RecToM: A Benchmark for Evaluating Machine Theory of Mind in LLM-based Conversational Recommender Systems

Mengfan Li[1*], **Xuanhua Shi**[1†], **Yang Deng**[2]
[1]National Engineering Research Center for Big Data Technology and System,
Services Computing Technology and System Lab, Cluster and Grid Computing Lab,
Huazhong University of Science and Technology
[2]Singapore Management University
{limf, xhshi}@hust.edu.cn, ydeng@smu.edu.sg

Abstract

Large Language models (LLMs) are revolutionizing the conversational recommender systems (CRS) through their impressive capabilities in instruction comprehension, reasoning, and human interaction. A core factor underlying effective recommendation dialogue is the ability to infer and reason about users' mental states (such as desire, intention, and belief), a cognitive capacity commonly referred to as *Theory of Mind* (ToM). Despite growing interest in evaluating ToM in LLMs, current benchmarks predominantly rely on synthetic narratives inspired by Sally-Anne test, which emphasize physical perception and fail to capture the complexity of mental state inference in realistic conversational settings. Moreover, existing benchmarks often overlook a critical component of human ToM: behavioral prediction, the ability to use inferred mental states to guide strategic decision-making and select appropriate conversational actions for future interactions. To better align LLM-based ToM evaluation with human-like social reasoning, we propose RECTOM, a novel benchmark for evaluating ToM abilities in recommendation dialogues. RECTOM focuses on two complementary dimensions: **Cognitive Inference** and **Behavioral Prediction**. The former focus on understanding *what has been communicated* by inferring the underlying mental states. The latter emphasizes *what should be done next*, evaluating whether LLMs can leverage these inferred mental states to predict, select, and assess appropriate dialogue strategies. Together, these dimensions enable a comprehensive assessment of ToM reasoning in CRS. Extensive experiments on state-of-the-art LLMs demonstrate that RECTOM poses a significant challenge. While the models exhibit partial competence in recognizing mental states, they struggle to maintain coherent, strategic ToM reasoning throughout dynamic recommendation dialogues, particularly in tracking evolving intentions and aligning conversational strategies with inferred mental states.

Datasets — https://github.com/CGCL-codes/RecToM

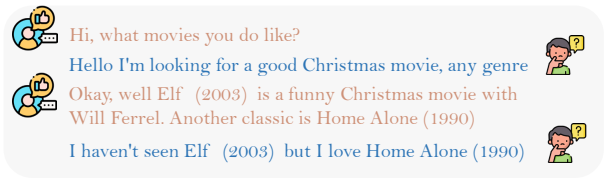
Introduction

Large Language Models (LLMs) have significantly advanced conversational recommender systems (An et al. 2025; He et al. 2025; Huang et al. 2025; Qin et al. 2024; Li et al. 2025), enabling significant proficiency in response generation that closely resembles human dialogue. A key capability that supports effective conversational recommendations is the ability to understand and anticipate others' thoughts, desires, and intentions, which is an ability widely recognized in cognitive science as the "Theory of Mind" (ToM) (Kosinski 2023; Zhang et al. 2025). Investigating ToM in LLM-based conversational recommenders enables a nuanced evaluation of the models' competence in comprehending user preferences, predicting subsequent behaviors, and strategically adapting interactions, thereby improving user satisfaction and engagement in recommendation dialogues. This not only foster a deeper understanding of which specific aspects of LLMs drive effectiveness in conversational recommendation, but also identifiers critical gaps that require targeted improvements to align with human ToM, facilitating more engaging and satisfying user experiences.

Recent advancements in LLMs have fueled growing interest in evaluating their capacity for ToM reasoning (de Carvalho et al. 2025; Friedman et al. 2023). While several benchmarks (Gandhi et al. 2023; Xu et al. 2024; Wu et al. 2023; Jin et al. 2024) have been proposed to evaluate ToM in LLMs, they exhibit significant limitations for assessing conversational recommender systems. One limitation is that many existing works (Jin et al. 2024; Xu et al. 2024; Shi et al. 2025) rely on the Sally-Anne test and similar paradigms, which typically involve simplified scenarios, such as individuals entering a space, moving objects, and others arriving afterward. These setups lack engaging and naturalistic interactions, rendering them ill-suited for complex conversational recommendation systems. A further limitation lies in the predominant focus of current benchmarks (Chan et al. 2024; Jung et al. 2024) on retrospective reasoning about mental states (*e.g.*, beliefs, intentions, desires), based on dialogues that have already transpired. Such benchmarks fail to capture a core aspect of human ToM: the ability to use inferred mental states to guide strategic decision-making for future interactions.

To bridge this gap, we introduce RECTOM, a benchmark for evaluating the ToM capabilities of LLMs specifically within conversational recommender systems as shown in Figure 1. RECTOM situates LLMs in realistic social in-

[*]Work was done during a visit at SMU.
[†]Corresponding author.



Cognitive Inference

Belief reasoning
-Q: How does the recommender believe the seeker's attitude about the Elf (2003)?
-A: The movie is proposed by the recommender, the seeker has not seen it yet, and accepted the suggestion

Intention reasoning
-Q: What is the intention of the recommender in the "Okay, well Elf (2003) is a funny Christmas movie with Will Ferrel. Another classic is Home Alone (1990) " given the dialogue history?
-A: Provides a recommendation directly or by inquiring further
Explains a recommendation with background or reasoning
(Multiple choice & Coarse intention)
-A: The recommender provides recommendation by showing it directly.
The recommender explains recommendation based on the seeker's past preference.
(Multiple choice & Fine intention)

Desire reasoning
-Q: Is the seeker likely to watch the Home Alone (1990)
-A: yes

Behavior Prediction

Prediction reasoning
-Q: What strategy will RECOMMENDER use next?"
-A: Provides a recommendation directly or by inquiring further
Explains a recommendation with background or reasoning
(Multiple choice)

Judgement reasoning
-Q: Recommender will adopt "Explains a recommendation with background or reasoning" strategy to promote communication, Is this strategy effective?"
-A: yes

Figure 1: An example dialogue in RECTOM.

teractions featuring asymmetrical conversational roles (*i.e.*, recommender and seeker), enabling assessment of complex psychological reasoning. Specifically, the benchmark highlights two core reasoning types: (1) ***Cognitive Inference***, which assesses the LLMs' capability to accurately infer and explain the true mental states of the recommender and seeker, such as their desires, beliefs, and intentions, treating these mental states as theoretical constructs supporting observable behaviors, and (2) ***Behavioral Prediction***, which evaluates the LLMs' ability to apply inferred mental states to effectively anticipate conversational actions, such as predicting appropriate dialogue strategies or evaluating the effectiveness of proposed conversational strategies based on dialogue history.

Following existing ToM benchmarks (Chan et al. 2024; Yu et al. 2025), we also adopt the question answering (QA) data format for constructing our RECTOM benchmark, while there are several distinctive features specifically designed for conversational recommendation:

- *Multi-choice Strategy.* An utterance from either the recommender or seeker may express multiple distinct intentions within a single sentence. Table 1 presents the distribution of different question types and an analysis of their corresponding answer options.
- *Multi-granular Intention.* Intentions in dialogue are inherently hierarchical: a single utterance can convey both a high-level purpose and nuanced, context-dependent sub-intentions. Figure 2 illustrates the categorization of intentions into coarse-grained and fine-grained levels.
- *Multi-dimensional Belief.* In conversational recommendation systems, beliefs about an item (*e.g.*, a film) are not uni-dimensional, but rather involve multiple interrelated aspects, such as who introduces the film (seeker or recommender), whether the seeker has viewed it, and their level of preference or acceptance, all contribute to a more nuanced mental reasoning.
- *Multi-concurrent Desire.* Recommendation dialogues often involve the simultaneous pursuit of multiple goals, such as exploring diverse film options and comparing alternatives. RECTOM captures this complexity by modeling the seeker's concurrent inclinations toward each recommended item, reflecting coexisting preferences that require independent evaluation.

To the best of our knowledge, RECTOM is the first human-annotated conversational recommendation benchmark to introduce ToM evaluation for LLMs in realistic recommendation scenarios. Experiments on the state-of-the-art LLMs reveal several key findings regarding the modeling of ToM in CRS:

(1) *Increased option complexity hinders ToM reasoning in CRS.* LLMs exhibit the significantly lower accuracy on multiple choice questions compared to single choice ones, indicating that their ability to infer the intentions of dialogue participants deteriorates as the choice space becomes more complex. This limitation highlights a fundamental challenge in CRS: capturing the nuanced mental states in dynamic and multi-faceted interactions.

(2) *Fine-Grained intent discrimination remains a key bottleneck.* While LLMs perform well on coarse-grained intention classification, their performance drops notably on fine-grained tasks. This gap reflects a critical limitation in current CRS: the inability to effectively model the subtle and evolving preferences of participants during conversation, essential for accurate and context-aware recommendation.

(3) *LLMs exhibit early potential for multi-dimensional mental state reasoning.* Despite performance limitations, LLMs show some capacity to integrate multiple contextual cues into a coherent reasoning process. This indicates early potential for modeling complex seeker states in CRS, such as belief attribution, which are essential for generating contextually appropriate and personalized recommendations.

(4) *LLMs exhibit a systematic bias towards sycophantic or "pleasing" responses.* In open-ended scenarios, LLMs frequently produce responses that align with perceived participants preferences or expectations, even when such responses are not factually or logically sound. This tendency, consistent with the "Answer Sycophancy" phenomenon (Sharma et al. 2024), poses a serious risk in CRS, where overly agreeable affirmations can lead to suboptimal experiences when it comes to the judgement prediction of subsequent behaviors.

Question Type	Quantity	# Options	Answer Type
Desire (Seek)	1,448	2	single
Coarse Intention (Rec/Seek)	2,205/2,205	5/4	multiple
Fine Intention (Rec/Seek)	2,205/2,205	10/16	multiple
Belief (Rec)	1,762	7	single
Prediction (Rec/Seek)	2,098/2,149	5/4	multiple
Judgement (Rec/Seek)	2,098/2,149	2/2	single

Table 1: Options statistics of RECTOM benchmark

(5) *Chain-of-thought (CoT) prompting yield limited benefits in complex recommendation tasks.* Contrary to expectations, CoT provides only marginal gains in ToM reasoning within CRS, and in some cases leads to performance degradation. This indicates that current prompting strategies may fail to effectively scaffold coherent, multi-step reasoning about mental states in realistic, context-rich CRS.

Related Works

LLM-based Conversational Recommendation

Recent advances in LLMs have significantly influenced the development of CRS (An et al. 2025; He et al. 2025; Huang et al. 2025). Thanks to their strong language understanding and generation capabilities (Naveed et al. 2023), LLMs demonstrate promising performance in several key aspects of CRS, including response quality, natural language understanding, and personalized recommendation generation (Bao et al. 2023; Karanikolas et al. 2023; Deng et al. 2023).

While LLMs excel at generating fluent and seemingly intelligent responses, it remains unclear whether they can accurately model the underlying mental states (*e.g.*, intentions, beliefs and desires) of both the recommender and the seeker throughout the conversation, or whether they truly engage in socially aware and contextually appropriate decision-making is still an open question.

Theory of Mind (ToM) Benchmarks

ToM, the ability to attribute and reason about mental states, has gained increasing attention in both cognitive science and artificial intelligence (Kosinski 2023; Zhang et al. 2025; Gandhi et al. 2023). In recent years, several benchmarks have been proposed to evaluate ToM reasoning in LLMs. Notable examples include Hi-ToM, FANTOM, Persuasive-ToM, OpenToM, AutoToM, NegotiationToM, MumA-ToM, and MMToM-QA (Wu et al. 2023; Kim et al. 2023; Yu et al. 2025; Xu et al. 2024; Zhang et al. 2025; Chan et al. 2024; Shi et al. 2025; Jin et al. 2024), which assess the model's ability to understand beliefs, intentions, and desires through narrative comprehension or dialogue reasoning tasks.

While these efforts have advanced our understanding of ToM capabilities in language models, existing benchmarks primarily focus on general purpose (Xu et al. 2024; Zhang et al. 2025; Shi et al. 2025; Jin et al. 2024) or task-oriented settings (Chan et al. 2024; Yu et al. 2025), and often abstract away from the nuanced, domain specific of real world conversational systems. To date, RECTOM is the first and only benchmark that systematically evaluates ToM reasoning in the context of CRS, where effective interaction relies on understanding the underlying mental states of both participants. Furthermore, our benchmark further captures more complex psychological dynamics, such as asymmetric roles, hierarchical intention structures, and evolving preferences.

RECTOM Benchmark

Overview

By constructing the RECTOM benchmark, we aim to assess the theory of mind capabilities of LLMs by answering following questions: **(1) Can LLMs reason about mental states within a multiple choice context?** For instance, a single utterance may encode multiple intentions, *e.g.*, a seeker's statement might simultaneously convey a request for recommendations and a preference for horror genres. **(2) How consistent is their performance across different granularity levels of mental state inference?** For example, can models equally identify both coarse-grained intentions (*e.g.*, "request") and fine-grained nuances (*e.g.*, requesting preferences, seeking feedback, or asking clarifying questions)? **(3) Can they integrate multi-dimensional reasoning to understand the mental state comprehensively?** This question examines whether LLMs can synthesize diverse and interrelated factors that collectively shape an agent's internal state. For instance, in assessing the seeker's attitude toward a movie, the model must jointly consider who proposed the movie, whether the seeker has seen it, and whether they ultimately accepted or rejected it. **(4) Do LLMs exhibit a tendency to ingratiate through overly affirmative responses?** For instance, when evaluating the effectiveness of a proposed conversational strategy, do models provide reasoned and objective judgments based on deliberation, rather than merely catering with reflexive positive responses?

Data Collection

The multi-turn conversational recommendation data used in this work is derived from the REDIAL dataset (Li et al. 2018), a publicly available corpus centered on movie recommendation dialogues. In REDIAL, each dialogue involves two participants: seeker and recommender. Moreover, to ensure dialogue quality and meaningful interaction, we follow the selection protocol established by IARD (Cai and Chen 2020), and selects 253 satisfactory recommendation dialogues, those in which the seeker initially rejects a recommended movie but later accepts a subsequent suggestion, and 83 unsatisfactory dialogues, where no recommendation is accepted by the seeker.

We further process the selected dialogues through two manual refinement steps: (1) *Belief: identifying the final acceptance status of the recommended item.* For each dialogue, we locate the exact utterance where the seeker explicitly

Type	RECTOM Questions
Desire Reasoning	Is the <Seeker> likely to watch the <movie>?
Intention Reasoning	What is the intention expressed by the <Recommender/Seeker> in the <utterance> given the dialogue history?
Belief Reasoning	How does the <Recommender> believe the <seeker's> attitude about the <movie>?
Prediction Reasoning	What strategy will <Recommender/Seeker> use next?
Judgement Reasoning	<Recommender/Seeker> will adopt <strategy> to promote communication, Is this strategy effective?

Table 2: ToM questions from RECTOM benchmark

expresses their opinion on a recommended movie. (2) *Desire: annotating multi-dimensional desires.* We re-annotate three core dimensions for each movie mentioned, including **suggestion** (whether the movie was suggested by the recommender or initiated by the seeker), **seen** (whether the seeker has seen the movie), and **liked** (whether the seeker liked the movie or the recommendation). Specificlly, three PhD students (trained in ToM knowledge and prior psychology projects) annotated the data. Two annotators labeled initially, conflicts resolved by a third. The IAA score (Fleiss's K) (Fleiss 1971) is 0.79. Table 3 presents the statistical summary of the RECTOM benchmark.

Statistic	Value
# Dialogues	336
# Total turns	4,583
# Question types	10
# QA pairs	20,524
# Avg. Turns per dialogue	13.64
# Avg. Movies per dialogue	5.24

Table 3: RECTOM Benchmark Statistics

Table 2 illustrates the question types in RECTOM, organized into two reasoning categories: *Cognitive Inference* and *Behavioral Prediction*. The former targets mental-state attribution, encompassing questions about desires, intentions, and beliefs. The latter involves strategic prediction and effectiveness judgment, reflecting the application of inferred mental states to guide conversational actions. Notably, the recommender and seeker characterize asymmetrical social status in complex psychological activities. (*e.g.*, recommender is expected to infer the seeker's belief about the movies and while the reverse is not required). This reflects the realistic dynamics of recommendation dialogues, where the recommender, as the proactive agent initiating the interaction, must primarily evaluate the seeker's attitude and desires to guide effective communication.

Cognitive Inference

In RECTOM, cognitive inference is decomposed into three core components: desire, intention, and belief reasoning, corresponding to the Belief-Desire-Intention (BDI) model (Bratman 1987) of mental-state attribution.

Desire Reasoning Desire represents a motivational state that drives behaviors thought it does not imply a firm commitment (Kavanagh, Andrade, and May 2005; Malle and Knobe 2001). In the context of CRS, desire reflects a seeker's latent interest or inclination toward specific items, such as movies, which may evolve dynamically through the interaction. In RECTOM we evaluate whether LLMs can infer and track the evolving motivational state through questions of the form "Is the seeker likely to watch the <movie>?". These questions access whether the seeker is likely to engage with a particular movie, based on their expressed preferences and contextual cues, and are presented with binary choices ("yes" or "no").

Belief Reasoning Belief refers to a cognitive state in which an agent holds a specific understanding or assumption about another agent's perspective or attitude toward a proposition. In CRS, belief reasoning involves understanding how the *recommender* infers the *seeker's* attitude toward a suggested item. In RECTOM, we evaluate whether LLMs can infer the recommender's belief about the seeker's stance toward a recommended movie through belief reasoning questions. Inspired by the multi-dimensional annotation schema in REDIAL (Li et al. 2018), we decompose belief into three key dimensions: *Suggestion*:whether the movie was suggested by the recommender or initiated by the seeker, *Seen*: whether the seeker has seen the movie and *liked*: whether the seeker liked the movie or the recommendation. This design require models to interpret contextual cues, such as explicit preferences, indirect feedback, and prior statements, and to dynamically update beliefs as the conversation progresses.

Intention Reasoning Intention refers to an agent's deliberate commitment to perform an action, typically grounded in their beliefs and desires, and directed toward achieving a specific goal (Phillips, Wellman, and Spelke 2002). In conversational systems, modeling intention is essential for understanding the purpose behind each utterance, especially in goal oriented interactions such as recommendation dialogues. In *RecToM*, we evaluate whether LLMs can identify the intentions underlying the utterances of both the recommender and the seeker, using the questions of the form "What is the intention expressed by the Recommender/seeker in the <utterance >, given the dialogue history?". Models are required to reason about both coarse-grained and fine-grained intention categories, reflecting increasing levels of specificity in communicative purpose. The

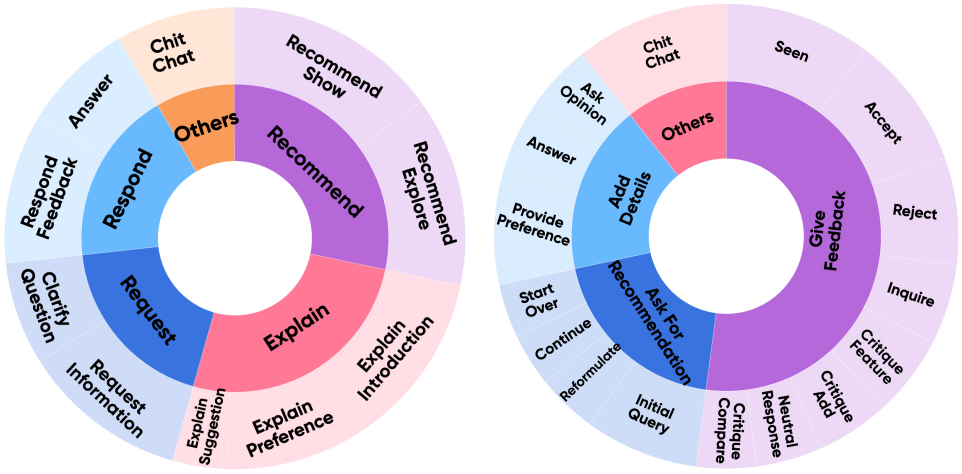


Figure 2: Coarse-Grained and Fine-Grained Intention Classification for Recommenders (Left) and Seekers (Right) in the RECTOM Benchmark, with segment sizes in the doughnut charts reflecting the frequency of each intention category.

full set of intention options, detailed in Figure 2, includes 10 fine-grained categories for the recommender and 16 for the seeker, capturing a wide range of conversational strategies. These questions require the model to understand not only *what is being communicated*, but also *why*, a core component of advanced ToM reasoning in conversational AI.

Behavioral Prediction

While cognitive inference plays the crucial role in understanding the mental states of recommender and seeker, it is equally important to leverage these inferred states to inform action. specifically, to guide the selection of effective recommendation strategies and evaluate the effectiveness on the dialogue outcomes.

Prediction Reasoning Prediction reasoning involves anticipating the dialogue strategies that the recommender or seeker is likely to employ in the next turn. This is operationalized through questions of the form: "What strategy will recommender/seeker use next?". Given the diversity of possible dialogue strategies and the potential for multiple strategies to be expressed within a single utterance, this task is framed as a multiple choice problem. Successfully answering these questions require LLMs to infer the current conversational state and generate plausible predictions about future interactions. This, in turn, influences the dynamic evolution of the recommender's and seeker's beliefs, desires and intentions, making prediction reasoning a key component of effective and proactive dialogue modeling.

Judgment Reasoning Judgment Reasoning assesses the model's ability to evaluate the effectiveness of a given dialogue strategy in advancing the conversation. In this task, a strategy is randomly specified, and models are asked to judge its likely impact using questions of the form: using the form of: "The recommender/seeker will adopt <strategy>to promote communication. Is this strategy effective? " Answering correctly requires higher-order reasoning about the participants' current beliefs, intentions, and dialogue context, going beyond surface-level understanding to assess strategic appropriateness. This capability is critical for adaptive conversational agents, enabling them to reflect on and improve their interaction strategies to enhance long-term user engagement.

By integrating both cognitive and behavioral inference, RECTOM offers a comprehensive evaluation of ToM capabilities in conversational recommendation settings. It bridges the gap between cognitive theory and real world applications in conversational artificial intelligence, moving beyond mere comprehension of mental states to modeling their role in strategic interaction.

Experiments

Baseline Models

We evaluate RECTOM on five state-of-the-art LLMs from diverse sources with varying levels of reasoning abilities. **Deepseek-V3** (Liu et al. 2024): a robust Mixture-of-Experts (MoE) language model featuring a total 671 billion total parameters, with 37 billion activated for each token processing. **GPT-4o-mini** (OpenAI 2024b) and **GPT-4o** (OpenAI 2024a): both are multimodal, multilingual generative pretrained transformer models developed by OpenAI (Abacha et al. 2024). **Gemini 2.5 Flash-Lite** (Comanici et al. 2025): developed by Google and designed to provide ultra-low-latency performance and high throughput per dollar. **Claude 3.5 Haiku** (Anthropic 2024): a fast and cost-effective language model from Anthropic.

Following established practices in the theory of mind literature (Sabour et al. 2024; Kim et al. 2023) we evaluate these models with two types of prompting strategies: (1) vanilla zero-shot prompting directly asks LLMs to select the answer (single or multiple options) without providing any explanation. (2) Chain-of-thought (CoT) prompting, adapted from (Kojima et al. 2022; Wei et al. 2022), in which the model is instructed with the prompt "Let's think step by step." to encourage explicit reasoning. The final answer is

Model	Cognitive Inference						Behavioral Prediction			
	Recommender			Seeker			Recommender		Seeker	
	Fine Intention	*Coarse Intention*	Belief	*Fine Intention*	*Coarse Intention*	Desire	*Prediction*	Judgement	*Prediciton*	Judgement
Random Guess	*0.10*	*3.23*	14.29	*0.00*	*6.67*	50.00	*3.23*	50.00	*6.67*	50.00
Human	*64.32*	*86.31*	96.84	*59.92*	*82.74*	98.25	*87.44*	96.37	*85.18*	97.23
DeepSeek-v3	*29.71*	*44.26*	69.86	***33.20***	*59.32*	86.05	***26.84***	**39.18**	*48.02*	35.60
GPT-4o-mini	*27.80*	*38.01*	52.50	*31.43*	*54.24*	88.60	*18.88*	32.22	*11.91*	31.97
GPT-4o	*32.61*	*40.45*	74.74	*28.84*	***64.22***	**92.27**	*24.07*	33.84	***49.23***	32.34
Gemini 2.5 Flash-Lite	*25.90*	*37.64*	63.73	*22.31*	*58.50*	89.78	*22.07*	36.80	*47.98*	36.11
Claude 3.5 Haiku	*25.74*	*36.55*	61.58	*29.11*	*45.35*	90.68	*21.64*	32.41	*14.15*	33.04
DeepSeek-v3+cot	***33.02***	***46.21***	**79.46**	*29.61*	*58.59*	76.10	*19.54*	37.94	*38.11*	35.55
GPT-4o-mini+cot	*26.17*	*42.90*	63.11	*28.53*	*55.15*	86.12	*18.73*	31.46	*15.87*	31.69
GPT-4o+cot	*28.44*	*42.40*	75.20	*25.44*	*54.10*	87.78	*21.54*	32.94	*27.97*	32.39
Gemini 2.5 Flash-Lite+cot	*24.31*	*41.63*	74.80	*27.66*	*56.78*	87.36	*24.83*	36.32	*42.25*	37.69
Claude 3.5 Haiku+cot	*23.78*	*41.27*	72.19	*25.91*	*51.47*	78.73	*7.24*	35.32	*10.33*	**39.04**
Model Average	*27.74*	*41.13*	68.72	*28.20*	*55.77*	86.35	*20.54*	34.84	*30.59*	34.53

Table 4: Main results of models on RECTOM (accuracy in %). The best results are **bold-faced**, and the second-best are underlined, *italics* indicate multiple choice questions.

then extracted via string matching from a fixed output format. the temperature for all model generations is set to 0.7 to balance creativity and determinism.

Main Results

LLMs demonstrate notable yet uneven performance across cognitive inference and behavioral prediction tasks in the CRS. While most models significantly outperform random guessing, indicating a basic capacity to extract and reason about mental states such as beliefs and intentions from dialogue context, substantial gaps remain compared to human-level performance. Even with the zero-shot CoT prompting, improvements are marginal and inconsistent. The comprehensive evaluation on RECTOM, summarized in Table 4, reveals critical insights into the capabilities and challenges of LLMs in ToM reasoning within CRS.

First, cognitive load from multiple choice formats impairs decision accuracy. On multi-choice tasks requiring discrimination among numerous plausible mental state attributions, LLMs' performance is markedly low (see Table 4 results in *italics*), averaging only **27.74%** of fine-grained intention reasoning for the recommender role. In contrast, performance on single choice tasks, such as belief reasoning (**68.72%**) and desire reasoning (**86.35%**), is substantially higher. This pronounced performance gap highlights the difficulty LLMs face in managing increased cognitive load when distinguishing between nuanced and plausible alternatives, particularly in complex, multiple choice inference scenarios.

Second, a significant granularity gap in intention inference reveals fundamental representational deficits. While LLMs achieve moderate accuracy in coarse-grained intention classification (*e.g.*, GPT-4o: **64.22%** for seeker), performance sharply declines in fine-grained tasks (*e.g.*, GPT-4o: **28.84%** for seeker), exposing their limited capacity to capture the subtle, context-dependent evolution of participant preferences. This deficit hinders the delivery of truly adaptive and personalized recommendations in CRS.

Third, LLMs exhibits a non-trivial yet limited capacity for multi-dimensional belief inference in CRS. The top-performing model (Deepseek-v3 + cot) achieves **79.46%** accuracy, substantially exceeding the random baseline (**14.29%**) and outperforming smaller models such as GPT-4o-mini (**52.50%**). This indicates that, under favorable conditions, sufficient model scale and structured prompting, LLMs can integrate conversational history and social cues to form coherent, albeit imperfect, inferences about the recommender's beliefs regarding seeker's attitudes.

Fourth, the efficacy of CoT prompting in realistic conversational reasoning workflows is limited and inconsistent. Despite its success in other domains, CoT yields only marginal improvements in this CRS context, such as a **+1.95** percentage point (pp) gain for DeepSeek-v3 in coarse grained intention inference of recommender, and **+0.46%** pp for GPT-4o in belief reasoning, and no improvement or even degradation in many cases (*e.g.*, GPT-4o in coarse grained intention of seeker: **64.22%** $\rightarrow$ **54.10%**). This variability suggests that CoT does not reliably enhance multi-step reasoning in complex, context-sensitive dialogues and may introduce noise or redundant reasoning that disrupts decision accuracy.

These results highlight the gap between surface-level language understanding and the deeper, inference-driven reasoning necessary for effective ToM in CRS.

In-depth Analysis

Judgement Reasoning Bias. As shown in Table 4, in the binary classification of judgement reasoning task, LLMs perform below random guess, indicating a systematic distortion driven by preference-conforming bias. To further as-

Model	Prediction Bias(↓)	FPR(↓)	Recall (↑)
GPT-4o	94.90	94.45	5.55
GPT-4o+CoT	94.08	94.44	5.56
DeepSeek-v3	88.42	85.84	14.16
DeepSeek-v3+CoT	88.43	86.62	13.38
Claude 3.5	97.86	97.64	2.36
Claude 3.5+CoT	90.09	89.87	10.13
Gemini 2.5	87.94	97.23	12.77
Gemini 2.5+CoT	**84.51**	**85.03**	**14.97**
GPT-4o-mini	98.52	98.27	1.73
GPT-4o-mini+CoT	94.91	95.45	4.55
GPT-4o	99.07	98.91	1.09
GPT-4o+CoT	98.93	98.77	1.23
DeepSeek-v3	93.67	92.56	7.44
DeepSeek-v3+CoT	94.45	92.96	7.04
Claude 3.5	97.63	97.34	2.66
Claude 3.5+CoT	86.03	**84.44**	**15.56**
Gemini 2.5	93.35	91.95	8.05
Gemini 2.5+CoT	**85.14**	84.59	15.41
GPT-4o-mini	99.81	99.73	0.27
GPT-4o-mini+CoT	99.62	99.51	0.49

Table 5: Performance of judgement reasoning task. Results are reported in percentage (%). *Claude 3.5* denotes the Claude 3.5 Haiku; *Gemini 2.5* refers to Gemini 2.5 Flash-Lite. The upper section presents results for the Recommender; the lower section for the Seeker. Best results are highlighted in **bold**; second-best in *italics*.

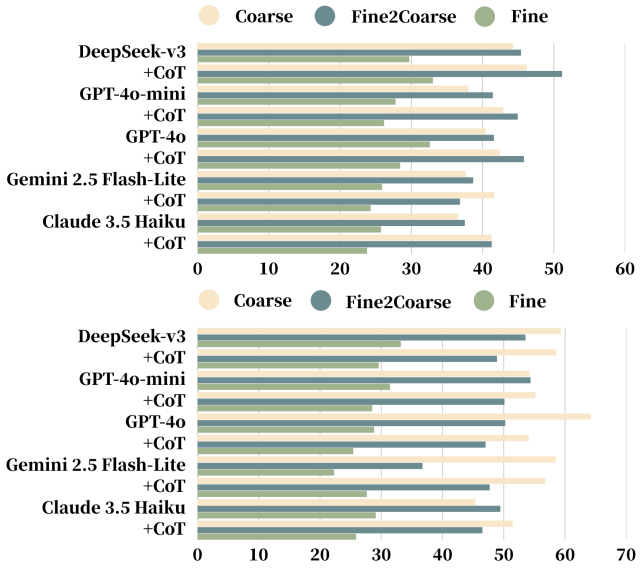


Figure 3: Intention reasoning compared across 10 models (accuracy in %), *Fine2Coarse intention* reflect the accuracy of mapping fine-grained intentions to their predefined coarse-grained categories. The upper section reports results for the recommender; the lower section for the seeker.

sess the LLMs' tendency toward generating overly positive "Yes" responses (a "pleasing" bias), we analysis the confusion matrix using three metrics: *Prediction Bias* (lower is better), defined as $(TP + FP)/(TP + FP + TN + FN)$, measures the proportion of "Yes" predictions; False Positive Rate (FPR, lower is better), calculated as $FP/(FP + TN)$, quantifies the misclassification of actual "No" instances as "Yes"; and Recall for "No" (higher is better), computed as $TN/(FP + TN)$, reflects the model's accuracy in identifying correct "No" responses. As shown in Table 5, the LLMs exhibits a high Prediction Rate of "Yes" (∼**93.37%**) indicating a strong default toward affirmative responses regardless of ground truth. This is further evidenced by an extremely high FPR of ∼**93.28%**, meaning nearly all true"No" instances are incorrectly classified as "Yes". Complementing this, the recall for "No" is only ∼**7.22%**, confirming the LLMs' near inability to correctly identify and respond with "No" when required. Together, these results reveal a severe affirmative bias, consistent with "answer sycophancy" (Sharma et al. 2024), where the LLMs prioritizes favoring agreement over accuracy, undermining its reliability in judgment tasks.

Fine-grained Intention Error Analysis. As shown in Figure 3, the Fine2Coarse task involves a human-annotated mapping: given the LLMs' fine-grained intention outputs, we manually map them to their corresponding coarse-grained categories using predefined rules, thereby eliminating potential model-induced mapping errors.

Empirical analysis reveals fine-grained accuracy is consistently lower than Fine2Coarse accuracy across all models (e.g., DeepSeek-v3: **29.71%** vs. **45.40%** for Recommender; GPT-4o: **28.84%** vs. **50.25%** for Seeker). Fine2Coarse accuracy approximates or is close to coarse-grained accuracy (e.g., DeepSeek-v3+CoT: **51.16%** vs. **46.21%** for Recommender), indicating most fine-grained outputs, though imprecise, still fall within the correct coarse-grained category. The poor performance in fine-grained intention classification stems not from misalignment with the coarse-grained direction (as evidenced by robust Fine2Coarse results) but from weak ability to discriminate between fine-grained options within the same coarse category. Models struggle to pinpoint the exact fine-grained intent, despite correctly identifying the broader coarse-grained scope.

Conclusion

This work introduces RECTOM, a benchmark designed to evaluate the Machine Theory of Mind in LLMs within conversational recommendation systems. The core of RECTOM lies in its structured assessment of cognitive inference and behavioral prediction, characterized by four key dimensions: **multi-choice strategy reasoning**, **multi-granular intentions**, **multi-dimensional beliefs**, and **multi-concurrent desires**. Through comprehensive experiments, we evaluate state-of-the-art LLMs on this benchmark, revealing critical insights into their strengths and limitations in modeling human-like mental state reasoning in realistic CRS.

Acknowledgments

This work was supported by the Major Program (JD) of Hubei Province (No.2023BAA024).

References

Abacha, A. B.; Yim, W.-w.; Fu, Y.; Sun, Z.; Yetisgen, M.; Xia, F.; and Lin, T. 2024. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv:2412.19260*.

An, G.; Zou, J.; Wei, J.; Zhang, C.; Sun, F.; and Yang, Y. 2025. Beyond whole dialogue modeling: Contextual disentanglement for conversational recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 31–41.

Anthropic. 2024. Claude 3.5 Haiku. https://www.anthropic.com/claude/haiku. Accessed: 2024-10.

Bao, K.; Zhang, J.; Zhang, Y.; Wang, W.; Feng, F.; and He, X. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, 1007–1014.

Bratman, M. 1987. Intention, plans, and practical reason.

Cai, W.; and Chen, L. 2020. Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations. In *Proceedings of the twenty-eighth ACM Conference on User Modeling, Adaptation and Personalization, Genoa, Italy, July*, 33–42. ACM.

Chan, C.; Jiayang, C.; Yim, Y.; Deng, Z.; Fan, W.; Li, H.; Liu, X.; Zhang, H.; Wang, W.; and Song, Y. 2024. NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA*, 4211–4241. Association for Computational Linguistics.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv:2507.06261*.

de Carvalho, G. A. L.; Igeri, S. B.; Healey, J.; Bursztyn, V.; Demeter, D.; and Birnbaum, L. A. 2025. A Flash in the Pan: Better Prompting Strategies to Deploy Out-of-the-Box LLMs as Conversational Recommendation Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, 8385–8398.

Deng, Y.; Zhang, W.; Xu, W.; Lei, W.; Chua, T.; and Lam, W. 2023. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *ACM Trans. Inf. Syst.*, 41(3): 77:1–77:25.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Friedman, L.; Ahuja, S.; Allen, D.; Tan, Z.; Sidahmed, H.; Long, C.; Xie, J.; Schubiner, G.; Patel, A.; Lara, H.; et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv:2305.07961*.

Gandhi, K.; Fränken, J.; Gerstenberg, T.; and Goodman, N. D. 2023. Understanding Social Reasoning in Language Models with Language Models. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems New Orleans, LA, USA*, 13518–13529.

He, Z.; Xie, Z.; Steck, H.; Liang, D.; Jha, R.; Kallus, N.; and McAuley, J. 2025. Reindex-then-adapt: Improving large language models for conversational recommendation. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 866–875.

Huang, X.; Lian, J.; Lei, Y.; Yao, J.; Lian, D.; and Xie, X. 2025. Recommender ai agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems*, 43(4): 1–33.

Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.; Hu, Z.; Ullman, T. D.; Torralba, A.; Tenenbaum, J. B.; and Shu, T. 2024. MMToM-QA: Multimodal Theory of Mind Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand*, 16077–16102. Association for Computational Linguistics.

Jung, C.; Kim, D.; Jin, J.; Kim, J.; Seonwoo, Y.; Choi, Y.; Oh, A.; and Kim, H. 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, Miami, FL, USA*, 19794–19809. Association for Computational Linguistics.

Karanikolas, N.; Manga, E.; Samaridi, N.; Tousidou, E.; and Vassilakopoulos, M. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, 278–290.

Kavanagh, D. J.; Andrade, J.; and May, J. 2005. Imaginary relish and exquisite torture: the elaborated intrusion theory of desire. *Psychological review*, 112(2): 446.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R. L.; Kim, G.; Choi, Y.; and Sap, M. 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*, 14397–14413. Association for Computational Linguistics.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*.

Kosinski, M. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv:2302.02083.

Li, C.; Deng, Y.; Hu, H.; Kan, M.; and Li, H. 2025. ChatCRS: Incorporating External Knowledge and Goal Guidance for LLM-based Conversational Recommender Systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 295–312. Association for Computational Linguistics.

Li, R.; Kahou, S. E.; Schulz, H.; Michalski, V.; Charlin, L.; and Pal, C. 2018. Towards Deep Conversational Recommendations. In *"Proceedings of the thirty-first Annual Conference on Neural Information Processing Systems*, 9748–9758.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv:2412.19437*.

Malle, B. F.; and Knobe, J. 2001. The Distinction between Desire and Intention: A Folk-Conceptual Analysis. Intentions and Intentionality: Foundations of Social Cognition. BF Malle, LJ Moses and DA Baldwin.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

OpenAI. 2024a. GPT-4o. https://platform.openai.com/docs/models\#gpt-4o. Accessed: 2024-09.

OpenAI. 2024b. GPT-4o-mini. https://platform.openai.com/docs/models\#gpt-4o-mini. Accessed: 2024-09.

Phillips, A. T.; Wellman, H. M.; and Spelke, E. S. 2002. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85(1): 53–78.

Qin, P.; Huang, C.; Deng, Y.; Lei, W.; and Chua, T. 2024. Beyond Persuasion: Towards Conversational Recommender System with Credible Explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4264–4282. Association for Computational Linguistics.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J. M.; Zhou, J.; Sunaryo, A. S.; Lee, T. M. C.; Mihalcea, R.; and Huang, M. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 5986–6004. Association for Computational Linguistics.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2024. Towards Understanding Sycophancy in Language Models. In *The 12th International Conference on Learning Representations, Vienna, Austria*. OpenReview.net.

Shi, H.; Ye, S.; Fang, X.; Jin, C.; Isik, L.; Kuo, Y.-L.; and Shu, T. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, volume 39, 1510–1519.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*.

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 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, Singapore*, 10691–10706. Association for Computational Linguistics.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 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, Bangkok, Thailand,*, 8593–8623. Association for Computational Linguistics.

Yu, F.; Jiang, L.; Huang, S.; Wu, Z.; and Dai, X. 2025. PersuasiveToM: A Benchmark for Evaluating Machine Theory of Mind in Persuasive Dialogues. arXiv:2502.21017.

Zhang, Z.; Jin, C.; Jia, M. Y.; and Shu, T. 2025. AutoToM: Automated Bayesian Inverse Planning and Model Discovery for Open-ended Theory of Mind. arXiv:2502.15676.