The Essence of Contextual Understanding in Theory of Mind: A Study on Question Answering with Story Characters

Chulun Zhou[1]* **Qiujing Wang**[2]* **Mo Yu**[2]† **Xiaoqian Yue**[1] **Rui Lu**[1]
Jiangnan Li[2], **Yifan Zhou**[2] **Shunchi Zhang**[2] **Jie Zhou**[2] **Wai Lam**[1]†
[1] The Chinese University of Hong Kong
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{clzhou,wlam}@se.cuhk.edu.hk, {yuexiaoqian,luruihfbp}@link.cuhk.edu.hk
qw26@njit.edu, szhan256@jhu.edu,
{moyumyu,jiangnanli,geniuszhou,withtomzhou}@tencent.com

Abstract

Theory-of-Mind (ToM) is a fundamental psychological capability that allows humans to understand and interpret the mental states of others. Humans infer others' thoughts by integrating causal cues and indirect clues from broad contextual information, often derived from past interactions. In other words, human ToM heavily relies on the understanding about the backgrounds and life stories of others. Unfortunately, this aspect is largely overlooked in existing benchmarks for evaluating machines' ToM capabilities, due to their usage of short narratives without global context, especially personal background of characters. In this paper, we verify *the importance of comprehensive contextual understanding about personal backgrounds in ToM* and assess *the performance of LLMs in such complex scenarios*. To achieve this, we introduce CHARTOM-QA benchmark, comprising 1,035 ToM questions based on characters from classic novels. Our human study reveals a significant disparity in performance: the same group of educated participants performs dramatically better when they have read the novels compared to when they have not. In parallel, our experiments on state-of-the-art LLMs, including the very recent o1 and DeepSeek-R1 models, show that LLMs still perform notably worse than humans, despite that they have seen these stories during pre-training. This highlights the limitations of current LLMs in capturing the nuanced contextual information required for ToM reasoning.

1 Introduction

Theory of Mind (ToM) is a psychological term that refers to the process to understand oneself or others by ascribing mental states (typically including the dimensions of belief, intention, emotion, desire, etc) to them (Premack and Woodruff, 1978). For human, ToM holds an important position in daily social interactions, as it enables people to explain and predict the behavior of others with no direct access to their minds (Apperly, 2010).

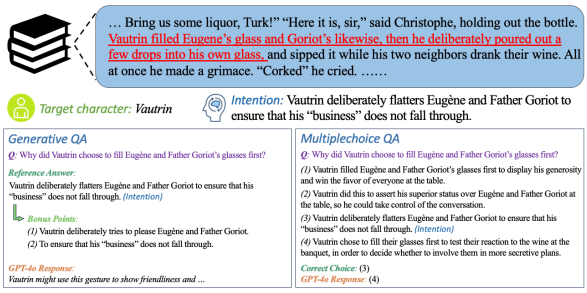

Figure 1: An example from CHARTOM-QA benchmark. LLMs make their responses given a question and corresponding story plot in generative and multichoice QA settings.

Recently, a range of emerging Large Language Models (LLMs) have shown remarkable performance in solving complex tasks and generating human-like language. To further enhance LLMs' capability to understand human, it is fascinating to explore the extent to which LLMs capture the ToM capabilities and ultimately to equip them with human-level ToM. To this end, researchers have developed numerous benchmarks to evaluate the ToM capabilities of LLMs (Grant et al., 2017; Nematzadeh et al., 2018; Sap et al., 2019; Le et al., 2019; Gandhi et al., 2023; Wu et al., 2023; Kim et al., 2023; Kosinski, 2023; Zhou et al., 2023; Shapira et al., 2024; Xu et al., 2024a; Chan et al., 2024; Wu et al., 2024; Chen et al., 2024). These benchmarks simulate the experimental design in psychological or cognitive research (Wimmer and Perner, 1983; Dyck et al., 2001), and contain elaborately crafted, story-based tasks that target ToM dimensions originally defined in psychology.

Despite their wide coverage of ToM dimensions, these existing benchmarks all largely overlook the

The work described in this paper is substantially supported from Tencent Rhino-Bird Research Fund (Project Code: TT2419882).

*Equal contribution. †: Co-corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22612–22631
July 27 - August 1, 2025

importance of **comprehensive contextual understanding** in human ToM. In daily lives, humans understand the minds of others based on long historical contexts of their personal backgrounds, rather than with only local circumstances on the spot. This is natural as human behaviors and mental states are also greatly influenced by global background factors of a person, such as social relationships, interactions with others, their personality and past experiences. However, the testing scenarios in existing ToM datasets are usually short of capturing these factors, as they are typically depicted with brief text pieces consisting of only superficial character actions and surrounding environment.

In this paper, we propose **CHARTOM-QA** benchmark ("CHAR" *stands for* "Character") that evaluates the capability of LLMs on understanding ToM of characters in famous novel books.[1] In CHARTOM-QA, the task takes the form of ToM-related question answering (QA) about characters within story plots. This setting naturally addresses the aforementioned challenges of most existing datasets due to the intrinsic features of story plots in novel books: **1)** diverse social scenarios; **2)** rich in complex social relationships and interactions; **3)** high relevance to the whole book storyline. Thus, it alleviates heavy reliance on pre-determined rules to generate testing scenarios and raises higher requirements for comprehensively understanding context when evaluating ToM capability of current LLMs.

For the dataset creation, we adopt an AI-assisted human annotation strategy that exploits publicly available book notes written by online reading app users. As users read, they can underline a specific text fragment and write their comments as book notes. These notes are often closely related to the selected fragments and their surrounding plots, containing valuable interpretation, analysis and thoughts of readers. We first collect a large number of user notes and their linked text from novel books. Then, as in most previous ToM assessments (Nematzadeh et al., 2018; Sap et al., 2019; Tracey et al., 2022; Gandhi et al., 2023; Ma et al., 2023b; Wu et al., 2023), we keep the notes reflecting certain ToM dimensions of appointed characters, including belief, intention, emotion and desire. Based on these notes, GPT-4o is leveraged to generate ToM descriptions of the four dimensions about given characters. Afterwards, these descriptions are used to construct ToM-related questions with answers. Finally, the experiments are conducted in both generative and multichoice QA settings, where models are asked to make their responses given a question and corresponding story plot, as shown in Figure 1. During the whole process, strict validation procedures are carried out by experts with strong literature background.

For evaluation, in generative QA, the qualities of responses are assessed based on the annotated answers. Particularly, traditional token-based metrics (*e.g.* Rouge (Lin, 2004)) and embedding-based metrics (*e.g.* cosine similarity between sentence embeddings) cannot give a fine-grained assessment that reflects how well a response matches the answer. Therefore, we design an evaluation protocol that mimics the process of grading papers. Specifically, we extract critical points from the answers, which are expected to be included in responses as bonus points. Then, GPT-4o is used as evaluator for assessment based on these bonus points. Besides, the evaluator also serves like a criticizer to point out defects existing in responses as penalty if there is any inappropriate statement. The bonus point coverage and penalty rate are used to comprehensively assess the qualities of model responses. In multichoice QA, we elaborately construct distraction choices and measure the model accuracy.

We experiment with several current popular LLMs in both generative and multichoice QA settings, where different lengths of plot window are exposed to models. The experimental results show that GPT-4o consistently outperforms other LLMs across the four ToM dimensions in both settings. Moreover, human studies are conducted using multichoice QA. The results reveal that (1) Humans who have read the book greatly outperform the state-of-the-art LLM, while those who have not perform only on par with it. (2) Human ToM largely benefits from familiarity with novels. (3) Humans achieve higher accuracy after reading longer contexts. In contrast, the performances of LLMs almost stay stable with different lengths of plot window. This indicates that the ToM capability of current LLMs still struggles at dealing with complex scenarios and capturing nuanced historical contextual information for robust ToM comprehension.

2 Related Work

With the advent of LLMs, many benchmarks have been developed to evaluate the ToM capability of

[1] Our data can be downloaded at Github https://github.com/Encyclomen/CharToM-QA and Huggingface https://huggingface.co/datasets/ZeroXeno/CharToM-QA.

LLMs by simulating the cognitive experiments originally in psychology (Grant et al., 2017; Nematzadeh et al., 2018; Sap et al., 2019; Le et al., 2019; Gandhi et al., 2023; Kosinski, 2023; Wu et al., 2023; Kim et al., 2023; Gandhi et al., 2023; Zhou et al., 2023; Yu et al., 2024; Shapira et al., 2024; Xu et al., 2024a; Chan et al., 2024; Wu et al., 2024; Chen et al., 2024). Nematzadeh et al. (2018) proposed ToMi for evaluating the capability of models to reason about beliefs. Based on ToMi, Sap et al. (2019) introduced Social IQA that tests social and emotional intelligence of models. Wu et al. (2023) constructed Hi-ToM to assess higher-order ToM capability of models that involves recursive reasoning on others' beliefs. Kim et al. (2023) build FANTOM to stress-test ToM within conversational contexts. Xu et al. (2024a) constructed OpenToM for assessing ToM with narrative stories containing explicit personality traits or preferences. Wu et al. (2024) created COKE consisting of cognitive chains that evaluate machine ToM capability to comprehend human activities. Chen et al. (2024) introduced ToMBench that encompasses 8 tasks and 31 abilities in social cognition.

Despite existing ToM benchmarks have covered most dimensions defined in psychology, the constructing procedures and testing scenarios of these datasets are still hindered by inherent limitations. **1)** The story generation procedures of these benchmarks rely heavily on pre-determined rules and templates (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2019; Wu et al., 2023), constraining it from simulating diverse scenarios in reality. This also increases the risk of incorporating predictable regularities and spurious correlations into datasets, bringing about so-called "Clever Hans" phenomenon (Lapuschkin et al., 2019). **2)** The testing scenarios in these datasets are usually short of complex social relationships and interactions so that the importance of comprehensive contextual understanding is almost ignored. Appendix C shows a more detailed comparison of CharToM-QA benchmark with existing benchmarks.

3 Problem Formulation of ToM in Global Contexts

Problem Formulation Similar to existing works on story understanding (Kočiskỳ et al., 2018; Pang et al., 2022; Xu et al., 2022; Yu et al., 2023), our task adopts a question-answering (QA) format. We denote the *global context* of a book as G, which in practice can be the list of all consecutive paragraphs of the book. Each QA pair (q, a) is associated with a *plot window* $W \subset G$, which is a book snippet. The task is then to answer the question according to W, with the necessary usage of the contextual knowledge from G[2]:

$$P(a|q, W, G). \quad (1)$$

ToM Dimensions Studied In this paper, we consider a set of ToM dimensions prevalently studied in previous work in Section 2, *Belief*, *Intention*, *Emotion* and *Desire*. The importance of these dimensions in daily lives have been discussed in (Apperly, 2010). We follow the standard definitions of these dimensions in (Apperly, 2010), with the detailed definitions provided in Appendix B.

4 CharToM-QA: Assessing the Contextual Aspect of ToM

We create the CharToM-QA benchmark consisting of questions about the minds of appointed characters in novel stories. The benchmark is designed to assess the ToM capability of LLMs in understanding characters within global contexts. The definitions of the four studied ToM dimensions are given in Appendix B. The questions and answers are annotated by four human experts in literature, with an AI-assisted annotation strategy that leverages massive publicly available book notes from online reading platforms.

The dataset construction undergoes three steps: (1) book notes collection & filtering; (2) key note extraction & paraphrasing; (3) question generation & verification. Our dataset supports evaluation in both generative and multichoice QA settings.

4.1 Book Notes Collection & Filtering

In recent online reading apps (*e.g.* Douban, Kindle), users can underline a text fragment and take notes while reading, as shown in Figure 2. These notes usually contain valuable interpretation, analysis and thoughts of readers about the selected fragment and corresponding story plot. Following (Wan et al., 2019; Yu et al., 2023), we use the user notes as a proxy to assist the human annotation of our dataset. We first download a set of public books available in the Gutenberg project and use

[2]It is worth noting that, like (Kočiskỳ et al., 2018; Pang et al., 2022), the problem allows to infer the thoughts of book characters with plots either after W (*retrospect*), or before W (*predict*). Both types of contexts naturally occur as part of the long-dependency reasoning in human ToM in daily life.

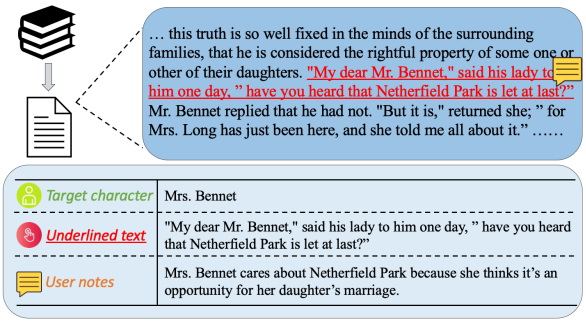

Figure 2: Illustration of notes from users on reading apps. A user underlines a text fragment (red) and writes his note about the character "Mrs. Bennet".

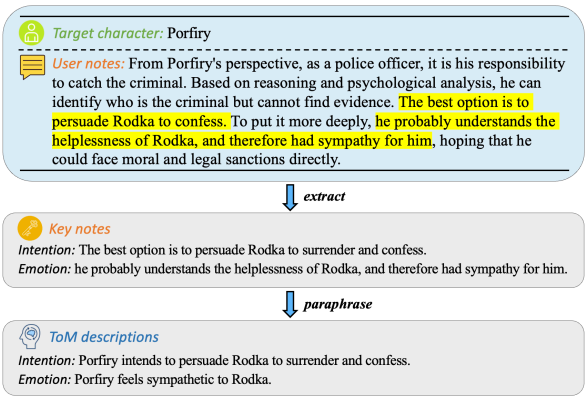

Figure 3: The extract-then-paraphrase approach to fetch ToM descriptions of a target character. The highlighted text pieces are the key notes extracted by annotators, which are then paraphrased into complete statements.

their Chinese-translated versions with valid usage licenses. Then, we collect a large number of user notes and their linked text fragment from the Internet with a list of main characters appearing in each book. The books with top 20 most user notes are chosen, as listed in Appendix A.

Next, we filter the notes and keep those that can express any one of the four ToM dimensions of characters within story plots. The prompt used for note filtering is detailed in Appendix F. To ensure the validity of these filtered notes, we also manually check their correctness. Specifically, we sample 100 notes from four books with two annotators who are familiar with the books. For each note, the annotators are required to judge whether it appropriately describes the ToM of the appointed character. The results show that 1) 80% notes are judged as accurate; 2) 13% of the notes are open to subjective interpretation by different readers, yet they remain consistent with the book's content and cannot be falsified; 3) only 7% notes are inappropriate, which demonstrates the overall validity of our filtered notes. Overall, 93% of notes can be treated as accurate.

4.2 Construction of Generative QA Task

Key Note Extraction & Paraphrasing (*i.e.*, Answer Generation). After collection and filtering, we have obtained a pool of highly ToM-related user notes. From these notes, we aim to fetch concrete ToM descriptions of characters within story plots. Specifically, we take a two-stage approach consisting of extraction and paraphrasing operations, as the case in Figure 3.

At the extraction stage, the annotators are instructed to extract the critical part of a note that explicitly reflects one of the four ToM dimensions of a target character, which we call as "*key note*". The basic principle of this stage is that annotators should only consider the literal meaning of a note with no additional inference. We elaborately craft examples for every ToM dimension and write a detailed manual to guide annotators. An annotator is only considered qualified after passing a labeling trial. They are given necessary information linked with a user note, including the target character, the underlined text and its surrounding context. After extraction, the key notes are kept for each ToM dimension. Nevertheless, the extracted key notes are often in a form of fragmented expressions that cannot be immediately used as complete ToM descriptions. Thus, a paraphrasing stage is required to convert fragmented parts into complete statements. For each note, we prompt GPT-4o to conduct paraphrasing based on key notes. The details of our used prompt template are presented in Appendix G. Annotators also check the paraphrased descriptions and make minor edit if necessary.

Although it is feasible for us to merely ask annotators or GPT-4o to directly produce ToM descriptions using the user notes, the advantages of such extract-then-paraphrase operation lie in the following two aspects. First, it relieves the burden of annotators to write complete descriptions totally from scratch. Second, GPT-4o can be more focused on the critical parts of notes without being distracted by other trivial details, reducing the possibility to produce unwanted content.

Question Generation & Verification. In this step, we aim to construct ToM-related questions using those ToM descriptions. Figure 4 depicts the process of question generation. GPT-4o serves as

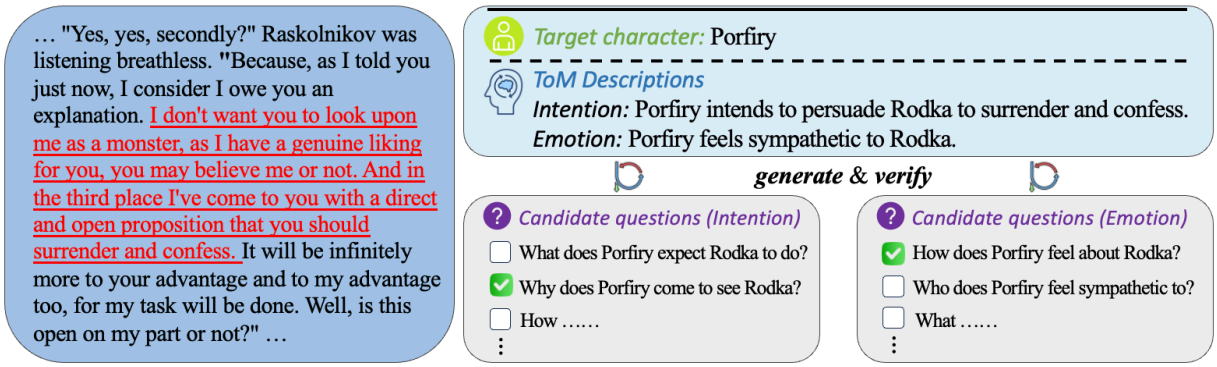

Figure 4: Given the corresponding story plot, GPT-4o is used to generate candidate questions about the target character using ToM descriptions. For each dimension, at most 4 candidates are kept after verification. Then, annotators manually choose the best one from these candidates.

a question generator to produce a set of candidate questions asking about target characters. During generation, along with the ToM description, we also provide GPT-4o with corresponding underlined text and its surrounding context of current story plot. For each ToM description, GPT-4o is asked to generate several candidate questions that adhere to the following requirements:

- The ToM description of target character should directly be the fluent and logically correct answer to the question.
- The question must be closely related to current story plot.
- The question should focus on core content rather than trivial details in the plot.

Subsequently, a verification procedure is involved to preliminarily filter out inappropriate candidates that violate the above requirements. We find that it is better not to let GPT-4o be aware of the story plot. This is because GPT-4o tends to build trivial connections between question and answer using any possible detail in the story regardless of importance, preserving many low-quality candidates. The prompt templates used for generation and verification are given in Appendix H. In this way, we keep at most 4 candidate questions for each ToM description. Then, annotators are responsible for choosing the best one out of these candidates, where slight modifications by human are allowed during examination.

4.3 Construction of Multichoice QA Task

Besides the aforementioned Generative QA setting, we support multichoice QA evaluation with elaborately constructed distraction choices for each question. Given the question-answer pair within a story plot, GPT-4o is used to produce distraction choices that are seemingly plausible but actually unreasonable. Appendix J provides the details of constructing such difficult distraction choices. Finally, a question is paired with four choices including the correct answer and three incorrect ones, as shown in the Appendix M.

5 Evaluation Protocols for Generative QA

Until now, we have acquired ToM-related question-answer pairs along with their linked story plots, which can be used to conduct evaluation in a generative QA setting. Models are given story plots and asked questions to make responses whose qualities are assessed using corresponding answers as reference. Conventionally, token-based metrics (*e.g.* Rouge (Lin, 2004)) and embedding-based metrics (*e.g.* cosine similarity between sentence embeddings) are mainly used for assessing generative QA tasks. However, these two types of metrics have their inherent limitations. For token-based metrics, they cannot deal with the cases where **1)** same meanings are expressed with different tokens; **2)** non-critical tokens dominate in a sentence. Meanwhile, embedding-based metrics only coarsely measure the semantic similarity but cannot explicitly indicate the pros and cons of a response.

Therefore, inspired by the process of grading papers, we design an evaluation protocol that inspects the bonus points and penalty of a model response. Specifically, for each question, we extract the critical points of its reference answer with the assistance of GPT-4o, which are expected to be included as bonus points in a response. The core requirements for bonus point extraction are: **1)** Bonus points must be derived from reference answers with no hallucination; **2)** Different bonus points should orient to different aspects of an answer. The statistics about the bonus points of questions in each

dimension is given in Appendix D. During evaluation, a GPT-4o evaluator measures the coverage of bonus points as an indicator of response quality. Moreover, it also criticizes the defects in a response as penalty if there is any inappropriate statement. In this way, both bonus point coverage and penalty rate are used as a comprehensive assessment of response quality. The results in Appendix E also demonstrate that our designed evaluation protocol is more correlated with human judgments than conventional metrics. Particularly, the prompts for extracting bonus points and conducting such evaluation protocol are detailed in Appendices I and L, respectively. Appendix M gives illustrative cases of such evaluation protocol.

6 Experiments

6.1 Setup

We conduct experiments on our generative and multichoice QA tasks in English and Chinese. Models are instructed to respond with vanilla prompts as in Appendix K. To investigate the capability of the model to exploit contextual information for ToM comprehension, we vary the context lengths of story plots exposed to models.

For human study, we recruit 8 native Chinese-speaking graduate students to do multichoice questions on a subset of books and use their results for human-LLMs comparison. Because it is almost infeasible for a person to write answers to all questions, we only acquire human baselines in multichoice QA on a subset, as in (Thai et al., 2022).

Evaluated LLMs In our experiments, we totally evaluate six current popular LLMs, including GPT-4o, GPT-3.5-Turbo-1106, Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen2-7B-Instruct (Yang et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and InternLM2-7B-Chat (Cai et al., 2024). We strictly call official APIs or download model weights from Huggingface[3] repositories with no violation of their terms.

Metrics For generative QA, as discussed in Section 5, we exploit GPT-4o as evaluator to assess the qualities of model responses using our proposed evaluation protocol. In our protocol, Bonus Point Coverage (BPC) and Penalty Rate (PR) assess response quality from different perspectives. BPC is calculated as the percentage of bonus points that are judged as being included in responses. PR is the rate of model responses that are judged as defective. Note that a response that includes all bonus points can still contain defects. During evaluation, as a necessary constraint, LLMs are required to produce responses of roughly the same length as corresponding answers. Otherwise, in extreme cases, LLMs could produce arbitrarily long responses to cover as many bonus points as possible. For multichoice QA, we just use the standard accuracy.

[3]https://huggingface.co/models

6.2 Main Results with LLMs

In this section, we show the performance of LLMs on generative QA and multichoice QA of our CHARTOM-QA benchmark.

Generative QA. For generative QA, LLMs are evaluated with vanilla prompting in English and Chinese. Table 1 gives the results in terms of BPC and PR. For fair comparison, LLMs are required to produce responses whose token numbers are no more than 1.5 times longer than those of answers. It can be seen that GPT-4o achieves the best performances across all dimensions in both English and Chinese. But it only covers roughly 35-50% (at most 58.5%) bonus points in every dimension. Meanwhile, PRs of all models are mostly at very high levels, reflecting the non-negligible existence of defects in model responses. After inspection, we find that some models tend to produce excessively long responses that are truncated due to the length constraint. Besides, producing longer responses would increase the risk of including inappropriate statements, which also exacerbates the phenomenon of high PR. Moreover, longer context exposed to models mostly reduce PRs of these models but fail to bring consistent improvements in term of BPC, which will be further discussed in Section 6.4. These results indicate the difficulty of our benchmark and there is still large room of improvement for current LLMs.

Multichoice QA. For multichoice QA, Table 2 gives the performances of all models in term of accuracy. It can be seen that GPT-4o also outperforms other models in both English and Chinese experiments. Similarly, the impact of context lengths is still inconsistent. These outcomes basically echo the results in generative QA experiments.

6.3 Human Performance

In this section, we conduct human study to explore how humans perform on our task, and to assess the role of long-context dependency in human ToM

Bonus Point Coverage% (English) ↑												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**41.4**	**43.4**	**44.4**	**34.3**	**35.7**	**34.6**	**34.0**	**37.5**	**36.1**	**51.9**	**50.0**	**52.8**
GPT-3.5-Turbo-1106	36.0	38.6	35.4	29.4	30.5	31.3	28.1	31.0	32.0	47.2	43.9	42.5
Llama-3.1-8B-Instruct	30.2	37.0	37.9	24.5	25.3	30.0	28.7	30.3	31.0	35.8	38.7	40.1
Qwen2-7B-Instruct	42.1	40.2	40.8	26.4	27.8	31.1	26.0	31.4	32.7	40.1	47.2	39.6
Mistral-7B-Instruct-v0.3	33.4	38.9	34.1	23.2	24.3	27.5	25.4	26.4	26.4	38.7	37.3	38.7
InternLM2-7B-Chat	25.4	32.8	28.6	17.7	28.1	24.0	21.9	27.4	26.3	34.4	34.9	35.4
Penalty Rate% (English) ↓												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**65.7**	**50.7**	**51.7**	**70.0**	**59.1**	**60.0**	**69.1**	**61.4**	**61.1**	**60.3**	**52.3**	**47.7**
GPT-3.5-Turbo-1106	75.6	66.7	69.7	79.1	72.3	73.2	81.9	74.9	77.1	66.9	71.5	70.2
Llama-3.1-8B-Instruct	88.1	85.6	83.6	91.4	77.7	84.1	84.9	81.9	82.3	83.4	80.1	82.1
Qwen2-7B-Instruct	83.6	79.1	80.6	85.0	83.2	85.0	89.0	84.9	80.1	84.1	78.8	79.5
Mistral-7B-Instruct-v0.3	81.6	72.6	75.1	85.0	83.2	80.9	89.4	84.2	83.2	80.1	83.4	75.5
InternLM2-7B-Chat	86.1	84.6	86.6	94.1	82.7	84.5	92.2	86.8	85.3	90.1	84.1	86.1
Bonus Point Coverage% (Chinese) ↑												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**50.2**	**51.4**	**51.1**	**38.4**	**42.5**	**40.9**	**42.6**	**41.6**	**43.7**	**59.4**	**58.5**	**61.3**
GPT-3.5-Turbo-1106	37.3	37.3	36.7	25.9	24.3	22.6	32.6	36.7	35.4	45.3	41.5	45.3
Llama-3.1-8B-Instruct	33.4	34.4	36.0	20.7	21.3	22.9	31.7	30.4	32.1	35.8	44.3	40.1
Qwen2-7B-Instruct	35.7	40.5	39.9	24.5	25.1	29.2	37.3	35.6	36.3	48.6	50.5	46.2
Mistral-7B-Instruct-v0.3	30.5	32.2	30.2	13.1	21.3	18.8	25.7	27.3	27.3	33.5	36.3	37.7
InternLM2-7B-Chat	35.0	35.4	32.8	26.7	26.2	22.3	22.0	20.6	19.9	38.2	37.3	37.3
Penalty Rate% (Chinese) ↓												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**48.3**	**43.3**	**35.3**	**58.2**	**41.4**	**45.0**	**45.8**	**40.6**	**39.5**	**43.7**	**33.8**	**33.1**
GPT-3.5-Turbo-1106	56.2	52.7	49.3	70.0	58.6	58.6	57.9	48.8	49.7	51.0	46.4	44.4
Llama-3.1-8B-Instruct	63.7	54.2	55.7	78.2	68.2	65.9	56.6	51.4	52.2	66.2	53.0	58.3
Qwen2-7B-Instruct	63.2	57.7	52.2	72.7	60.0	61.4	52.9	44.7	44.7	54.3	47.7	43.0
Mistral-7B-Instruct-v0.3	71.6	64.2	61.2	82.3	73.6	68.2	65.0	56.8	57.5	70.2	64.2	60.9
InternLM2-7B-Chat	68.7	62.7	61.7	74.1	69.1	64.1	57.5	54.4	52.3	57.0	49.0	45.7

Table 1: Generative QA performances of LLMs in terms of bonus point coverage and penalty rate. For each ToM dimension, the context lengths of story plots exposed to LLMs are set to 0, 1k and 2k tokens.

when understanding fictional characters. Our study addresses the following research questions:

- *RQ1: What is the performance gap between humans with and without knowledge on the historical contexts of the characters?* This question explores the key hypothesis of this paper regarding the long-dependency nature of human ToM. To investigate this, we divided the human annotators into two groups: those who have read the books and are familiar with the appointed characters (***w/.* history**) and those who have not read the books (***w/o.* history**). The performance gap between the two groups highlights the importance of long-context dependency in human ToM.
- *RQ2: Can humans outperform LLMs?* We compare the performance of the two human groups described earlier with GPT-4o on the same samples, to understand how human ToM can help to better solve our task.
- *RQ3: Can human performance benefit from longer input windows?* To investigate this, each annotator is initially asked to answer a question using a limited plot window (**c=0**). Then, they are provided with extended plot window and decide whether to revise their choices (**c=2k**).

To enable direct comparison with LLMs, we use the multichoice QA setting and recruit 8 educated participants, all of whom are either PhD candidates or hold PhD degrees across various disciplines. We sampled 150 questions from 5 books, ensuring that each question is assigned to one person who had read the book and one who had not. Each participant is given an *equal number of questions about books they had read and had not*. Combined with the two variations of plot windows, the participants totally make choices on 600 questions.

The result of this study is presented in Table 3, from which we draw the following conclusions:

- *Answer to RQ1:* Human ToM is largely impacted by familiarity with the target characters. A gap of ∼21% is observed between the two groups.
- *Answer to RQ2:* Humans who had read the book

Accuracy% (English)												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**58.7**	**56.7**	**54.7**	**52.3**	**53.6**	**51.4**	**53.1**	**51.0**	**51.8**	**56.3**	**56.6**	**55.0**
GPT-3.5-Turbo-1106	48.3	42.3	41.8	41.4	44.1	41.8	46.2	47.7	50.1	46.4	49.3	47.0
Llama-3.1-8B-Instruct	38.8	34.8	33.8	39.1	35.9	37.3	47.5	47.7	47.3	40.4	45.7	42.4
Qwen2-7B-Instruct	44.3	41.3	36.8	38.2	37.3	36.8	46.0	46.7	47.1	44.4	39.1	41.7
Mistral-7B-Instruct-v0.3	38.8	36.8	38.8	36.4	35.5	33.6	41.7	43.4	45.6	41.1	43.7	43.7
InternLM2-7B-Chat	39.8	34.3	35.3	30.0	36.8	35.0	41.0	43.4	45.1	45.0	44.4	40.4
Accuracy% (Chinese)												
Model	Belief			Intention			Emotion			Desire		
	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k	c=0	c=1k	c=2k
GPT-4o	**52.2**	**52.2**	**54.7**	**50.0**	**47.3**	**50.5**	**51.8**	**52.1**	**53.8**	**54.3**	**51.7**	**53.6**
GPT-3.5-Turbo-1106	37.8	37.3	40.8	33.6	37.3	41.4	43.2	47.7	49.0	35.8	41.1	34.4
Llama-3.1-8B-Instruct	37.3	35.8	39.8	37.3	38.6	35.9	43.6	45.6	44.9	37.7	35.8	40.4
Qwen2-7B-Instruct	43.3	39.3	41.1	38.2	35.5	48.7	46.0	46.0	49.0	35.8	39.7	41.1
Mistral-7B-Instruct-v0.3	36.8	30.8	32.3	28.6	31.8	32.7	39.1	39.7	42.1	27.2	35.1	35.8
InternLM2-7B-Chat	38.3	40.3	41.8	37.3	31.4	33.2	44.7	47.3	50.5	41.1	39.7	40.4

Table 2: Multichoice QA performances of LLMs in terms of accuracy with vanilla prompting. For each ToM dimension, the context lengths of story plots exposed to LLMs are set to 0, 1k and 2k tokens.

Accuracy% (Chinese)										
	w/. history		*w/o.* history		GPT-4o		o1		DS-R1	
	c=0	c=2k	c=0	c=2k	c=0	c=2k	c=0	c=2k	c=0	c=2k
all	59.3	**62.0**	41.3	48.7	47.3	42.0	51.3	43.3	54.0	48.7
indir.	54.8	60.6	33.7	40.4	45.1	37.5	46.2	39.4	50.4	40.2

Table 3: Accuracy of multichoice QA (Chinese) by different participants familiar or unfamiliar with selected books. We compare with GPT-4o, o1 and DS-R1 on the same subset. *indir.* refers to indirect questions.

greatly outperform the state-of-the-art LLM, while those who had not perform only on par with it. It suggests that the key advantage of human ToM in our task lies in the ability to integrate the characters' global historial contexts.

- *Answer to RQ3:* Humans achieves better accuracy when provided with longer contexts. This conforms to the general expectation that longer contexts can 1) help people familiar with the book recall the exact position of the context within the book; 2) provide additional information for those unfamiliar with the book. In contrast, the LLM does not exhibit a similar pattern.

6.4 Analysis and Discussions

Can Chain-of-Thought (CoT) help? We further explore whether CoT could bring improvement of ToM comprehension. For this study, we use the recently released strong OpenAI o1 and DeepSeek-R1 (DS-R1) (Guo et al., 2025) models on the human study subset of 150 instances. As shown in Table 3, both o1 and DS-R1 outperform GPT-4o, showing the benefit from stronger reasoning ability. However, they still notably lag behind humans. As the CoT of DS-R1 is visible, we closely examine its generated thoughts on randomly selected 100 samples. We find that 76% of the thoughts initially recalled either the character's personal background or prior plot events that are relevant to the question. Among the remaining 24% of cases, three-fourths of questions are about emotions, which is as expected because emotions are more likely to be temporary mental states that rely less on global contexts. These findings further highlights the importance of comprehensive contextual understanding for ToM reasoning, a promising direction for enhancing future LLM.

LLMs Struggles at Exploiting Longer Plot Windows. While many studies have shown the superior capability of LLMs in handling long inputs, even up to 100k tokens (Dubey et al., 2024; Hurst et al., 2024), we find they fail to utilize longer plot windows in our tasks. In both generative and multichoice QA settings, the performances of LLMs almost remain stable as the plot window is enlarged.[4]

This failure can be attributed to two potential reasons. First, despite the ability to process long inputs, LLMs struggle with capturing dependencies between paragraphs and enhancing the understanding of individual paragraphs based on these dependencies (Xu et al., 2024b). Second, the LLMs may solve our task largely with their memory of the books and reviews, exhibiting the *Clever Hans* phenomenon (Shapira et al., 2024) – as the evaluated books and their reviews have been exposed to the LLMs during pre-training. It is likely that the LLMs can recall the associated plot details and

[4]It is worth noting that PRs steadily drop as LLMs read longer plot windows, showing that they do incorporate contextual information from story plots.

knowledge with both short and long windows. We leave this study of these reasons in future work.

LLMs Perform Poorly on Indirect Questions. To gain further insight on the LLM performance, we categorize our questions into two types, *direct* and *indirect* questions, for analysis. Direct question refers to one with the answer or its synonymous expressions explicitly appear in the given story plot. We manually labeled the 150 instances used in human study, resulting in 104 indirect questions, with the performance on this subset shown in Table 3.

As expected, both humans unfamiliar with the books and the LLMs drop dramatically on the subset of indirect questions. Furthermore, it is noteworthy that both o1 and DS-R1 models also perform worse on the indirect questions than on the full set. As indirect questions typically require broader global context of the whole stories rather than only superficial textual clues in the given plot windows, it suggests that the ToM capabilities of current LLMs are still limited especially in the aspect of comprehensive contextual understanding.

7 Conclusion

We introduce CHARTOM-QA benchmark that aims to evaluate machines' ToM capability in scenarios with complex social relationships and interactions. CHARTOM-QA focuses on the aspect of contextual understanding in comprehending ToM of characters in novels, which is largely ignored in existing benchmarks. The tasks take the forms of generative and multichoice QA. For generative QA, we design an evaluation protocol inspecting bonus points and penalty in model responses. Experimental results reveal that current advanced LLMs, including o1 and DS-R1, still lag behind humans and struggle at effectively exploiting context to comprehend ToM of characters. Overall, our CHARTOM-QA benchmark highlights the importance of comprehensive contextual understanding for advancing ToM reasoning in LLMs.

Limitations

Our proposed CHARTOM-QA benchmark is built from the notes written by readers when reading classic novels. For the creation, as we adopt an AI-assisted annotation strategy, the annotation results will be affected by the personal understanding from the annotators and GPT-4o. Thus, it requires the annotators to basically know the story of these books. Besides, the benchmark includes the ToM dimensions of belief, intention, emotion and desire, which are prevalently studied in previous work. However, there are still other ToM dimensions that can be investigated, such as knowledge and thoughts.

We also design an evaluation protocol that assesses model responses by BPC and PR, where GPT-4o is used as an evaluator. Such evaluation would cost much if the dataset is very large or the stories are excessively long. For those who cannot afford massively calling OpenAI APIs, other locally-deployed open-source LLMs can also be used to make assessments.

Ethical Consideration

The construction of CHARTOM-QA benchmark involves the usage of public books and notes written by reading app users. During the process of using these data, we strictly adhere to the guidelines in Internet Research Ethics. All the books are available in the Gutenberg project and their Chinese-translated version are properly used with valid license. For the notes, we collect and use them with necessary procedures to protect the privacy of online reading app users who wrote the notes. For annotators participating in the benchmark construction, they are adequately paid according to their working hours. In our human study, we do not collect any personally identifiable information of the participants. The participants are only asked to respond to multichoice questions and not required to provide any new written input.

Potential Risks

As our benchmark is derived from classic novels, the stories are essentially fictional and might have limitations of the time. Meanwhile, the notes written by users reflect their personal interpretation of the story, which may contain content expressing negative emotions. However, please note that the construction approaches and evaluation protocols in this paper can also be used in other type of tasks.

References

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 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, November 12-16, 2024*, pages 4211–4241. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15959–15983. Association for Computational Linguistics.

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

Murray J Dyck, Kara Ferguson, and Ian M Shochet. 2001. Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? *European child & adolescent psychiatry*, 10:105–116.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*. cognitivesciencesociety.org.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14397–14413. Association for Computational Linguistics.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 5871–5876. Association for Computational Linguistics.

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

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023a. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning, CoNLL 2023, Singapore, December 6-7, 2023*, pages 15–26. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023b. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2392–2400. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

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, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2257–2273. Association for Computational Linguistics.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10438–10451. Association for Computational Linguistics.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. Relic: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7500–7518. Association for Computational Linguistics.

Jennifer Tracey, Owen Rambow, Claire Cardie, Adam Dalton, Hoa Trang Dang, Mona T. Diab, Bonnie J. Dorr, Louise Guthrie, Magdalena Markowska, Smaranda Muresan, Vinodkumar Prabhakaran, Samira Shaikh, and Tomek Strzalkowski. 2022. Best: The belief and sentiment corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2460–2467. European Language Resources Association.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. COKE: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15984–16007. Association for Computational Linguistics.

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, Singapore, December 6-10, 2023*, pages 10691–10706. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024a. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8593–8623. Association for Computational Linguistics.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024b. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5822–5838, Bangkok, Thailand. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14784–14802.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, and Jie Zhou. 2024. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. In *Forty-first International Conference on Machine Learning*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How far are large language models from agents with theory-of-mind? *CoRR*, abs/2310.03051.

Appendix

A Book List

We keep the books with top 20 user notes for the construction of CHARTOM-QA benchmark. These books are publicly accessible and span different historical periods, genres, and topics. Table 4 lists the chosen books and the number of questions belonging to each book in the descending order. It is also notable that our dataset has included a number of unique target characters (150 in total) from these books. These characters are significantly diverse, crafted by different authors with varying literary styles. Particularly, these characters possess distinct personalities, encompassing various types and fundamentally different personal backgrounds. These elements could introduce diversity to our dataset, effectively mitigating potential biases related to literary style and topic.

Book	*#Question*
Madame Bovary	167
The Count of Monte-Cristo	101
Crime and Punishment	94
Of Human Bondage	88
Pride and Prejudice	82
Anna Karenina	79
War and Peace	53
Jane Eyre	49
Wuthering Heights	42
The Brothers Karamazov	37
Anne Of Green Gables	33
Little Women	32
The Idiot	30
Twenty Thousand Leagues under the Sea	29
Les Miserables	23
Notre-Dame de Paris	22
Oliver Twist	21
Father Goriot	19
Tess of the d'Urbervilles	19
The Red and the Black	15
Total	1,035

Table 4: The book list and the number of questions from each book.

B ToM Dimensions Studied

In this paper, we concern four ToM dimensions of characters within story plots including *Belief*, *Intention*, *Emotion* and *Desire*. The definitions of these dimensions are as follows

- **Belief**: Beliefs encompass both objective facts and subjective perceptions concerning the existence or truth of something.
- **Emotion**: Emotions are strong feelings deriving from one's circumstances, mood, or relationships with others. And emotions are variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure.
- **Intention**: Intentions are blueprints that steer actions, encompassing both future plans and the motivations driving current behaviour.
- **Desire**: Desires encompass both physical needs and psychological yearnings. Desires incline people toward action and fulfilling desires is pleasurable. Their fulfillment is normally experienced as pleasurable in contrast to the negative experience of failing to do so.

C Comparison of Benchmark Characteristics

In Table 5, we compare the characteristics of our CHARTOM-QA and other ToM benchmarks. First, our CHARTOM-QA provides both multichoice QA and generative QA tasks. Second, as the classic novel stories are directly used, CHARTOM-QA needs no additional story creation procedure, unlike most of other benchmarks. Meanwhile, it naturally possesses high-quality diverse social scenarios with complex social relationships and interactions due to the intrinsic features of human-written novels. Third, for the problem construction, most other benchmarks involve templates for generating questions or answers. Although this kind of template-based approach can produce a relatively large number of problems, it could bring shortcuts and spurious correlations into benchmarks. In contrast, we adopt an AI-assisted human annotation strategy elaborated in Section 4 without any predefined template. This simultaneously ensures the efficiency and quality of our annotation process.

Last and most importantly, in the aspect of character background, CHARTOM-QA is the only benchmark that incorporates character's previous experience depicted in story. Most of other ToM benchmarks merely provide depictions of character's current situation or some pre-defined background information such as conversation topics, preference items. As demonstrated by the human study described in Section 6.3, such personal background is necessary for comprehensive contextual understanding about characters to correctly answer ToM-related questions in CHARTOM-QA. Specifically, participants with similar educational levels who have read the book (and thus possess knowledge of long personal backgrounds) perform significantly better than those who have not read

Benchmarks *w/.* Template-based Problem Construction									
	Task Form	Story Creation			Problem Construction			Character Background	*#Eval_Q*
		Template	AI	Human	Template	AI	Human		
ToMi (Nematzadeh et al., 2018)	*MQA*	✓			✓			*N/A*	6,000
Social IQA (Sap et al., 2019)	*MQA*	–	–	–	✓		✓	*N/A*	2,224
Hi-ToM (Wu et al., 2023)	*MQA*	✓			✓			*N/A*	1,000
FanToM(Kim et al., 2023)	*MQA,GQA*	✓	✓		✓	✓		*N/A*	2,118
BigToM (Gandhi et al., 2023)	*MQA*	✓	✓		✓	✓		*N/A*	5,000
NegotiationToM (Chan et al., 2024)	*MQA*			✓	✓		✓	*pre-defined character's preference items*	13,800
OpenToM (Xu et al., 2024a)	*MQA*	✓	✓		✓		✓	*pre-defined personlity traits and preferences*	16,008
COKE* (Wu et al., 2024)	*NLG*	–	–	–	✓	✓	✓	*N/A*	-
Benchmarks *w/o.* Template-based Problem Construction									
	Task Form	Story Creation			Problem Construction			Character Background	*#Eval_Q*
		Template	AI	Human	Template	AI	Human		
FauxPas-EAI (Shapira et al., 2023)	*MQA,GQA*			✓			✓	*N/A*	80
ToMChallenges (Ma et al., 2023a)	*MQA,GQA*	✓		✓			✓	*N/A*	360
ToMBench (Chen et al., 2024)	*MQA*		✓	✓		✓	✓	*N/A*	2,860
CharToM-QA (Ours)	*MQA,GQA*			✓		✓	✓	*all of the character's previous experience depicted in story*	1,035

Table 5: Comparison among CharToM-QA and other ToM benchmarks in terms of task form, story generation, problem construction, character background and the number of evaluation questions ($\#Eval\_Q$). In the column of "Task Form", *MQA*, *GQA* and *NLG* refer to multichoice QA, generative QA and natural language generation tasks, respectively. "✓" denotes that a benchmark possesses corresponding properties. "*N/A*" means the story merely depicts character's current situation without extra character background. "-" in the $\#Eval\_Q$ column means the task form does not support the counting of questions equivalent to standard QA settings. *The COKE and Social IQA datasets are essentially about thoughts on input events/situations so the concept of *stories* does not apply.

	belief	intention	emotion	desire
$\#Question$	201	220	463	151
$\#Bonus\ Point$	311	367	700	212
$\#Q_{\#bp=1}$	108	106	267	95
$\#Q_{\#bp=2}$	77	84	159	51
$\#Q_{\#bp>2}$	16	30	37	5

Table 6: Statistics of questions and their bonus points in each ToM dimension.

	BPC(%)	RougeL(%)	BertScore(%)
GPT-4o	**39.8**	**21.5**	**56.6**
GPT-3.5-Turbo-1106	34.0	19.6	51.6
Llama-3.1-8B-Instruct	33.5	16.2	50.8
Qwen2-7B-Instruct	34.9	15.9	50.5
Mistral-7B-Instruct-v0.3	29.9	19.0	49.7
InternLM2-7B-Chat	27.6	17.7	49.7
*Human Correlation**	*0.92*	*0.72*	*0.81*

Table 7: Generative QA performances of LLMs in terms of BPC, RougeL and BertScore with the context length being 2k. "*": The human correlation of these metrics are calculated on the subset used in our human study.

when given the same contexts. It shows that our CharToM-QA highlights the importance of comprehensive contextual understanding for ToM reasoning, an aspect overlooked by previous benchmarks.

D Statistics of Questions & Bonus Points

CharToM-QA consists of ToM-related questions about characters in the dimensions of belief, intention, emotion and desire. For the evaluation of generative QA, the bonus points for each question are extracted for subsequent assessment. In Table 6, we present the statistics of questions and bonus points in each dimension. Since most questions have less than 3 bonus points, we group them into "$Q_{\#bp=1}$", "$Q_{\#bp=2}$" and "$Q_{\#bp>2}$", which refers to questions with one, two and more than two bonus points.

E Correlations of Evaluation Metrics with Humans on Generative QA

For Generative QA, besides the designed LLM-based evaluation protocol, we also use conventional token-based (RougeL) and embedding-based (BertScore) metrics for assessing the quality of model responses. In Table 7, we report the experimental results in terms of BPC, RougeL and BertScore when the context length is 2k. The reported values are the weighted averages of corresponding metrics over the four ToM dimensions. From the table, our LLM-based metric gives similar relative rankings among different LLMs as conventional metrics. Moreover, on the subset used in our human study, we let human experts assess model responses by giving scores ranging from 0 (nonsense) to 1.0 (perfect). The Pearson Correlations of human judgments with BPC, RougeL and BertScore are 0.92, 0.72 and 0.81, respectively. These results further justify the reliability of our designed evaluation protocol.

F Prompt for Book Note Filtering

For book notes filtering, we incorporate the above definitions into the prompt and keep those notes that can reflect any one of the four ToM dimension

Assuming you are an expert in psychology and literary. Reading app users will take notes while reading novels, which may include the analysis about characters. Please use theory of mind to determine whether the user notes about the novel plots explicitly describe the ${*dimension*} of the target characters in the plot.

[Definition of ${*dimension*}]: ${*definition*}
[User Note]: ${*user_note*}
[Target Character]: ${*target_character*}

If "Yes", describe what the ${*dimension*} of ${*target_character*} is; If not, give your reason. Note that (1) We need the ${*dimension*} of the characters in the novel expressed by readers, not that of the reader. (2) Only consider the content about the ${*dimension*} explicitly described in the note. Do not guess or make any unnecessary prediction.

Please answer in the following format:
"Yes / No"
"${*dimension*} of ${*target_character*} / Reason"

Table 8: The prompt for book note filtering.

of characters. The prompt for filtering book notes is presented in Table 8.

G Prompt for ToM Paraphrasing

In Section 4.2, after the key note is extracted, we use GPT-4o to paraphrase it into a complete description about the ToM of the target character within the story plot. The prompt for paraphrasing is presented in Table 9.

H Prompt for Question Generation & Verification

In Section 4.2, we leverage GPT-4o to generate a set of candidate questions, to which the ToM description of the target character is the answer. Then, at most 4 candidate questions are kept for each ToM description after verification. The prompts for question generation and verification are presented in Table 10 and Table 11, respectively.

I Prompt for Extracting Bonus Points

In CHARTOM-QA benchmark, for each question, we provide the bonus points that are expected to be included in model responses. The prompt for extracting bonus points is presented in Table 12.

Assuming you are an expert in psychology and literary. Reading app users will take notes while reading novels, which may include the analysis about characters. The critical fragments of a user note that can reflect the ${*dimension*} of the target character in the novel plot are extracted as key note. Please use the theory of mind to paraphrase the fragmented key note into a complete ToM description about the ${*dimension*} of ${*target_character*}. You are also given the original note and its corresponding underlined text.

[Underlined Text]: ${*underlined_text*}
[User Note]: ${*user_note*}
[Key Note]: ${*key_note*}
[Target Character]: ${*target_character*}

Note that (1) We need the ${*dimension*} of the target character in the novel, not that of the reader. (2) Only consider the content about the ${*dimension*} explicitly described in the note. Do not guess or make any unnecessary prediction. (3) Do not always copy the content of the key note. You are encouraged to make necessary polishing without changing the original meaning. (4) Straightforwardly describe ${*dimension*} and avoid providing additional explanations.

The ToM description about the ${*dimension*} of ${*target_character*} is:

Table 9: The prompt for key note paraphrasing.

J Construction of Difficult Distraction Choices for Multichoice QA

Furthermore, to increase task difficulty, we also exploit the model responses obtained in the evaluation of generative QA. Concretely, the defects of those responses pointed out by the evaluator are used as potential misleading directions. During the construction of distraction choices, GPT-4o can selectively use these misleading directions for reference to produce more difficult content. In this way, our multichoice QA task becomes much more challenging. Table 13 gives the prompt for the construction of such distraction choices.

K Prompt for Responding to Questions

Table 14 and Table 15 give the prompts for responding to the questions in generative and multichoice

Assuming you are an expert in psychology and literary. Please raise questions related to the ${*dimension*} of the characters in the novel based on [Story Plot] and [Underlined Text] from the book ${*book_name*}.

[Story Plot]: ${*story_plot*}
[Underlined Text]: ${*underlined_text*}
[Target Character]: ${*target_character*}
[ToM Description]: ${*ToM_description*}

Note that (1) [ToM Description] should directly be the fluent and logically correct answer to your question. (2) The questions should be closely related to current story plot. (3) The question should focus on core content rather than trivial details in the plot. (4) Only raise one question.

The question you raise that meets the above requirements is:

Table 10: The prompt for question generation.

Assuming you are an expert in psychology and literary. You are reading the book ${*book_name*}. There are a piece of description about the ${*dimension*} of ${*target_character*} and a question related to the book story. Please judge whether [Candidate Question] is an appropriate question whose direct answer is [ToM description].

[Candidate Question]: ${*candidate_question*}
[ToM description]: ${*ToM_description*}

An appropriate question should satisfy the following requirements: (1) [Candidate Question] should not be too broad. (2) [ToM description] should be a fluent and logically correct answer to [Candidate Question].

Please answer in the following format:
"Appropriate / Inappropriate"
"Reason"

Table 11: The prompt for question verification.

QA during experiments, respectively.

L Prompt for Generative QA Evaluation

For generative QA evaluation, we leverage GPT-4o as an evaluator to assess the quality of model responses based on the bonus points extracted from the reference answer. Table 16 presents the prompt for inspecting which bonus point is included in a response. Besides, the evaluator also finds defects in a response as penalty. The bonus point coverage and penalty rate are the two metrics measuring the quality of a response. The prompt for finding defects in a response is presented in Table 17.

M Illustrative Cases

Table 18 gives a case of question about Vautrin's intention from the book "Father Goriot", accompanied with the responses of GPT-4o. For the generative QA, the response of GPT-4o covers one of the two bonus points. The penalty explains the defects existing in the response. For the multi-choice QA, the model selects the wrong candidate choice. Table 19 gives another case about Monte Cristo's belief from the book "The Count of Monte Cristo". Note that this question involves the second-order belief of the character, which can evaluate the second-order ToM capability of models. We can see that GPT-4o covers both bonus points in generative QA and select the correct choice in multichoice QA.

Assuming you are an expert in psychology and literary. Based on your profound understanding of the story in the book ${*book_name*}, extract the critical bonus points from the [Answer] for the [Question]. The bonus points are used to comprehensively assess the quality of responses from other respondents. Below are the [Story Plot], [Question] about the character ${*target_character*} and the [Reference Answer].

[Story Plot]: ${*story_plot*}
[Question]: ${*question*}
[Reference Answer]: ${*reference_answer*}

Note that (1) Bonus points must be derived from the reference answer without hallucination. Bonus points are the indispensable points in the reference answer, but should not include any unnecessary explanatory content. (2) Different bonus points should orient to different aspects of the reference answer. (3) The semantic granularity of the bonus points should not be too fine. Each bonus point must be a sentence that constitutes a complete semantic unit. Semantically coherent content in the reference answer should not be broken down into multiple bonus points.

Please use '\n' to separate multiple bonus points line by line and every point should conform to the following format:
[Bonus Point]: "Content of bonus point"

Table 12: The prompt for extracting bonus points from reference answers.

Assuming you are an expert in psychology and literary. Please construct appropriate multichoice questions for the given [Story Plot] from the the book ${*book_name*}.

[Story Plot]: ${*story_plot*}
[Question]: ${*question*}
[Reference Answer]: ${*reference_answer*}
[Misleadings]: ${*misleadings*}

Note that (1) Ensure that the reference answer is the only correct answer to the question. (2) The three distraction candidate choices should be seemingly plausible but actually unreasonable. (3) You can refer to the reference answer and misleadings to construct distraction choices. (4) Do not repeat the same content in different distraction choices.

Please provide distraction candidate choices that meet the above requirements and conform to the following format:
[Distraction Choice 1]: (Choice Content)
[Distraction Choice 2]: (Choice Content)
[Distraction Choice 3]: (Choice Content)

Table 13: The prompt for constructing difficult distraction choices in multichoice QA.

Assuming you are an expert in psychology and literary. Based on your profound understanding of the story in the book ${*book_name*}, Please answer the [Question]. The following items give the [Story Plot] and the [Question].

[Story Plot]: ${*story_plot*}
[Question]: ${*question*}

Note that you should respond to the question using only one concise sentence, with around ${*length_of_answer*} tokens.
Your response is:

Table 14: The prompt for responding to questions in generative QA.

Assuming you are an expert in psychology and literary. Based on your profound understanding of the story in the book ${*book_name*}, Please answer the [Question]. The following items give the [Story Plot], the [Question] and the four candidates from [Candidate Choices]..

[Story Plot]: ${*story_plot*}
[Question]: ${*question*}
[Candidate Choices]:
(1). ${*cand_choice_1*}
(2). ${*cand_choice_2*}
(3). ${*cand_choice_3*}
(4). ${*cand_choice_4*}

Note that
1. You should only choose one candidate from (1),(2),(3),(4).
2. Only output the index of your chosen candidate.
3. Do not include any other unnecessary content or symbol.

Your choice is:

Table 15: The prompt for responding to questions in multichoice QA.

Assuming you are an expert in psychology and literary. Based on your profound understanding of the story in the book ${*book_name*}, assess the quality of the response to the given [Question] according to the provided [Reference Answer] and corresponding [Bonus Points]. You are supposed to indicate which bonus points are explicitly or implicitly included in the response.

[Question]: ${*question*}
[Reference Answer]: ${*reference_answer*}
[Bonus Points]: ${*bonus_points*}
[Response]: ${*response*}

Note that (1) The reference answer and bonus points must be the core basis of assessment. (2) The response may be similar to a certain bonus point but with different wording. As long as it expresses a similar meaning to this bonus point with no error, the point can be considered to be validly included. (3) If the response does not include any bonus point, just output '[Included Bonus Points]: None'

Please conform to the following format:
[Included Bonus Points]: "<Indices of the included bonus points separated by commas, such as '1,2'>"

Table 16: The prompt for inspecting which bonus points are included in a response.

Assuming you are an expert in psychology and literary. Based on your profound understanding of the story in the book ${*book_name*}, according to the [Story Plot] and [Reference Answer], you are supposed to detect whether there is any defect existing in the [Response] to the [Question].

[Story Plot]: ${*story_plot*}
[Question]: ${*question*}
[Reference Answer]: ${*reference_answer*}
[Response]: ${*response*}

Note that (1) The reference answer and story plot must be the core basis for detecting defects in the response. (2) Defects refers to factual or logical errors in the response. (3) Do not reluctantly find defects when the response is totally reasonable. If there is no defect, just output '[Defect]: None'

Please conform to the following format:
[Defects]: "Content of defects in the response"

Table 17: The prompt for pointing out defects in a response.

Plot Window

..... Bring us some liquor, Turk!" "Here it is, sir," said Christophe, holding out the bottle. Vautrin filled Eugene's glass and Goriot's likewise, then he deliberately poured out a few drops into his own glass, and sipped it while his two neighbors drank their wine. All at once he made a grimace. "Corked" he cried. .....

Generative QA

Question: Why did Vautrin choose to fill Eugène and Father Goriot's glasses first?

Reference Answer: Vautrin deliberately flatters Eugène and Father Goriot to ensure that his "business" does not fall through.

Bonus Points:

(1) Vautrin deliberately tries to please Eugène and Father Goriot. ✔

(2) To ensure that his "business" does not fall through. ✗

GPT-4o Response: Vautrin might use this gesture to show friendliness and generosity, thereby fostering closer relationships and easing Eugène's distant attitude toward him, all while concealing his true intentions.

Penalty: The response mentioning "easing Eugène's distant attitude toward him" is neither explicitly stated nor implied in the excerpt. Vautrin's actions are more likely driven by a need to maintain control of the situation and conceal his true intentions, rather than simply addressing Eugène's cold demeanor.

Multichoice QA

Question:Why does Vautrin choose to fill Eugène and Father Goriot's glasses first?

Candidate Choices:

(1) Vautrin filled Eugène and Father Goriot's glasses first to display his generosity and win the favor of everyone at the table.

(2) Vautrin did this to assert his superior status over Eugène and Father Goriot at the table, so he could take control of the conversation.

(3) Vautrin deliberately flatters Eugène and Father Goriot to ensure that his "business" does not fall through.

(4) Vautrin chose to fill their glasses first to test their reaction to the wine at the banquet, in order to decide whether to involve them in more secretive plans.

Correct Choice: (3)

GPT-4o Response: (4)

Table 18: Case 1 from "Father Goriot". ✔ means the response of GPT-4o covers the bonus point. ✗ means the corresponding bonus point is not covered. The defects of response are also indicated in the penalty. For multichoice QA, the model selects the wrong choice.

Plot Window

..... It was M. de Villefort, and not the prefect, who returned to the house of M. de Villefort. The procureur felt more at ease, although he had learned nothing really satisfactory, and, for the first time since the dinner-party at Auteuil, he slept soundly. .....

Generative QA

Question: What does the Count of Monte Cristo believe Mr. Villefort's attitude towards him is, and what actions does he expect Villefort to take?

Reference Answer: The Count of Monte Cristo knows that Mr. Villefort must harbor suspicions about him and will inquire everywhere for information about him.

Bonus Points:

(1) Christophe the Count of Monte Cristo knows that Mr. Villefort must have doubts about him. ✔

(2) Mr. Villefort will inquire about news concerning the Count of Monte Cristo everywhere. ✔

GPT-4o Response: The Count of Monte Cristo believed that Mr. Villefort harbored suspicion towards him and would continue to monitor and investigate his actions through various means.

Penalty: None

Multichoice QA

Question: What does the Count of Monte Cristo believe Mr. Villefort's attitude towards him is, and what actions does he expect Villefort to take?

Candidate Choices:

(1) The Count of Monte Cristo believes that Monsieur de Villefort holds a cautious attitude towards him and will take careful actions to observe his movements.

(2) The Count of Monte Cristo plans to take a series of actions to seek revenge on Monsieur de Villefort and his family, and therefore believes that Villefort has already realized his plan.

(3) The Count of Monte Cristo knows that Mr. Villefort must harbor suspicions about him and will inquire everywhere for information about him.

(4) As no specific threat is discovered, the Count of Monte Cristo believes that Monsieur de Villefort will not take any immediate action and will simply maintain a watchful stance for the time being.

Correct Choice: (3)

GPT-4o Response: (3)

Table 19: Case 2 from "The Count of Monte Cristo". ✔ means the response of GPT-4o covers the bonus point. The response of GPT-4o covers all bonus points with no penalty in generative QA. It also selects the correct choice in multichoice QA. Note that this question involves the second-order ToM capability.