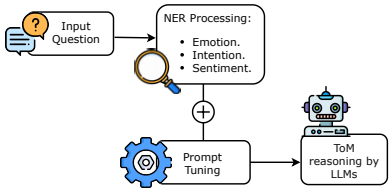


Figure 3: ToM reasoning process via prompt tuning.

- **Posts sentiment and emotion:** To measure the sentiment and emotions of the posts, we utilized a RoBERTa [28] base model that had been fine-tuned on the Go Emotions dataset [6], which is derived from Reddit posts. This choice was ideal for our study since Go Emotions can predict and rank a range of up to 28 distinct emotions based on the likelihood of their association with a given text input. For our analysis, we concentrated on the highest-ranked emotion identified in both the titles and bodies of the posts. Furthermore, to precisely capture and analyze the spectrum of emotions conveyed by questioners in their inquiries, we utilized Named Entity Recognition (NER) techniques. The NER is an information extraction technique that identifies and categorizes named entities in text into predefined categories. Here, we employed NER to specifically identify and classify emotional expressions within the text, enhancing our understanding of the distinct emotional states expressed by questioners. Additionally, we categorized these 28 emotions into three sentiment groups—positive, negative, or neutral—based on which sentiment classification aligned most closely with each emotion, following the methodologies outlined in [34]. To better understand the emotions and sentiments in the posts and responses generated by both humans and LLMs, we conducted a series of analyses, with results presented in Table 3. We used a pairwise t-test to assess the statistical significance of differences in sentiments and emotions, considering p-values of 0.05 or lower as significant. Notably, human responses showed a broader emotional range compared to LLMs, which tended to express concentrated emotions, primarily admiration, approval, neutrality, and curiosity. It's also important to highlight that the Zephyr-7B LLM exhibited a notably limited emotional range, with significant findings in only one emotion, suggesting it is less emotionally developed than more advanced models like GPT-4 and Llama2-Chat-13B.
- **Posts intention:** To measure the intentions behind the posts, we utilized a DistilBERT model [39] fine-tuned on the Clinc OOS dataset [19]. This dataset includes 150 intent classifications designed for task-oriented dialogue systems, sourced from a variety of platforms including Quora and Wikipedia. Similar to how we extracted emotions, we also employed NER techniques to identify intentions behind the questions. When the NER identifies intents as 'OOS' (Out of Scope),

Table 3: Emotion sentiment significance results for top emotions in reasoning answers generated by Human, GPT-4, Llama2-Chat-13B, and Zephyr-7B. Statistically significant values are shown in bold.

Emotion	Positive	Negative	Neutral
Human			
Admiration	**+6.42e-08**	2.53e-01	5.26e-02
Caring	2.58e-01	**+2.44e-04**	2.01e-01
Approval	**+1.56e-02**	9.54e-01	2.52e-01
Neutral	**-2.51e-02**	3.43e-01	1.33e-01
Desire	8.72e-01	**+1.43e-03**	1.36e-01
Disappointment	6.01e-01	**+1.82e-02**	1.61e-01
Anger	**+2.28e-05**	5.42e-01	9.73e-02
Sadness	7.63e-01	**+1.87e-02**	1.95e-01
GPT-4			
Admiration	**+2.13e-04**	1.26e-01	3.22e-01
Caring	4.01e-01	**+2.20e-02**	1.47e-01
Approval	**+4.52e-02**	**-2.73e-02**	9.16e-01
Neutral	**-7.67e-03**	4.64e-01	3.77e-01
Curiosity	6.02e-02	**+4.56e-02**	7.34e-02
Llama2-Chat-13B			
Joy	**+3.78e-03**	6.94e-01	2.52e-01
Admiration	**+4.60e-14**	1.10e-01	**-6.46e-03**
Approval	**+3.31e-04**	1.06e-01	3.66e-01
Neutral	**-5.77e-04**	9.57e-01	1.21e-01
Curiosity	**+1.16e-05**	9.55e-01	**-4.19e-02**
Annoyance	2.48e-01	**+1.11e-16**	**-9.97e-04**
Zephyr-7B			
Joy	4.11e-01	**+7.41e-03**	1.05e-01

Prompt Template

If you were a user on Reddit exploring the 'Change My View' subreddit, a platform dedicated to facilitating reasoned responses aimed at changing the questioner's perspective, imagine encountering a post titled "**{title}**" with the following content: "**{Body}**". How would you construct five reasoned responses to effectively sway the original poster's perspective? Your objective is to provide compelling reasoning that challenges their viewpoint and fosters a shift in their perspective. Approach this task step by step.

Consider the user's intent "**{user's intention}**" and emotion. "**{user's emotion}**", along with the sentiment of the post "**{post sentiment}**", as you construct your reasoning. Think step by step.

Figure 4: Final prompt template.

indicating that they do not align with predefined categories, we particularly focus on extracting tokens that carry the highest 'OOS' weight. This method enhances our accuracy in identifying intents that fall outside the standard classifications.

After extracting sentiment, emotion, and intention from questions, we refined the prompt by incorporating this contextual data. According to [25], such contextual information enhances output quality. The updated prompt template is illustrated in Figure 4. This comprehensive prompt allows for a rich exploration of how these mental states affect model performance in reasoning. We then asked the LLMs to regenerate reasoning responses based on the updated prompt and the given mental states. Annotators were tasked with evaluating the "Reasoning Correctness" using the previously defined criteria and scale. In this round, we achieved a Cohen's kappa score of 0.7204 and an overlap rate of 86.243% for the Dichotomous scale which met the acceptance rate. The results showed an improvement in reasoning quality: 29.43% for Zephyr-7B, 35.09% for Llama2-Chat-13B, and 43.21% for GPT-4. In the comparative analysis between the first and second rounds, the delta improvements observed were as follows: Zephyr-7B exhibited an increase of 5.91%, Llama2-Chat-13B showed a gain of 6.9%, and GPT-4 demonstrated an enhancement of 7.52%. These findings are presented in Table 2.

In summary, this evaluation phase involved human annotators assessing the impact of incorporating mental states—like sentiment, emotion, and intent—into LLM reasoning for open-ended questions. Drawing on the results presented in Table 2 and the observed delta improvement, we conclude that while incorporating mental states into reasoning generation enhances the performance of LLMs in producing responses to open-ended questions, these models still fall short of generating high-quality reasoning responses.

4.2 Metric-based Evaluation

Human evaluation shows that incorporating mental states improves LLM reasoning in open-ended questions, but the quality remains inadequate. In this section, we aim to quantitatively assess how closely LLM-generated responses mimic the reasoning patterns typically found in human responses. To do this, we conduct a detailed analysis by comparing LLM outputs to human reasoning, using advanced metrics for measuring semantic similarity and lexical overlap, following previous reasoning studies like [31, 38, 59].

In our analysis, we use human-written responses as benchmarks to evaluate the outputs of LLMs. For each post, we compare five top-scored human-written answers with five answers generated by the LLMs. We employ evaluation metrics to construct a 5x5 score matrix for each comparison, where each entry represents the comparison between one human answer and one LLM answer. We then apply a max-mapping strategy, matching each LLM-generated answer with the human-written answer that yields the highest score. The average of these maximum scores is calculated to obtain the final scores, providing a detailed quantitative analysis of how closely LLM outputs align with human reasoning. The evaluation metrics used include:

1. ROUGE-L [24]: ROUGE-L is a conventional metric that assesses F1 N-gram overlaps between a candidate sequence and, ideally, several reference sequences; **2. BLEURT** [42]: BLEURT is a metric developed using BERT to detect nuanced semantic similarities between sentences; **3. BERTScore** [62]: The BERTScore utilizes pre-trained BERT embeddings to calculate the semantic similarity between the generated responses and the reference. Here we use "all-mpnet-base-v2" and cosine similarity to evaluate semantic similarity in responses; **4. MoverScore** [63]: MoverScore is a reference-based evaluation metric that utilizes Earth Mover's Distance to compare a candidate sentence with its reference, using contextual word representations.

The results of our analysis, summarized in Table 4, establish a baseline for this research. Here, GPT-4 emerges as the standout model, consistently outperforming the others across all metrics. It scored 0.279 on ROUGE, 0.411 on BLEURT, 0.593 on BERTScore, and 0.310 on MoverScore, confirming its ability to generate responses that are contextually relevant, aligned with the nuances of human reasoning and emotional characteristics of speech.

The second-best performances varied among the other models: Llama2-Chat-13B showed strong results with the second-highest scores in ROUGE, BERTScore, and MoverScore, reflecting its capability to produce responses with significant lexical and semantic resemblance to human output. Meanwhile, Zephyr-7B claimed the second spot in the BLEURT metric, indicating a reasonable level of semantic understanding, though not as pronounced as GPT-4 or Llama2-Chat-13B.

This layered analysis reveals that while no models currently achieve perfect, high-quality human-like reasoning, GPT-4 demonstrates better performance in generating reasoning responses compared to the other two models. This finding aligns with the outcomes of human assessments, indicating that recent advancements in model training are progressively reducing the cognitive discrepancies between human and machine-generated reasoning.

Table 4: The baseline results for ROUGE-L, BLEURT, BERTScore, and MoverScore were obtained from various LLMs. The highest scores for each model are highlighted in bold, and the second highest with underline.

	Zephyr-7B	Llama2-Chat-13B	GPT-4
ROUGE-L	0.240	<u>0.244</u>	**0.279**
BLEURT	<u>0.397</u>	0.396	**0.411**
BERTScore	0.582	<u>0.585</u>	**0.593**
MoverScore	0.300	<u>0.301</u>	**0.310**

To determine whether the differences in performance between LLMs across various evaluation metrics are statistically significant,

we employ a pairwise t-test approach. Specifically, we regard the differences as statistically significant if the resulting p-values are less than or equal to 0.05. The p-values for each metric are listed in Table 5, with statistically significant results marked by asterisks.

Analysis shows significant performance differences between models, particularly in ROUGE-L and BLEURT metrics. Specifically, GPT-4 consistently demonstrates statistically significant differences compared to both Zephyr-7B and Llama2-Chat-13B in ROUGE-L and BLEURT, indicating its better performance in these areas. For the BLEURT metric, the comparison between GPT-4 and Zephyr-7B is highly significant, with a p-value below 0.001, suggesting a robust disparity in BLEURT performance favoring GPT-4. However, in the BERTScore metric, none of the model comparisons reached statistical significance, indicating comparable performance across all models in this metric. Lastly, in the MoverScore metric, significant differences were also observed between GPT-4 and the other two models, with more pronounced significance in the comparisons with Llama2-Chat-13B. These findings underscore GPT-4's generally higher ability to reason across most assessed metrics, except for BERTScore, where all models performed similarly.

Table 5: T-test comparison of LLMs showing statistically significant differences across metrics, indicated by asterisks.

	Zephyr-7B	Llama2-Chat-13B	GPT-4
ROUGE-L			
Zephyr-7B	–	0.217	0.011 *
Llama2-Chat-13B	0.217	–	0.026*
GPT-4	0.011 *	0.026*	–
BLEURT			
Zephyr-7B	–	0.926	0.0008 ***
Llama2-Chat-13B	0.926	–	0.019*
GPT-4	0.0008 ***	0.019*	–
BERTScore			
Zephyr-7B	–	0.631	0.402
Llama2-Chat-13B	0.631	–	0.812
GPT-4	0.402	0.812	–
MoverScore			
Zephyr-7B	–	0.128	0.034*
Llama2-Chat-13B	0.128	–	0.009 **
GPT-4	0.034*	0.009 **	–

p-values codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

To evaluate the effectiveness of LLMs in understanding individual mental states for ToM reasoning, we systematically tuned the prompt template to integrate these mental state dimensions, following the format outlined in Section §4.1.2. Initially, we modified the prompt template similarly to the one shown in Figure 4, to reflect the sentiment of the question. Subsequent adjustments were made to accommodate the expressed emotion, and the prompt was refined to align with the underlying intention of the question. Ultimately, we incorporated all three mental states—sentiment, emotion, and intention—into the prompt template (Figure 4), providing a comprehensive assessment of the LLMs' reasoning capabilities. The impact of incorporating these elements is detailed in Table 6, with an extensive analysis of these findings provided in the following sections.

4.2.1 ***LLM reasoning with a focus on "Sentiment"****:* To assess how understanding the sentiment of questions affects reasoning capabilities, this phase integrates sentiment information into the prompt template. As described in Section §4.1.2, we categorize questions into positive, negative, and neutral sentiments. We then tune the prompt template based on these categories and instruct LLMs to generate reasoning responses accordingly.

In this evaluation, GPT-4 clearly outperforms all other models across various metrics, thereby consolidating its status as the top-performing model in sentiment-based ToM reasoning. GPT-4 consistently excels at interpreting and responding based on the sentiment of posts. However, in MoverScore, the margin of its lead narrows, with GPT-4 scoring 0.316, just slightly ahead of Zephyr-7B's 0.310. Meanwhile, Llama2-Chat-13B consistently ranks as the second-best performer in these metrics, although it falls behind in MoverScore. This indicates that while Llama2-Chat-13B is reliable, it does not lead the field in reasoning among the models tested.

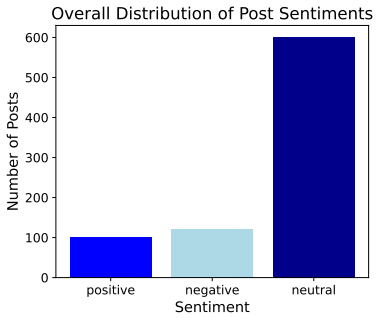


Figure 5: The overall distribution of sentiments of the posts.

When comparing the evaluation metrics of the sentiment-based reasoning output to the original reasoning output (Table 6), we did not observe significant differences. To investigate further, we analyzed the sentiment of the posts, as shown in Figure 5. The lack of improvement could be due to the predominance of 'neutral' classifications. The overwhelming presence of neutral sentiments in the dataset may dilute the impact of sentiment-based enhancements, as neutral sentiments typically provide fewer distinct cues for the model to leverage in its reasoning processes.

4.2.2 ***LLM reasoning with a focus on "Emotion"****:* This step involves incorporating questioners' emotions into the prompt for reasoning generation, based on the extracted NER data. Similar to our analysis of sentiment, we tuned the prompts according to the extracted emotions and instruct the LLMs to generate reasoning responses accordingly.

As shown in Table 6, in the realm of emotion-based tuning, GPT-4 continues to perform better than the other two LLM models, scoring the highest across all evaluation metrics. Notably, it achieves a ROUGE-L score of 0.291 and a BLEURT score of 0.456, underscoring its adeptness at integrating emotional context into its reasoning. The competition, particularly from Llama2-Chat-13B, remains intense, with close scores such as 0.286 in ROUGE-L and 0.441 in BLEURT, reflecting a tight race in model performance when emotion is a significant factor. These scores suggest that all models exhibit improved ToM reasoning when emotional information was integrated into the prompt.

4.2.3 ***LLM reasoning with a focus on "Intention"****:* To explore the effect of understanding the intentions behind questions on the

reasoning capabilities of LLMs, we incorporate these intentions into the prompt template for reasoning generation, using intentions extracted via NER techniques. This step involves prompt tuning based on these intentions, guiding LLMs to generate responses that consider the specified intent.

Based on the result it is evident that GPT-4 outperforms other models, achieving the highest scores across all evaluation metrics: ROUGE-L at 0.300, BLEURT at 0.452, BERTScore at 0.631, and MoverScore at 0.368. Notably, Zephyr-7B achieves the second-highest score in ROUGE-L, while Llama2-Chat-13B secures the second position in the remaining metrics.

By comparing these outcomes with the performance metrics from the original model of reasoning (Table 6), a clear trend of improvement is observable across all metrics for all LLMs tested. This improvement underscores the significant impact that incorporating the explicit intention into the prompts improves LLMs' ability to approximate human-like reasoning. Such findings validate the effectiveness of integrating intent into ToM reasoning.

Table 6: The results from ROUGE-L, BLEURT, BERTScore, and MoverScore, obtained across various LLMs, were generated with prompt tuning based on users' emotion, intent, and sentiment of posts. The highest scores for each model are highlighted in bold, and the second highest with underline.

	Zephyr-7B	Llama2-Chat-13B	GPT-4
Sentiment			
ROUGE-L	0.244	0.250	**0.286**
BLEURT	0.391	0.401	**0.428**
BERTScore	0.586	0.592	**0.599**
MoverScore	0.310	0.303	**0.316**
Emotion			
ROUGE-L	0.284	0.286	**0.291**
BLEURT	0.440	0.441	**0.456**
BERTScore	0.621	0.622	**0.627**
MoverScore	0.342	0.346	**0.350**
Intention			
ROUGE-L	0.299	0.291	**0.300**
BLEURT	0.442	0.445	**0.452**
BERTScore	0.629	0.630	**0.631**
MoverScore	0.357	0.360	**0.368**
Emotion + Sentiment + Intention			
ROUGE-L	0.312	0.299	**0.357**
BLEURT	0.460	0.463	**0.480**
BERTScore	0.641	0.647	**0.661**
MoverScore	0.381	0.385	**0.402**

4.2.4 ***LLM reasoning with a focus on "Sentiment + Emotion + Intention"****:* In the combined tuning scenario, we use the prompt template from Section § 4.1.2 to incorporate sentiment, emotion, and intention into reasoning. This comprehensive prompt allows for a thorough exploration of how these intertwined mental states affect model performance.

The scores evaluating the reasoning capabilities of the models are shown in Table 6. Based on these results, GPT-4 demonstrates superior performance across all metrics. It achieves a ROUGE-L score of 0.327 and a BLEURT score of 0.472, showcasing its ability to incorporate and reason based on mental states. Llama2-Chat-13B often achieves the second-highest scores, including a BERTScore of 0.647 and a MoverScore of 0.385. These scores reflect its ability to adapt to the complexities of considering emotion, sentiment, and intention simultaneously. Though it does not outperform GPT-4, Llama2-Chat-13B demonstrates capabilities in generating reasoning responses.

Table 7: Delta improvement in reasoning capabilities of LLMs through prompt tuning based on posts' sentiment, emotion, and intention.

	Zephyr-7B	Llama2-Chat-13B	GPT-4
ROUGE-L	↑ 7.29%	↑ 5.51%	↑ 7.88%
BLEURT	↑ 6.30%	↑ 6.73%	↑ 6.94%
BERTScore	↑ 5.91%	↑ 6.29%	↑ 6.83%
MoverScore	↑ 8.14%	↑ 8.45%	↑ 9.27%

For a deeper analysis of the performance enhancements from incorporating a combination of emotion, sentiment, and intention into LLM reasoning prompts, we computed the delta changes between the combined format and the baseline results in Table 4. The delta results, displayed in Table 7, illustrate performance improvements across all evaluated metrics for each model.

In the ROUGE-L metric, all models showed notable improvements when prompted with combined mental states. GPT-4 led with a 7.88% increase, followed by Zephyr-7B with 7.29% and Llama2-Chat-13B with 5.51%. For the BLEURT metric, GPT-4 again led with a 6.94% improvement, indicating its ability to generate semantically rich text. Llama2-Chat-13B improved by 6.73%, and Zephyr-7B by 6.30%. In BERTScore, GPT-4 saw a 6.83% rise, Llama2-Chat-13B improved by 6.29%, and Zephyr-7B by 5.91%. MoverScore showed the most pronounced improvements, with GPT-4 increasing by 9.27%, Llama2-Chat-13B by 8.45%, and Zephyr-7B by 8.14%. These gains highlight the models' enhanced ability to capture fine-grained details of mental state information in the prompt template, which is crucial for advanced language understanding.

The analysis clearly shows that integrating emotion, sentiment, and intention into LLM prompts improves model performance, as evidenced by the human evaluation (Section § 4.1.2). This integration not only improves the overall quality of the generated text but also aligns it more closely with human cognitive and emotional processing, making the outputs more natural. The consistent improvements across different models and metrics validate the effectiveness of prompt tuning based on mental states in advancing LLMs' capabilities in natural language understanding and reasoning.

5 DISCUSSION

5.1 LLMs' capability in zero-shot reasoning with open-ended questions (RQ1)

The analysis of annotations by human evaluators indicates that the reasoning responses to open-ended questions generated by LLMs do not meet high-quality standards, as detailed in Table 2. This finding underscores a critical challenge in the field of artificial intelligence, specifically in the development and training of LLMs. Human evaluators consistently report that responses from these models, while often structurally sound and linguistically coherent, lack the depth, nuance, and contextual awareness inherent in human reasoning [9, 11, 53].

Moreover, as demonstrated in Table 4, the quantitative metrics used to evaluate model performance corroborate these observations by empirically highlighting the existing gaps, which align with the

results of the human evaluation. These metrics, which assess aspects ranging from lexical similarity to semantic coherence, consistently reveal that despite their sophistication, LLMs still fall short of the natural reasoning processes humans use in responding to open-ended questions.

In conclusion, although LLMs are capable of mimicking human-like text generation, they still cannot fully replicate human-like ToM reasoning in open-ended questions. These results do not align with those of Bubeck et al. [5], Kosinski [18], who suggested that modern LLMs reached high ToM reasoning ability. We contend that these assertions are overly broad, derived from a limited focus on one aspect of ToM, and based on a minimal set of examples. Furthermore, their experiments were conducted using multiple-choice and short-answer questions, whereas we explored LLMs' ToM abilities with open-ended questions, which more closely mirror real-life scenarios. In accordance with [14, 40, 41, 46, 57], our analysis indicates that even the most advanced models are inadequate in ToM reasoning.

5.2 LLM alignment with human ToM reasoning (RQ2)

Based on the results presented in Table 4 and 6, our analysis indicates that there is not a substantial alignment between human and LLM ToM reasoning capabilities when addressing open-ended questions. While LLMs mimic some aspects of human reasoning, they still fall short of fully replicating the depth and nuance inherent in human cognitive processes [10, 14, 55]. This suggests that despite advancements in LLM capabilities, significant discrepancies remain, underscoring the ongoing challenge of achieving a true equivalence in reasoning between humans and LLMs for open-ended questions.

5.3 The impact of a questioner's mental state on LLM performance in ToM reasoning (RQ3)

The integration of emotional and intentional information into LLMs prompt has enhanced their capacity for reasoning, refining their effectiveness not only in multiple-choice and short-answer queries [9, 14] but also in open-ended questions. As shown in Table 6 and 2, by integrating mental state information, LLMs are able to generate higher-quality reasoning responses.

This improvement is evident in enhanced performance metrics, showing that LLMs are increasingly able to process complex emotional and intentional information. This development is essential for ToM reasoning, enabling the models to handle open-ended queries with a depth similar to human interactions.

Despite advancements, challenges persist in achieving human-like ToM reasoning with LLMs. Prompt tuning has improved LLMs' ability to generate refined reasoning responses, yet it has not fully closed the gap between model outputs and the nuanced responses typical of human reasoning. This ongoing disparity is primarily due to the inherent subjectivity in reasoning, which allows for a wide range of valid responses influenced by individual perspectives and details [7, 13]. To minimize subjectivity and boost evaluation reliability, we analyzed five different answers per query, covering diverse human reasoning and reducing biases. Despite this, subjectivity still impacts the results, likely explaining the gaps in LLM responses compared to human reasoning. Additionally, disparities inherent within the LLM models may also contribute to this shortfall.

6 CONCLUSION

In the evolving landscape of AI, the pursuit of integrating ToM reasoning into LLMs poses opportunities alongside challenges. This study explores how well LLMs' ToM reasoning aligns with human reasoning in handling open-ended questions. Despite advancements in LLMs, our findings reveal a clear gap in their ToM reasoning capabilities for open-ended questions.

Our comprehensive evaluation, utilizing data from Reddit's ChangeMyView subreddit, has elucidated the limitations of LLMs. Despite the proficiency of models such as Zephyr-7B, Llama2-Chat-13B, and GPT-4 in structured tasks like summarization [3, 17, 48], question answering [1, 47], and language translation [15, 29], they exhibit deficiencies in ToM reasoning with open-ended questions. Our findings indicate that these LLMs struggle to produce high-quality reasoning that aligns with human capabilities in such contexts. The significant discrepancy between human and LLM reasoning in ToM can be partly attributed to the inherent subjectivity involved in ToM reasoning [7]. To mitigate this subjectivity, we analyzed five different reasoning strategies for each open-ended question. However, despite this methodological diversification, our results reveal that LLM outputs still diverge considerably from human reasoning, underscoring a critical limitation in their current ToM capabilities.

Our study confirmed that integrating human intentions and emotions through prompt tuning enhances LLMs' ToM reasoning abilities. By embedding individual emotions, intentions, and sentiments within the prompts, we improved the LLMs' ability to generate responses that more closely mirror human ToM reasoning, as demonstrated in [9]. Despite advancements, the reasoning results still need advancements to fully align with human-level reasoning in open-ended questions.

7 LIMITATIONS AND FUTURE WORK

This study shows that incorporating mental states through prompt tuning improves ToM reasoning in LLMs of open-ended questions, yet it faces limitations. The first limitation involves subjectivity. Although we consider five different reasoning answers to cover a diverse spectrum and minimize subjectivity, subjectivity remains a potential factor in the differences observed between LLMs and human reasoning. Moreover, while integrating individual intentions and emotions has proven to boost LLMs' ToM reasoning capabilities, it remains uncertain whether LLMs can inherently account for these mental states in their reasoning processes without explicit prompt tuning. In future work, we will investigate this.

Another limitation of this study stems from its data source; we gathered open-ended questions and responses from the Reddit community. While these responses are tailored to Reddit's unique cultural context, they have been generalized to reflect the broader global community in our analysis. In this categorization, Reddit users are simply referred to as 'humans'. Moreover, it is crucial to recognize that our data is exclusively from English language interactions, potentially limiting the relevance of our findings to LLM behavior across different languages.

ACKNOWLEDGMENTS

This work was supported by University of Washington Bothell Scholarship, Research, and Creative Practice (SRCP) Grant Program.

REFERENCES

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 8–17.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Toufique Ahmed and Premkumar Devanbu. 2022. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–5.

[4] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (12 2008), 555–596. https://doi.org/10.1162/coli.07-034-R2 arXiv:https://direct.mit.edu/coli/article-pdf/34/4/555/1808947/coli.07-034-r2.pdf

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[6] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

[7] Elsa Ermer, Scott A Guerin, Leda Cosmides, John Tooby, and Michael B Miller. 2006. Theory of mind broad and narrow: Reasoning about social exchange engages ToM areas, precautionary reasoning does not. *Social Neuroscience* 1, 3-4 (2006), 196–219.

[8] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12619–12640. https://doi.org/10.18653/v1/2023.findings-emnlp.840

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

[10] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.

[11] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8154–8173. https://doi.org/10.18653/v1/2023.emnlp-main.507

[12] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science* 15, 1 (2023), 29.

[13] Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

[14] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14397–14413. https://doi.org/10.18653/v1/2023.emnlp-main.890

[15] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*. 1–42.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[17] Zahra Kolagar and Alessandra Zarcone. 2024. Aligning Uncertainty: Leveraging LLMs to Analyze Uncertainty Transfer in Text Summarization. In *Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*. 41.

[18] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* 4 (2023), 169.

[19] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1311–1316. https://doi.org/10.18653/v1/D19-1131

[20] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5872–5877. https://doi.org/10.18653/v1/D19-1598

[21] Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in 'theory of mind'. *TRENDS in Cognitive Sciences* 8, 12 (2004).

[22] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 180–192. https://doi.org/10.18653/v1/2023.emnlp-main.13

[23] Terence X Lim, Sidney Tio, and Desmond C Ong. 2020. Improving multi-agent cooperation using theory of mind. *arXiv preprint arXiv:2007.15703* (2020).

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[25] Zhicheng Lin. 2024. How to write effective prompts for large language models. *Nature Human Behaviour* (2024), 1–5.

[26] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive Verification of Chain-of-Thought Reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=I5rsM4CY2z

[27] Xiangyang Liu, Tianqi Pang, and Chenyou Fan. 2023. Federated Prompting and Chain-of-Thought Reasoning for Improving LLMs Answering. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 3–11.

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

[29] Yuan Lu and Yu-Ting Lin. 2023. Characterised LLMs Affect its Evaluation of Summary and Translation. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Bali, Indonesia, 184–192. https://doi.org/10.18653/v1/2023.eval4nlp-1.15

[30] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1011–1031. https://doi.org/10.18653/v1/2023.findings-emnlp.72

[31] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language Models of Code are Few-Shot Commonsense Learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1384–1403. https://doi.org/10.18653/v1/2022.emnlp-main.90

[32] Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=pTHfApDakA

[33] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2392–2400. https://doi.org/10.18653/v1/D18-1261

[34] Yogendra Narayan Prajapati, Sandeep Yadav, Sandhya Sharma, Umesh Kumar Patel, and Swati Tomar. 2023. Analysis of Underlying Emotions in Textual Data Using Sentiment Analysis Which Classifies Text In to Positive, Negative or Neutral Sentiments. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.

[35] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10066–10086. https://doi.org/10.18653/v1/2023.emnlp-main.622

[36] Yewen Pu, Kevin Ellis, Marta Kryven, Josh Tenenbaum, and Armando Solar-Lezama. 2020. Program synthesis with pragmatic communication. *Advances in Neural Information Processing Systems* 33 (2020), 13249–13259.

[37] Leonardo Ranaldi and Andre Freitas. 2024. Aligning Large and Small Language Models via Chain-of-Thought Reasoning. In *Proceedings of the 18th Conference of*

the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1812–1827. https://aclanthology.org/2024.eacl-long.109

[38] Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. 2024. One Law, Many Languages: Benchmarking Multilingual Legal Reasoning for Judicial Support. In *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*. https://openreview.net/forum?id=7vkz7cKd1X

[39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[40] 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*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3762–3780. https://doi.org/10.18653/v1/2022.emnlp-main.248

[41] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13960–13980. https://doi.org/10.18653/v1/2023.acl-long.780

[42] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[43] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2257–2273. https://aclanthology.org/2024.eacl-long.138

[44] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the Emerging Norms of Using Large Language Models in Social Computing Research. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 569–571. https://doi.org/10.1145/3584931.3606955

[45] Winnie Street. 2024. LLM Theory of Mind and Alignment: Opportunities and Risks. *arXiv preprint arXiv:2405.08154* (2024).

[46] Eliza Strickland. 2023. AI Outperforms Humans in Theory of Mind Tests: Large Language Models Convincingly Mimic the Understanding of Mental States. https://spectrum.ieee.org/theory-of-mind-ai. Accessed: 2024-05-20.

[47] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In *International Semantic Web Conference*. Springer, 348–367.

[48] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine* 6, 1 (2023), 158.

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian CantonFerrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, , Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian XiangKuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

[50] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).

[51] Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

[52] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, Jing Jiang, David Reitter, and Shumin Deng (Eds.). Association for Computational Linguistics, Singapore, 389–402. https://doi.org/10.18653/v1/2023.conll-1.25

[53] Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11865–11881. https://doi.org/10.18653/v1/2023.findings-emnlp.795

[54] Hongru WANG, Rui Wang, Fei Mi, Yang Deng, Zezhong WANG, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=FRRlmKxuf2

[55] Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Can LLMs Reason with Rules? Logic Scaffolding for Stress-Testing and Improving LLMs. *arXiv preprint arXiv:2402.11442* (2024).

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[57] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10691–10706. https://doi.org/10.18653/v1/2023.findings-emnlp.717

[58] Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2024. Investigating the effectiveness of task-agnostic prefix prompt for instruction following. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19386–19394.

[59] Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissi, Siddharth Verma, Zhijing Jin, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2023. ALERT: Adapt Language Models to Reasoning Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1055–1081. https://doi.org/10.18653/v1/2023.acl-long.60

[60] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. 2023. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare* 11, 20 (2023), 2776.

[61] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating Large Language Models at Evaluating Instruction Following. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=tr0KidwPLc

[62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

[63] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622* (2019).

[64] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023. How FaR Are Large Language Models From Agents with Theory-of-Mind? *arXiv preprint arXiv:2310.03051* (2023).

[65] Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11136–11155. https://doi.org/10.18653/v1/2023.acl-long.624