XToM: Exploring the Multilingual Theory of Mind for Large Language Models

Chunkit Chan[♠*] **Yauwai Yim**[♠*] **Hongchuan Zeng**[§*] **Zhiying Zou**[§] **Xinyuan Cheng**[†]
Zhifan Sun[◇] **Zheye Deng**[♠] **Kawai Chung**[♠] **Yuzhuo Ao**[♠] **Yixiang Fan**[♠]
Cheng Jiayang[♠] **Ercong Nie**[†‡] **Ginny Y. Wong**[♣] **Helmut Schmid**[†‡]
Hinrich Schütze[†‡] **Simon See**[♣] **Yangqiu Song**[♠]
[♠]HKUST, Hong Kong [♣]NVIDIA AI Technology Center (NVAITC), USA
[§]Shanghai Jiao Tong University, China [◇]Technische Universtität Darmstadt, Germany
[†]LMU Munich, Germany [‡]Munich Center for Machine Learning, Germany
{ckchancc, yqsong}@cse.ust.hk

Abstract

Theory of Mind (ToM)—the ability to infer mental states in others—is pivotal for human social cognition. Existing evaluations of ToM in LLMs are largely limited to English, neglecting the linguistic diversity that shapes human cognition. This limitation raises a critical question: *can LLMs exhibit Multilingual Theory of Mind—the capacity to reason about mental states across diverse linguistic contexts?* To address this gap, we present XToM, a rigorously validated multilingual benchmark that evaluates ToM across five languages and incorporates diverse, contextually rich task scenarios. Using XToM, we systematically evaluate LLMs (e.g., DeepSeek R1), revealing a pronounced dissonance: while models excel in multilingual language understanding, their ToM performance varies across languages. Our findings expose limitations in LLMs' ability to replicate human-like mentalizing across linguistic contexts[1].

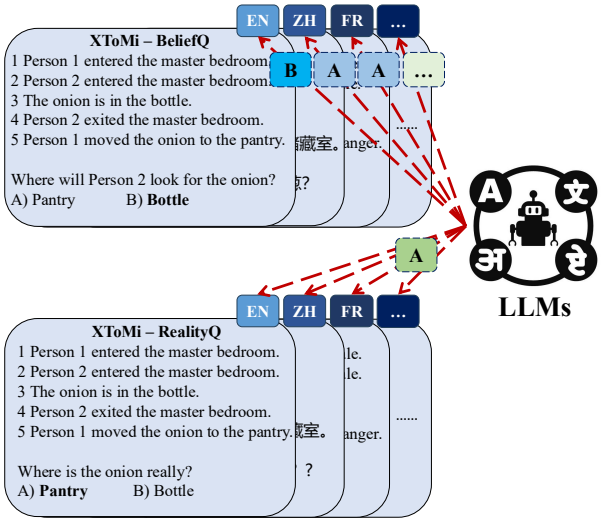


Figure 1: The example of a belief question and a fact question (i.e., RealityQ) from XToM for assessing LLMs' multilingual theory of mind ability. Ground truth is in bold, and LLMs answer correctly for fact questions while responses (i.e., selected choice in dash box) vary in belief questions for different languages.

1 Introduction

Theory of Mind (ToM), the capacity to infer and attribute the mental states of others, is a cornerstone of human social cognition, enabling individuals to navigate complex interpersonal interactions by understanding beliefs, intentions, and emotions (Premack and Woodruff, 1978; Ma et al., 2023b). Numerous scenarios rely on the ToM modeling of the mental states of others, such as multiagent-based simulation (Pynadath and Marsella, 2005; Yim et al., 2024), planning (Favier et al., 2023), negotiation (Yang et al., 2021), and various forms of reasoning and decision-making (Pereira et al., 2016; Rusch et al., 2020). Large language models (LLMs) have demonstrated remarkable proficiency in extensive natural language processing (NLP) tasks (Bubeck et al., 2023), but their capacity for ToM reasoning remains contentious. Some previous research believes that LLMs already exhibit a high level of competence in addressing ToM tasks (Strachan et al., 2024; Bubeck et al., 2023; Kosinski, 2023), while other studies express doubt and develop benchmarks to illustrate that LLMs do not possess proficient ability in ToM tasks (Sap et al., 2022; Ullman, 2023; Shapira et al., 2024). This contentious landscape reflects a deeper issue: existing ToM research for LLMs has been predominantly confined to English-language contexts, creating a critical blind spot in our understanding of these models' cognitive capabilities.

Existing monolingual benchmarks, though valuable for initial assessment, fail to capture the linguistic diversity inherent to human cognition, potentially inflating performance estimates based on language-specific artifacts rather than genuine rea-

[*] Equal contribution.

[1]Code and data are available at https://github.com/HKUST-KnowComp/XToM.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17681–17716
July 2-7, 2026 ©2026 Association for Computational Linguistics

soning abilities. This limitation prevents us from understanding whether LLMs' apparent ToM capabilities represent robust, language-invariant cognitive processes or merely statistical patterns learned from English training data. Previous research highlights a nuanced interplay between multilingualism and Theory of Mind (ToM), with bilingual individuals often demonstrating enhanced mental-state reasoning compared to monolinguals. Studies reveal that early bilingual exposure correlates with superior ToM performance, as bilinguals exhibit heightened empathy and outperform monolinguals on tasks requiring belief attribution (Nguyen and Astington, 2014; Javor and Javor, 2016; Schroeder, 2018; Buac and Kaushanskaya, 2020; Zhu et al., 2021). Given that state-of-the-art large language models (LLMs) such as GPT-4o (Hurst et al., 2024) and DeepSeek R1 (DeepSeek-AI et al., 2025) are pre-trained on vast multilingual corpora and have demonstrated their superior cross-lingual capabilities (Chirkova and Nikoulina, 2024; Pires et al., 2019; Wu and Dredze, 2019), a critical question emerges: *Do LLMs exhibit Multilingual Theory of Mind—the ability to reason about mental states consistently across diverse linguistic contexts?*

Investigating this intersection is essential for advancing our fundamental understanding of whether LLMs' apparent "cognitive" capacities are truly language-agnostic or merely artifacts of English-centric training data—a gap underscored by critiques of their superficial statistical learning (Bender and Koller, 2020; Shapira et al., 2024). To address this, we developed XToM, a high-quality, human-validated multilingual ToM benchmark that encompasses five languages and diverse task scenarios. Our comprehensive evaluation of state-of-the-art models reveals a striking dissonance: despite LLMs excelling in multilingual language understanding, discrepancies across different languages in XToM. As illustrated in Figure 1, LLMs answer correctly and consistently in a multilingual fact question, while responses vary in a multilingual belief question. These findings suggest that their apparent reasoning abilities may be surface-level rather than rooted in robust, language-invariant cognition. This paper aims to bridge the gap in multilingual ToM research by introducing three key contributions:

- **High-Quality Benchmark Creation**: We develop XToM, a high-quality multilingual ToM benchmark encompassing five languages and diverse task scenarios, human-annotated through the Multidimensional Quality Metrics (MQM) framework to ensure robustness and reliability. XToM advances ToM evaluation by integrating linguistic diversity to address critical limitations of prior benchmarks.
- **Exploration Across Language**: We systematically explore LLMs' ToM capabilities across languages, exposing critical limitations in their multilingual reasoning. Our findings underscore a fundamental disconnect: LLMs' statistical mastery of language patterns does not equate to human-like social cognition.
- **Empirical Insights**: We undertake the necessary empirical experiments to evaluate LLMs on the XToM benchmark and conduct extensive in-depth analysis to explore the LLMs' empirical performance under various settings.

2 Related Work

2.1 Theory of Mind Benchmarks

Early computational efforts built upon foundational psychological experiments (i.e., the Sally-Anne test (Baron-Cohen et al., 1985)), including the ToM-bAbi dataset (Grant et al., 2017) and its refinement into ToMi (Le et al., 2019), laid the groundwork for assessing false belief understanding as a core aspect of ToM. More recently, several new benchmarks have emerged to address limitations in previous ToM evaluations, offering more realistic and challenging scenarios. FANToM (Kim et al., 2023) assesses ToM capabilities in dialogue contexts, while TOMBench (Chen et al., 2024) provides a comprehensive framework with 31 social cognition abilities across eight tasks. While these advancements have significantly deepened our understanding of ToM in AI systems, exploring multilingual ToM capabilities remains underexplored, presenting a critical gap that our work aims to address. More related works on the theory of mind are provided in Appendix B.1. It is worth noting that a prior work (Sadhu et al., 2024) developed a multilingual theory of mind benchmark (i.e., Multi-TOM) by leveraging LLMs to translate the TOMBench (Chen et al., 2024) into various languages automatically. Their work focused on how cultural contexts relate to the theory of mind, whereas our research aims to explore LLM's multilingual theory of mind ability. Moreover, a key distinction lies in the translation quality assurance pro-

cess: Multi-TOM relies on the back-translation of many samples using Google Translate to preserve the core narrative. In contrast, our benchmark employs experts to manually correct and evaluate the translations based on the Multidimensional Quality Metrics (MQM) framework, which assesses quality across more diverse dimensions than Multi-TOM.

2.2 Multilingual Capabilities of LLMs

State-of-the-art large language models (LLMs), such as GPT-4o (Hurst et al., 2024), LLaMA (Dubey et al., 2024), Mistral (Jiang et al., 2023b), and DeepSeek R1 (DeepSeek-AI et al., 2025), are pre-trained on multilingual corpora, leveraging linguistic similarities and shared representations to enhance low-resource language performance (Zeng et al., 2025; Wendler et al., 2024; Dumas et al., 2025). Despite their cross-lingual capabilities (Chirkova and Nikoulina, 2024; Pires et al., 2019; Wu and Dredze, 2019), performance disparities persist, prompting efforts to quantify and mitigate them (Li et al., 2024c; Kumar et al., 2024; Zeng et al., 2024). Multilingual models have demonstrated strong reasoning abilities, including the capacity to reason in underrepresented languages (Shi et al., 2022). Techniques like chain-of-thought prompting (Qin et al., 2023) and preference optimization (She et al., 2024) have further improved reasoning performance in non-dominant languages. However, the ability of large language models to exhibit multilingual Theory of Mind reasoning across diverse linguistic contexts remains largely underexplored, and further investigation in this area is necessary.

3 XToM

The construction pipeline of XToM is systematically organized into four distinct phases to ensure linguistic diversity and high-quality outputs. These phases include §3.1 **Source Data Sampling**, §3.2 **Preprocessing and Translation**, §3.3 **Human Annotation**, and §3.4 **Quality Evaluation**.

XToM. We use $\mathcal{M}$ to denote a multilingual parallel dataset for XToM and define an XToM instance M as a set of story or dialogue S, question Q, and answer A. Each instance M has multiple semantically equivalent versions in different languages. Let us denote a *language* by $l \in \mathcal{L}$ where $\mathcal{L} = \{en, zh, de, fr, ja\}$ and $|\mathcal{L}|$ is the number of languages of interest. $\{en, zh, de, fr, ja\}$ represent English, Chinese, German, French, and Japanese, respectively. Then, M^l is the instance M in the language l. For example, M^{en} and M^{fr} denote the instance with the same meaning but in English (en) and French (fr), respectively. Therefore, the multilingual dataset $\mathcal{M}$ consists of $S \times |\mathcal{L}| \times K$ questions and answers. K is the number of questions and answers in a story or dialogue with the same language. Finally, we can formally describe a multilingual dataset $\mathcal{M}$ for XToM:

$$\forall M \in \mathcal{M},\ \forall (l_x, l_y) \in \mathcal{L}^2,\ \ \forall i \in \mathbb{N}_{\leq K},\ M_i^{l_x} \bowtie M_i^{l_y}\ . \tag{1}$$

We use the notation $\bowtie$ to indicate two instances in different languages (e.g., l_x and l_y) are semantically equivalent to each other. i is the index of question and answer and $\mathbb{N}_{\leq K}$ is a natural number smaller than or equal to K.

3.1 Source Dataset Sampling

To ensure the quality, diversity, and representativeness of the source data used for dataset construction, we systematically curated a balanced subset of 300 stories and dialogues from three distinct, well-established benchmarks, ranging from theoretical to application of the Theory of Mind (ToM). These three well-established datasets, ToMi (Le et al., 2019), FANToM (Kim et al., 2023), and NegotiationToM (Chan et al., 2024c), are utilized to create three subsets of XToM (i.e., XToMi, XFANToM, and XNegotiationToM). More details on the source dataset sampling are provided in Appendix A.1.

Potential Contamination To prevent the potential contamination issue in sampled ToM benchmarks, we follow the established protocols by prior works (Golchin and Surdeanu, 2024; Li and Flanigan, 2024) to verify contamination issues. For FANToM and NegotiationToM, none of the LLMs' generated responses matched with the data instance by using two established protocols across various languages, which indicates most instances of the XToM benchmark (i.e., two subtasks XFANToM and XNegotiationToM) are not identified as suffering from the data contamination issue, and the experimental results of the paper are reliable and valuable. An interesting finding is that some LLM-generated responses match some ToMi story patterns or even data instances. However, to ensure the representativeness of the dataset, we still have to collect the ToMi instances for reference purposes, as ToMi is a classical and widely used ToM

	Metric	en→zh GPT-4o | DeepL | Human	en→de GPT-4o | DeepL | Human	en→fr GPT-4o | DeepL | Human	en→ja GPT-4o | DeepL | Human
XToMi	LASER	0.924 | 0.920 | **0.938**	**0.950** | 0.948 | 0.948	0.947 | **0.955** | **0.955**	0.927 | **0.932** | 0.927
	$P_{\text{BERTScore}}$	0.870 | 0.868 | **0.877**	**0.913** | 0.912 | 0.912	0.899 | **0.901** | 0.900	**0.870** | 0.863 | **0.870**
	$R_{\text{BERTScore}}$	0.864 | 0.864 | **0.878**	**0.913** | 0.911 | 0.911	**0.886** | **0.886** | 0.885	**0.860** | 0.848 | **0.860**
	$F_{\text{BERTScore}}$	0.867 | 0.865 | **0.884**	**0.913** | 0.911 | 0.911	0.892 | **0.893** | **0.893**	**0.865** | 0.855 | **0.865**
	Human Evaluation (MQM)	95.91 | - | **99.27**	94.09 | - | **99.27**	97.09 | - | **99.52**	97.92 | - | **98.39**
	Avg. Word Count	70.16	52.03	61.88	122.16
XFANToM	LASER	**0.958** | 0.950 | **0.958**	**0.979** | 0.978 | **0.979**	**0.974** | **0.974** | **0.974**	0.949 | 0.940 | **0.956**
	$P_{\text{BERTScore}}$	**0.841** | **0.841** | **0.841**	**0.883** | 0.881 | **0.883**	0.869 | **0.870** | 0.869	**0.831** | 0.826 | **0.831**
	$R_{\text{BERTScore}}$	0.846 | 0.845 | **0.847**	**0.883** | 0.881 | **0.883**	0.865 | **0.866** | 0.865	**0.834** | 0.820 | **0.834**
	$F_{\text{BERTScore}}$	**0.844** | 0.843 | **0.844**	**0.883** | 0.881 | **0.883**	0.867 | **0.868** | 0.867	**0.833** | 0.825 | **0.833**
	Human Evaluation (MQM)	99.21 | - | **99.99**	98.87 | - | **99.47**	99.47 | - | **99.91**	99.90 | - | **99.94**
	Avg. Word Count	1031.9	652.13	727.81	1522.57
XNegotiationToM	LASER	**0.904** | 0.901 | 0.891	**0.968** | 0.964 | **0.968**	**0.957** | 0.950 | 0.956	0.910 | 0.900 | **0.912**
	$P_{\text{BERTScore}}$	**0.858** | **0.858** | **0.858**	**0.904** | 0.901 | **0.904**	**0.893** | 0.891 | **0.893**	**0.844** | 0.831 | 0.831
	$R_{\text{BERTScore}}$	0.863 | 0.860 | **0.865**	**0.900** | 0.899 | **0.900**	0.877 | **0.878** | 0.877	**0.842** | 0.832 | 0.834
	$F_{\text{BERTScore}}$	**0.861** | 0.860 | **0.861**	**0.902** | 0.900 | **0.902**	**0.885** | 0.884 | **0.885**	**0.843** | 0.830 | 0.832
	Human Evaluation (MQM)	96.37 | - | **99.74**	95.00 | - | **99.75**	97.54 | - | **99.00**	98.06 | - | **99.99**
	Avg. Word Count	187.04	133.94	150.1	313.25

Table 1: Benchmark quality evaluation for different translation methods across three partitions of XToM benchmarks (i.e., XToMi, XFANToM, and XNegotiationToM). GPT-4o, DeepL, and Human indicate GPT-4o translation, DeepL translation, and Human Annotation. $P_{\text{BERTScore}}$, $R_{\text{BERTScore}}$, and $F_{\text{BERTScore}}$ refer to the BERTScore Precision, BERTScore Recall, and BERTScore F1. en, zh, de, fr, and ja represent English, Chinese, German, French, and Japanese, respectively. The average word counts for the English versions of ToMi, FANToM, and NegotiationToM are 51.39, 635.68, and 139.12, respectively.

benchmark in the ToM field. More experimental details are reported in Appendix A.4.

3.2 Dataset Preprocessing and Translation

Preprocessing. Label consistency is critical to fairly assess LLMs' performance across different languages, and the translated instance with an inconsistent label may significantly affect the LLMs' performance during the evaluation. For instance, the intention label "Discover-Preference" from NegotiationToM may be translated to "Découvrir-Préférence" or "Découvrir la préférence" in French. To address this, we consulted native speakers of each target language to identify the most contextually appropriate and semantically accurate translation for a tailored set of labels and specific terms. Based on their feedback, we applied a mapping strategy to standardize the translation of labels across all languages, ensuring uniformity in the dataset.

Translation. Since the translation quality is crucial for ensuring the reliability of translated datasets and experimental results in our cross-linguistic research, we employ the two commonly used translation methods (i.e., GPT-4o (Hurst et al., 2024) and DeepL[2]) for translating the sampled data from the original language (i.e., English) to multiple target languages Chinese, German, French, and Japanese. After conducting an automatic translation quality evaluation, the results presented in Table 1 indicate that GPT-4o outperforms DeepL in overall translation quality on our dataset. Therefore, we chose the GPT-4o translated version for the human annotation phase. Two automatic translation quality evaluation metrics were used, which are LASER[3] (Artetxe and Schwenk, 2019a,b) and BERTScore[4] (Zhang et al., 2020).

3.3 Human Annotation

We employ three qualified human annotators for each language to refine the translated data, ensuring alignment with the predefined dimensions of the Multidimensional Quality Metrics (MQM) framework[5] (Burchardt, 2013; Mariana, 2014; Lommel et al., 2024). This process involves correcting errors and enhancing semantic accuracy, grammatical correctness, contextual coherence, and stylistic consistency. The MQM framework is a widely adopted standard for analytical translation quality evaluation (TQE), serving for human annotation and evaluation purposes, systematically assessing and counting translation errors across eight predefined dimensions (Freitag et al., 2021; Li et al., 2025b). These dimensions include Terminology, Accuracy, Linguistic Conventions, Style, Locale Conventions, Audience Appropriateness, Design and Markup, and Custom (e.g., label and name consistency, as discussed in §3.2). To maintain high annotation quality, each annotator undergoes a preliminary training phase for understanding the definitions and examples associated with

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

[3]https://github.com/facebookresearch/LASER

[4]https://github.com/Tiiiger/bert_score

[5]https://themqm.org/

each dimension. The qualification details of human annotators are outlined in Appendix A.2, and comprehensive details regarding the framework and its pre-defined dimensions are in Appendix A.3.

3.4 Human Evaluation

To guarantee the dataset quality and accurate alignment with established dimensions, we employ three qualified human annotators for each language to assess the translation quality of XToM by using the Multidimensional Quality Metrics (MQM) framework. The qualification details of human annotators can be found in Appendix A.2. The MQM framework computes the overall quality score (OQS) by counting the translation error across eight specified dimensions according to the formula shown in Table 2 in Appendix A.2. While the standard passing threshold for the OQS in the MQM framework is typically set at 90 overall quality score, we adopted a more stringent threshold of 95 to ensure higher-quality annotated data. This elevated threshold was used to determine whether a translated instance passes or fails.

Discussion on Length Effect on Translation Quality Evaluation To explore the translation quality of our dataset, we performed an automatic quality evaluation and human evaluation on the translated instances from various source datasets, which reveals a length effect on the automatic evaluation methods. The result is shown in Table 1. We observed that automatic metrics like LASER and BERTScore exhibit diminished effectiveness as word length increases, resulting in the LASER score and BERTScore of the human-corrected version being similar to the GPT-4o translated version in the XFANToM portion. This aligns with findings from prior works (Peng et al., 2024; Wang et al., 2021; Herold and Ney, 2023), which *highlighted the challenges of automatic context-aware evaluation in longer texts and the fact that it leads to inconsistent evaluation scores.* Moreover, BERTScore gives different correct translations of the same sentence with lower scores when that sentence contains difficult words that have little lexical meaning or have an ambiguous meaning (i.e., function words) (Hanna and Bojar, 2021), resulting in the BERTScore of the human corrected version being lower than the GPT-4o translated version for XToMi (en→de) and XNegotiationToM (en→ja). However, this issue is mitigated by using the MQM framework, which computes the overall quality score by using the *Per-Word Penalty* and taking the word count into consideration. We observed that the human-annotated version received a higher overall quality score than the GPT-4o translated version, demonstrating the high quality of the human-annotated version. By using the MQM framework, we regard the pass or fail label of data instances in human evaluation phrases as the label for computing the inter-annotator agreement. We observed high inter-annotator agreement, and the overall Fleiss's κ is 95.18% (Fleiss, 1971) for the XToM benchmark. The breakdown computation of κ is shown in Table 7 in the Appendix.

3.5 Dataset Statistics

XToM contains 900 stories from ToMi, FANToM, and NegotiationToM, which range from theoretical to application and provide contextual diversity for assessing the LLM performance across different languages. There are a total of 15,115 questions, which are 3,090 questions for XFANToM, 3,235 questions for XToMi, and 8,790 questions for XNegotiationToM. The detailed statistics are shown in Table 6 in the Appendix.

4 Experiments

4.1 Experimental Settings

In this work, we test ten recent state-of-the-art large language models: DeepSeek R1 (DeepSeek-AI et al., 2025), Qwen-2.5-7b-instruct (Yang et al., 2024), Qwen-2.5-72b-instruct (Yang et al., 2024), Qwen-3-235b-a22b (Yang et al., 2025a), llama-3.1-8b-instruct (Dubey et al., 2024), llama-3.3-70b-instruct (Dubey et al., 2024), mistral-7b-instruct-v0.3 (Jiang et al., 2023a), mixtral-8x22b-instruct-v0.1 (Jiang et al., 2023a), GPT-4o-11-20 (Hurst et al., 2024), and GPT-3.5-turbo (OpenAI, 2022). By following the common practices in the theory of mind field (Kim et al., 2023; Gandhi et al., 2023a; Shapira et al., 2023a), we test these models with two types of prompts: (1) One is zero-shot prompting, and we utilize the prompt template in the original paper. (2) Another one is the chain of thought (CoT) prompting method by following Wei et al. (2022) and using the prompt "let's think step by step." For the prompt template in each language, the human annotator was consulted to translate the English prompt template to another language. More configuration details can be found in Appendix C.1, and the prompt template and data example refer to Appendix C.2.

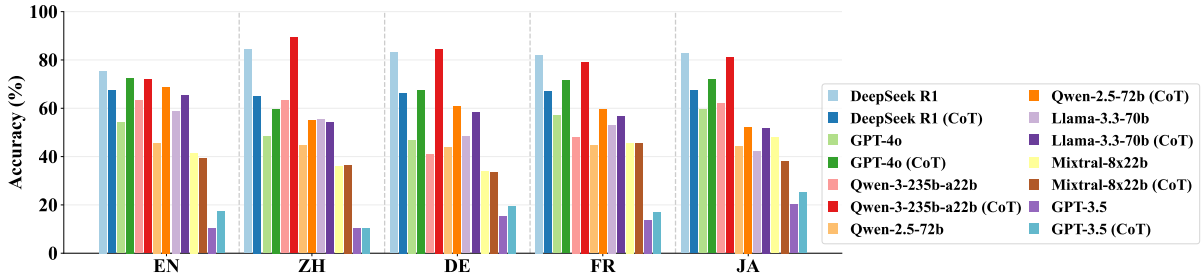


Figure 2: Performance comparison of various models on belief questions across languages in XFANToM.

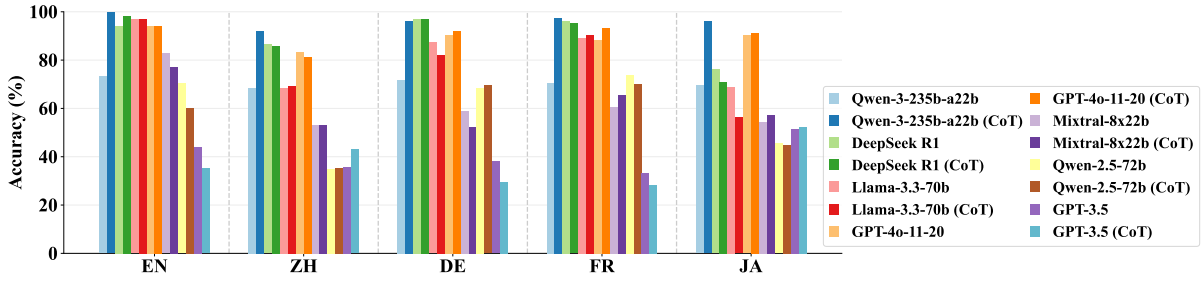


Figure 3: Performance comparison of various models on belief questions across languages in XToMi.

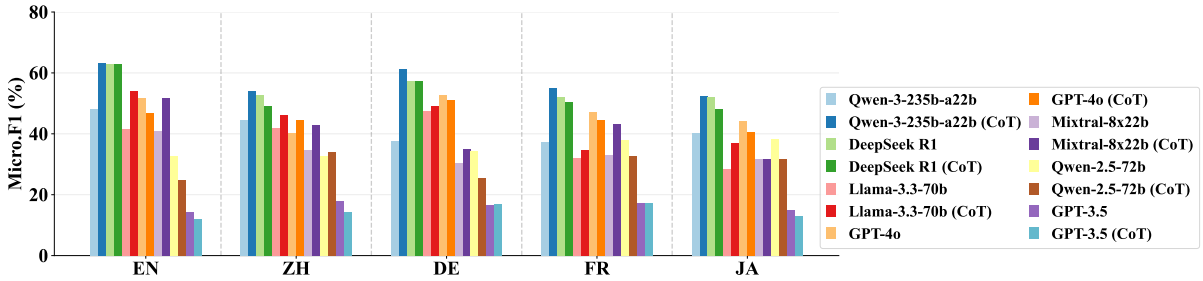


Figure 4: Performance comparison of various models on belief choice across languages in XNegotiationToM.

Evaluation Metrics We adopt the identical evaluation metrics utilized in all source benchmarks, except for fact questions, because we reformulate fact questions in XFANToM as the multiple-choice format for fair comparison purposes. Accuracy is used as the evaluation metric for multiple-choice belief and fact questions in XFANToM and XToMi. In XNegotiationToM, we report the exact match percentages across all three categories of high, medium, and low preferences for desire and belief classification. Only these three preferences that answer correctly count toward the correct answer. For intention classification, we report both micro and macro F1 scores to evaluate the model's performance across multiple labels comprehensively.

5 Main Results

Figures 2, 3, and 4 summarize the main results of the state-of-the-art large language models for belief questions in XToM, while Figures 15 and 16 report LLMs' performance on desire and intention questions in XNegotiationToM in Appendix D.1. Based on the evaluation result, we derive the following conclusions. The (1) **Language Model Scaling Effects.** Larger models consistently outperform their smaller models across all three subtasks. Smaller models, such as Mistral-7b-instruct-v0.3, exhibit significantly lower performance, suggesting that larger-scale training enhances ToM capabilities. The detailed performance of more models, including smaller-scale LLMs, is provided in Tables 8, 9, and 10 in Appendix D. (2) **Language variation generally impacts model performance.** Most LLMs' performance varies across different languages, and each LLM has its language preference. This discrepancy may arise from reporting bias (Gordon and Durme, 2013) inherent in the pre-training corpus, where certain languages are overrepresented (e.g., English) or inherent differences in linguistic complexity across languages. (3) **Qwen3-235b-a22b and DeepSeek R1 demonstrates superior performance.** Large reasoning models Qwen3-235b-a22b and DeepSeek R1 demonstrate superior performance compared to other models across most sub-tasks in XToM. The significant performance gap between these two models and other models suggests that the

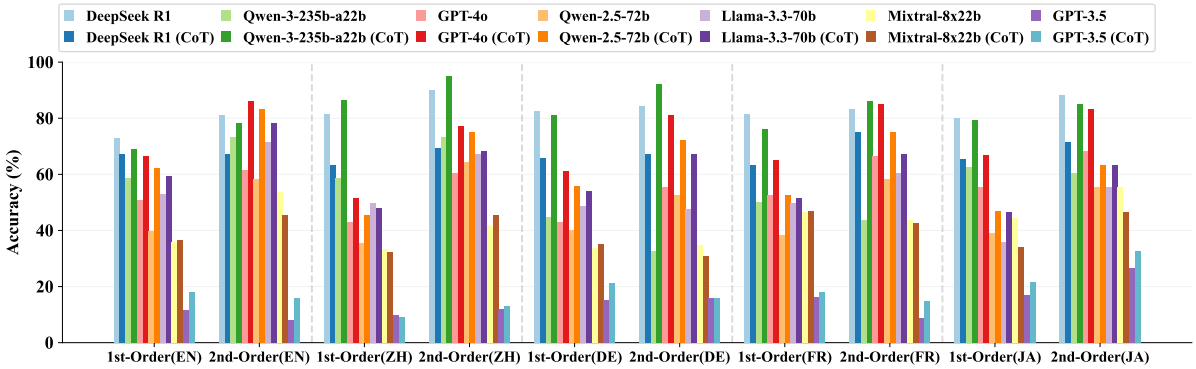


Figure 5: Performance comparison of different models on first- and second-order belief questions across languages in XFANToM.

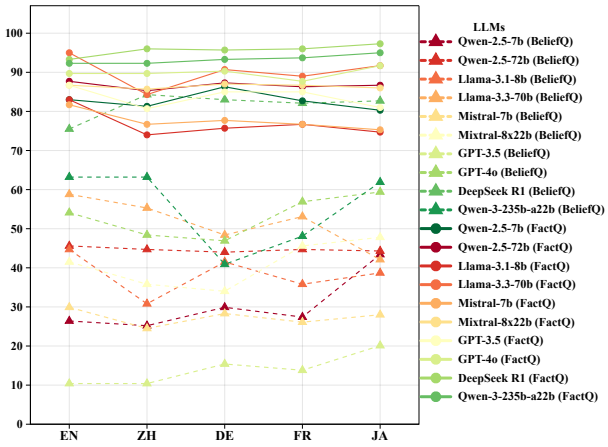


Figure 6: Comparison of LLMs' performance on belief questions (i.e., BeliefQ) versus fact questions (i.e., FactQ) in the XFANToM dataset.

reinforcement fine-tuning method enhances the ToM reasoning capabilities (Yang et al., 2025a; DeepSeek-AI et al., 2025). The comprehensive analysis of error analysis and case study is provided in Appendix D.3.

5.1 First Order versus Second Order ToM

To fully assess the ToM capability of large language models to understand others' mental states in each complexity, Figure 5 presents larger-scale LLMs' performance on first- and second-order false belief questions across various languages in XFANToM. The result illustrates that most models receive a higher accuracy in second-order belief questions compared to first-order belief questions across different languages, which is similar to the findings by Le et al. (2019) and Kim et al. (2023). The detailed performance of more models, including smaller-scale LLMs on first- and second-order false belief questions in XFANToM and XToMi across various languages, is displayed in Figures 17 and 18 in Appendix.

5.2 Most LLMs Lack Robust ToM Reasoning Across Linguistic Contexts

Fact questions in XToM require LLMs to understand the provided context (i.e., story or conversation) across different languages and retrieve the answer in the provided context, while belief questions require the models not only to understand the provided context across different languages but also to possess the theory of mind reasoning capability. Figure 6 presents the performance comparison of belief and fact questions in XFANToM. The results reveal that existing LLMs (except DeepSeek R1) lack the robust theory of mind capability across diverse linguistic settings. While all models achieve high accuracy in fact questions across different languages, proving their basic multilingual comprehension abilities, their performance is significantly lower in the multilingual theory of mind reasoning task. The substantial performance gap between the belief and fact questions, combined with our error analysis in Appendix D.3, identifies ToM reasoning errors as the primary failure mode. This pattern reflects reporting bias (Gordon and Durme, 2013) inherent in LLMs' pre-training corpora, where most datasets emphasize factual information rather than mental states, since mental states exist implicitly in human cognition and are less frequently documented in text.

5.3 Multilingual Consistency

To further analyze multilingual consistency across various LLMs, we assess whether models provide the identical answers to the same instances across different languages. For clarity, we present FANToM results in Figure 7 and include others in Appendix D.2. Models perform well in factQA, achieving high accuracy and consistency, but struggle in BeliefQA, often giving inconsistent and incorrect responses. This suggests that the models'

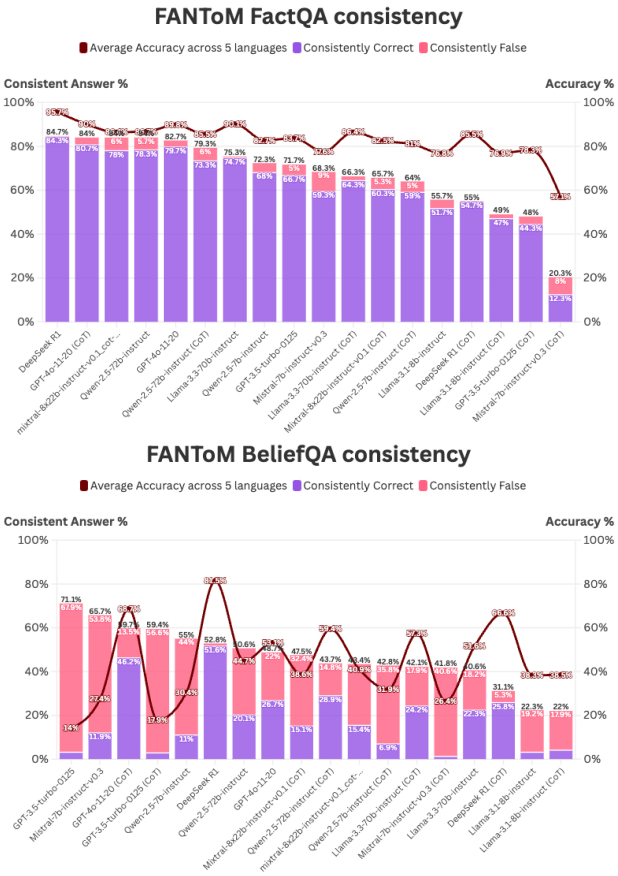


Figure 7: Consistency analysis across different languages. *Consistently correct* indicates that the LLM answers the same instance correctly across all different languages, while *consistently false* represents answering the same instance incorrectly across all languages.

ToM reasoning capability is significantly weaker than their information retrieval ability. Fact consistency is significantly higher than belief consistency, indicating that strong factual understanding does not guarantee robust ToM reasoning across languages. Correlation analysis reveals that fact accuracy and consistency are strongly correlated (i.e., correlation coefficient is 0.8898), while belief accuracy and consistency show virtually no correlation (i.e., correlation coefficient is -0.08). This stark difference highlights substantial variation in ToM reasoning abilities across different languages. Model scaling improves fact consistency but does not necessarily enhance belief consistency, demonstrating that increased model size alone is insufficient for improving robust cross-linguistic ToM capabilities. These findings suggest models perform well on multilingual fact understanding but lack robust cross-linguistic belief reasoning, underscoring the need for specialized ToM training to improve consistency and correctness across languages.

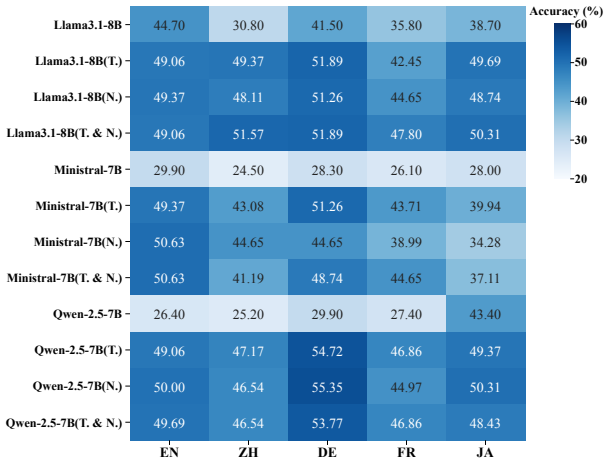


Figure 8: Performance comparison of task transfer on belief questions in XFANToM. *T.* and *N.* indicate the model fine-tune on XToMi and XNegotiationToM task and evaluated on the XFANToM Task.

5.4 Task Transfer Learning

This experiment is crucial for understanding how language models transfer their theory of mind capabilities across different tasks, revealing the effectiveness of specialized fine-tuning for belief reasoning. The results reported in Figure 8 demonstrate several notable patterns: First, base models without fine-tuning (i.e., Llama3.1-8B, Mistral-7B, and Qwen-2.5-7B) consistently underperform compared to their fine-tuned counterparts, with accuracy improvements of around 10-50% when models are fine-tuned on either XToMi or XNegotiationToM. Second, models fine-tuned on both datasets simultaneously (T. & N.) generally achieve the highest performance across languages, particularly evident in Llama3.1-8B(T. & N.), which reaches 51.57% accuracy in Chinese (ZH) and maintains strong performance across all languages. The combined fine-tuning approach consistently produced more robust multilingual belief reasoning, suggesting that diverse theory of mind training datasets create complementary benefits that enhance cross-task generalization.

5.5 Evaluation of LLMs' Multilingual Transfer Capabilities

By fine-tuning the same number of instances across various languages on open-source LLMs, we evaluate LLMs' multilingual transfer capabilities. A question is raised here according to the experimental findings: How to enhance the multi-language ToM ability? Therefore, we study whether learning from other languages can help LLM understand the

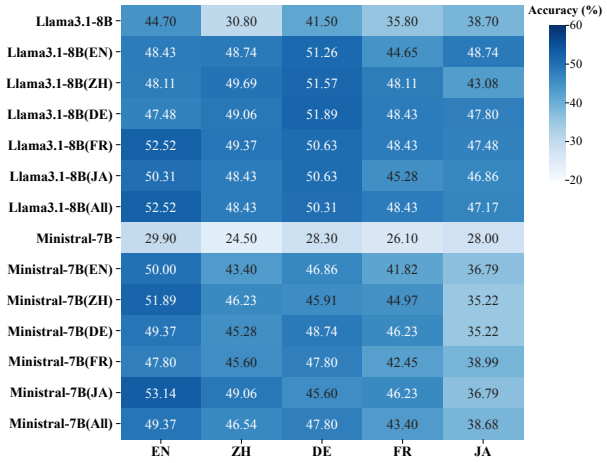


Figure 9: Performance comparison of cross-lingual transfer on belief questions in XFANToM. *All* indicates models fine-tuned on XFANToM dataset containing all five languages.

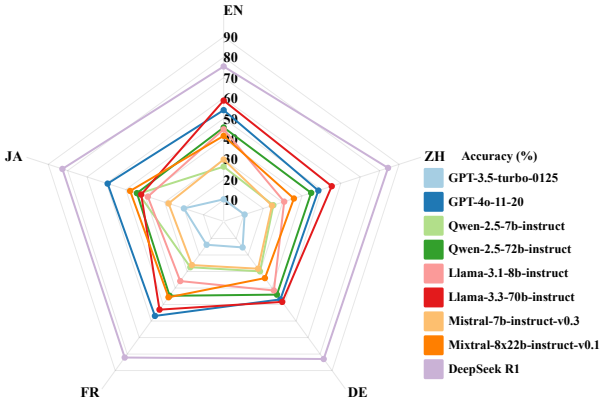


Figure 10: LLMs Language Preference in XFANToM.

Theory of Mind. Exploring the potential error that occurred in this situation, we want to find whether the gap existed between the model because of pre-training data, lack of source dataset, or lack of the ToM task dataset. Moreover, we intend to see whether it can enhance the LLM's performance. Figure 9 shows the results of LLaMA3.1-8B and Mistral-7B, which are fine-tuned on five monolingual datasets separately and then evaluated across all five languages on belief questions in XFANToM. The results indicate that cross-lingual transferability can be observed in most cases, meaning that training in another monolingual can lead to performance improvements on the test set of the target language. For example, training in French achieves the best performance in English, while training in English yields the best performance in Japanese.

Which is Better: Monolingual Fine-tuning or Multilingual Fine-tuning? We analyze the effectiveness of multilingual fine-tuning on belief questions through XFANToM. The results in Figure 9 indicate that some languages exhibit improved performance after multilingual fine-tuning. For instance, LLaMA3.1-8B outperforms monolingual training in both English and Japanese under the multilingual setting, while Mistral-7B surpasses monolingual performance in Chinese and French.

5.6 Do Different LLMs have Similar Language Preferences?

While many state-of-the-art LLMs claim multilingual proficiency, their actual performance varies significantly depending on the language, raising essential questions about their underlying linguistic biases. The experimental results in Figure 10 on the XFANToM reveal that many LLMs exhibit distinct language preferences, as evidenced by their varying performance across languages. For instance, GPT-3.5-turbo-0125 and GPT-4o-11-20 demonstrate a notable preference for Japanese but lag behind in Chinese, and the performance gap of GPT-4o-11-20 between Japanese and Chinese is 11%. Interestingly, DeepSeek R1 excels across multiple languages, achieving scores of 84.3% in Chinese and maintaining high performance in German, French, and Japanese as well, but lagging in English at 75.5%. Notably, the Qwen models exhibit varied results, with Qwen-2.5-72b-instruct performing uniformly around 45% across languages, while Qwen-2.5-7b-instruct shows less consistency. The Llama models present exhibits a clear preference for English, with Llama-3.3-70b-instruct scoring highest in English at 58.8% but lagging in Japanese at 42.1%. These findings emphasize that multilingual performance is highly model-dependent, with some LLMs favoring specific linguistic structures, potentially due to variations in pre-training corpus, model architecture, and tokenization strategies.

6 Conclusion

This work introduces XToM, a high-quality multilingual Theory of Mind benchmark tailored to assess the multilingual ToM ability of current LLMs. We performed comprehensive and detailed experiments to evaluate LLMs' capability on the XToM benchmark and expose critical limitations in LLMs' multilingual reasoning. LLMs are equipped with multilingual understanding ability but fail in multilingual ToM reasoning tasks.

Limitations

Limited Language Coverage and Source Dataset Coverage. The rigorous human annotation process and stringent data quality requirements present significant challenges in recruiting sufficiently qualified annotators to achieve broader language and source dataset coverage. Although XToM encompasses three existing benchmarks spanning the theoretical to the applied aspects of the theory of mind, many multifaceted dimensions of ToM remain unaddressed in our benchmark, including specialized assessments such as the Faux Pas test (Baron-Cohen et al., 1999) and OpenToM (Xu et al., 2024). Furthermore, since this study aims to construct a high-quality multilingual benchmark for exploring LLMs' multilingual theory of mind capabilities, the language coverage is inherently constrained by annotator availability and expertise. The difficulty in securing qualified annotators for diverse languages limits XToM's linguistic diversity to regions spanning Western Europe to East Asia, leaving other linguistic families underrepresented.

Potential Contamination Most of the existing available benchmarks in the NLP field were released prior to the initiation of the LLM training process, indicating that these datasets are likely to have been utilized during the pre-training phase and post-training phase of LLMs (Golchin and Surdeanu, 2024; Li and Flanigan, 2024). Since ToMi is constructed before the initiation of the LLM training process, the ToMi benchmark may suffer from the potential contamination issue. To ensure the representativeness of XToM, we still have to collect the ToMi for reference purposes, as this dataset is a classical and widely used ToM benchmark. However, for FANToM and NegotiationToM, none of the LLMs' generated responses matched with the sampled data across various languages, which indicates most instances of the XToM benchmark (i.e., two subtasks XFANToM and XNegotiationToM) are not identified as suffering from the data contamination issue, and the experimental results of the paper are reliable and valuable. Therefore, data contamination may not be a primary concern in XToM.

Ethics Statement

In this work, we conformed to recognized privacy practices and rigorously followed the data usage policy. This paper introduces a multilingual benchmark for assessing the multilingual theory of mind of large language models. We conduct a human evaluation to strictly assess the Audience Appropriateness dimension of the MQM framework, and we assign a high value of Severity Multiplier to filter out the offensive context in XToM. Therefore, we can foresee no immediate social consequences or ethical issues as we do not introduce social/ethical bias into the model or amplify any bias from the data. Moreover, the license of these datasets allows us to modify the data for evaluation and research, and this fulfills their intended use.

Resource Copyright This work presents a new resource: XTOM, which are multilingual extension of the FANToM, ToMi, and NegotiationToM respectively. All these three sources of data are publicly available for free, and we do not add any additional requirements for accessing our resources. We will highlight the sources of our data and ask users to cite the original papers when they use our extended versions for research.

Acknowledgements

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China. We also thank the support from NVIDIA AI Technology Center (NVAITC).

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. 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 3197–3203. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21:37–46.

Simon Baron-Cohen, Michelle O'riordan, Valerie E. Stone, R. Jones, and Kate C. Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning

autism. *Journal of Autism and Developmental Disorders*, 29:407–418.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

Milijana Buac and Margarita Kaushanskaya. 2020. Predictors of theory of mind performance in bilingual and monolingual children. *International Journal of Bilingualism*, 24(2):339–359.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Chunkit Chan, Jiayang Cheng, Xin Liu, Yauwai Yim, Yuxin Jiang, Zheye Deng, Haoran Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024a. Audience persona knowledge-aligned prompt tuning method for online debate. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3851–3858. IOS Press.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024b. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024c. 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.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023a. Self-consistent narrative prompts on abductive natural language inference. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 1040–1057. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 35–57. 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. *ArXiv*, abs/2402.15052.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11518–11537. Association for Computational Linguistics.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. *Preprint*, arXiv:2402.14778.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2024. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *CoRR*, abs/2404.14215.

Zheye Deng, Chunkit Chan, Tianshi Zheng, Wei Fan, Weiqi Wang, and Yangqiu Song. 2025. Structuring the unstructured: A systematic review of text-to-structure generation for agentic AI with a universal evaluation framework. *CoRR*, abs/2508.12257.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. *Preprint*, arXiv:2411.08745.

Anthony Favier, Shashank Shekhar, and Rachid Alami. 2023. Models and algorithms for human-aware task planning with integrated theory of mind. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023*, pages 1279–1286. IEEE.

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

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023a. Understanding social reasoning in language models with language models. *CoRR*, abs/2306.15448.

Kanishk Gandhi, Jan-Philipp Franken, Tobias Gerstenberg, and Noah D. Goodman. 2023b. Understanding social reasoning in language models with language models. *ArXiv*, abs/2306.15448.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alison Gopnik and Janet Wilde Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, 59 1:26–37.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *AKBC*, pages 25–30.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? *Cognitive Science*.

Michael Hanna and Ondrej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 507–517. Association for Computational Linguistics.

Yinghui He, Yufan Wu, 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 *Conference on Empirical Methods in Natural Language Processing*.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. *CoRR*, abs/2306.05183.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Rebeka Javor and R Javor. 2016. Bilingualism, theory of mind and perspective-taking: The effect of early bilingual exposure. *Psychology and Behavioral Sciences*, 5(6):143–148.

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. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

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. 2023b. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023c. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 3134–3154. Association for Computational Linguistics.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024a. ECON: on the detection and resolution of evidence conflicts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7816–7844. Association for Computational Linguistics.

Cheng Jiayang, Lin Qiu, Chunkit Chan, Xin Liu, Yangqiu Song, and Zheng Zhang. 2024b. Eventground: Narrative reasoning by grounding to eventuality-centric knowledge graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 6622–6642. ELRA and ICCL.

Cheng Jiayang, Qianqian Zhuang, Haoran Li, Chunkit Chan, Xin Liu, Lin Qiu, and Yangqiu Song. 2025. Integround: On the evaluation of verification and retrieval planning in integrative grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 13587–13602. Association for Computational Linguistics.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer David Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *ArXiv*, abs/2401.08743.

Cameron R. Jones, Sean Trott, and Benjamin K. Bergen. 2024. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (epitome). *Transactions of the Association for Computational Linguistics*, 12:803–819.

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.

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

Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Tanuja Ganu, Sunayana Sitaram, Kalika Bali, and Akshay Nambi. 2024. Bridging the gap: Dynamic learning strategies for improving multilingual performance in llms. *Preprint*, arXiv:2405.18359.

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.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18471–18480. AAAI Press.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions. *CoRR*, abs/2310.10383.

Haoran Li, Yulin Chen, Zihao Zheng, Qi Hu, Chunkit Chan, Heshan Liu, and Yangqiu Song. 2024a. Backdoor removal for generative large language models. *CoRR*, abs/2405.07667.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024b. Privlm-bench: A multi-level privacy evaluation benchmark for 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 54–73. Association for Computational Linguistics.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025a. Towards agentic RAG with deep reasoning: A survey of rag-reasoning systems in llms. *CoRR*, abs/2507.09477.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025b. Mqm-chat: Multidimensional quality metrics for chat translation. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 3283–3299. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024c. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Fangzhou Liang, Tianshi Zheng, Chunkit Chan, Yauwai Yim, and Yangqiu Song. 2025. Llm-hanabi: Evaluating multi-agent gameplays with theory-of-mind and rationale inference in imperfect information collaboration game. *CoRR*, abs/2510.04980.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2024. Constrained reasoning chains for enhancing theory-of-mind in large language models. In *PRICAI 2024: Trends in Artificial Intelligence - 21st Pacific Rim International Conference on Artificial Intelligence, PRICAI 2024, Kyoto, Japan, November 18-24, 2024, Proceedings, Part II*, volume 15282 of *Lecture Notes in Computer Science*, pages 354–360. Springer.

Arle Lommel, Serge Gladkoff, Alan K. Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas, AMTA 2024 - Volume 2: User Track, Chicago, USA, September 30 - October 2, 2024*, pages 75–94. Association for Machine Translation in the Americas.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *CoRR*, abs/2302.00539.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023a. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Conference on Computational Natural Language Learning*.

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, Singapore, December 6-10, 2023*, pages 1011–1031. Association for Computational Linguistics.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Yunxiang Mo, Tianshi Zheng, Qing Zong, Jiayu Liu, Baixuan Xu, Yauwai Yim, Chunkit Chan, Jiaxin Bai, and Yangqiu Song. 2025. Dixitworld: Evaluating multimodal abductive reasoning in vision-language models with multi-agent dixit gameplay. *CoRR*, abs/2510.10117.

Thien-Kim Nguyen and Janet Wilde Astington. 2014. Reassessing the bilingual advantage in theory of mind and its cognitive underpinnings. *Bilingualism: Language and Cognition*, 17(2):396–409.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024. Investigating length issues in document-level machine translation. *CoRR*, abs/2412.17592.

Gonçalo Duarte Garcia Pereira, Rui Prada, and Pedro Alexandre Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artif. Intell.*, 241:1–44.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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.

David V. Pynadath and Stacy Marsella. 2005. Psychsim: Modeling theory of mind with decision-theoretic agents. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pages 1181–1186. Professional Book Center.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Preprint*, arXiv:2310.14799.

Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. 2020. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488.

Jayanta Sadhu, Ayan Antik Khan, Noshin Nawal, Sanju Basak, Abhik Bhattacharjee, and Rifat Shahriyar. 2024. Multi-tom: Evaluating multilingual theory of mind capabilities in large language models. *CoRR*, abs/2411.15999.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings*

of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 3762–3780. Association for Computational Linguistics.

Scott R Schroeder. 2018. Do bilinguals have an advantage in theory of mind? a meta-analysis. *Frontiers in Communication*, 3:36.

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. 2023a. 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.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. How well do large language models perform on faux pas tests? In *Annual Meeting of the Association for Computational Linguistics*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *Preprint*, arXiv:2401.06838.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Haochen Shi, Tianshi Zheng, Weiqi Wang, Baixuan Xu, Chunyang Li, Chunkit Chan, Tao Fan, Yangqiu Song, and Qiang Yang. 2025. Inferencedynamics: Efficient routing across llms through structured capability and knowledge profiling. *Preprint*, arXiv:2505.16303.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *CoRR*, abs/2212.09292.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Minghan Wang, Jiaxin Guo, Yuxia Wang, Yimeng Chen, Chang Su, Hengchao Shang, Min Zhang, Shimin Tao, and Hao Yang. 2021. How length prediction influence the performance of non-autoregressive translation? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, pages 205–213. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. 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 2351–2374. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 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, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and

22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 681–693. Association for Computational Linguistics.

Yuqi Yang, Weiqi Wang, Baixuan Xu, Wei Fan, Qing Zong, Chunkit Chan, Zheye Deng, Xin Liu, Yifan Gao, Changlong Yu, Chen Luo, Yang Li, Zheng Li, Qingyu Yin, Bing Yin, and Yangqiu Song. 2025b. Sessionintentbench: A multi-task inter-session intention-shift modeling benchmark for e-commerce customer behavior understanding. *CoRR*, abs/2507.20185.

Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi Zheng, and Yangqiu Song. 2024. Evaluating and enhancing llms agent based on theory of mind in guandan: A multi-player cooperative game under imperfect information. *CoRR*, abs/2408.02559.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. *Preprint*, arXiv:2404.04748.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, and 4 others. 2025. From web search towards agentic deep research: Incentivizing search with reasoning agents. *CoRR*, abs/2506.18959.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

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. 2023a. How far are large language models from agents with theory-of-mind? *ArXiv*, abs/2310.03051.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2022. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Annual Meeting of the Association for Computational Linguistics*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. *ArXiv*, abs/2310.11667.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12901–12911. PMLR.

Qing Zong, Jiayu Liu, Tianshi Zheng, Chunyang Li, Baixuan Xu, Haochen Shi, Weiqi Wang, Zhaowei Wang, Chunkit Chan, and Yangqiu Song. 2025. Critical: Can critique help LLM uncertainty or confidence calibration? *CoRR*, abs/2510.24505.

A Appendix for XTOM

A.1 Appendix for Sampling Source Data

To ensure the quality, diversity, and representativeness of the source data used for dataset construction and to evaluate the multilingual Theory of Mind ability of LLMs, we systematically curated a balanced subset of stories and dialogues from three distinct, well-established benchmarks, ranging from theoretical to application of the Theory of Mind (ToM). These three well-established Theory of Mind datasets are ToMi (Le et al., 2019), FANToM (Kim et al., 2023), and NegotiationToM (Chan et al., 2024c), utilized to create three subsets of XToM (i.e., XToMi, XFANToM, and XNegotiationToM). For the ToMi dataset, we randomly selected 300 stories, including first-order false belief, second-order false belief, and reality question categories, yielding a total of 647 questions. For the FANToM dataset, we similarly sampled 300 stories, prioritizing the short conversation version due to its manageable dialogue length (averaging 13.8 turns per conversation) compared to the longer version. This subset encompassed

618 questions, including fact questions (FACTQ), belief questions (BELIEFQ), and information accessibility questions (INFOACCESSQ). Finally, from the NegotiationToM dataset, we randomly selected 300 real-world negotiation dialogues, each accompanied by a fixed set of questions probing the beliefs, desires, and intentions of both agents, resulting in 1,758 questions. This systematic sampling approach not only preserves the theoretical richness of the dataset but also ensures a comprehensive evaluation of ToM reasoning capabilities in LLMs across varying levels of cognitive complexity.

A.2 Appendix for Human Annotator Qualification

In the construction process of XToM, all human annotators are required to manually conduct the human annotation and evaluation of the translated instances. To ensure high-quality bilingual annotations and evaluation, we implemented a rigorous annotator qualification protocol. We recruited a total of 12 annotators, with three qualified human annotators for each language in the human annotation and evaluation process. All annotators were required to be native speakers of one target language while demonstrating certified proficiency in English through standardized assessments. For instance, annotators working with the Chinese version were native Chinese speakers who possessed either TOEFL or IELTS certification, demonstrating advanced English language competency. Prior to full-scale annotation, we conducted a preliminary quality assurance phase where annotators completed a pilot set of 20 instances. This initial phase served dual purposes: it allowed us to evaluate annotator reliability and provided an opportunity to address typical errors through targeted feedback, thereby standardizing the annotation process across all participants.

A.3 Appendix for Multidimensional Quality Metrics (MQM) Framework

The *Multidimensional Quality Metrics* (MQM) framework is a structured and flexible system for assessing translation quality. It defines a set of dimensions that categorize different aspects of translation errors, allowing for precise and customizable evaluations. The core dimensions of MQM include:

- **Terminology**: Ensures the correct and consistent use of terms. Errors include incorrect terminology usage, inconsistencies, and deviations from domain-specific terminology guidelines.
- **Accuracy**: Evaluates whether the translation correctly conveys the meaning of the source text. Issues include mistranslation, overtranslation, undertranslation, addition, omission, and untranslated content.
- **Linguistic Conventions**: Covers grammatical correctness and linguistic coherence, including errors in grammar, punctuation, spelling, and unintelligibility.
- **Style**: Assesses adherence to appropriate tone and register. This includes language register, awkward phrasing, unidiomatic expressions, and inconsistent style. We place particular importance on avoiding unidiomatic expressions in translations.
- **Locale Conventions**: Ensures compliance with region-specific norms, such as number format, measurement format, time format, date format, address format, telephone format, and shortcut keys.
- **Audience Appropriateness**: Examines whether the translation is suitable for its target audience, considering readability, domain-specific terminology, and adaptation to audience expectations. In our case, we particularly emphasize culture-specific references and the avoidance of offensive content.
- **Design and Markup**: Focuses on structural and formatting aspects, identifying errors such as incorrect text formatting, misplaced tags, broken links, and layout inconsistencies.
- **Custom**: Allows for the definition of additional dimensions tailored to specific translation tasks or industry requirements, ensuring adaptability for specialized evaluations. In our scenario, we ask annotators to ensure that object labels are translated consistently across the answer, question, and story within a single data instance. In NegotiationToM, we focus on maintaining consistency in the translation of water, food, firewood, and intention labels. In ToMi, we emphasize the consistency of objects mentioned in both the story and the

Step	Overall Quality Score Calculation	Formulas
1	Absolute Penalty Total (APT)	$\sum_{i,j}$ Error Count$_{ij}$ $\times$ Severity Multiplier$_j$ $\times$ Error Type Weight$_i$ Where: i = index for Error Types, j = index for Severity Level.
2	Per-Word Penalty Total (PWPT)	$\frac{\text{Absolute Penalty Total}}{\text{Evaluation Word Count}}$
3	Overall Quality Score (OQS)	$(1 - \text{PWPT}) \times$ Maximum Score Value

Table 2: Overall Quality Score Calculation.

Dataset	Prompt Template
EN	Give a `{ToMi/FanToM/NegotiationToM}` Test example and its answer.
ZH	给出一个`{ToMi/FanToM/NegotiationToM}`测试示例及其答案。
DE	Geben Sie ein Beispiel für den `{ToMi/FanToM/NegotiationToM}`-Test und die dazugehörige Antwort.
FR	Donnez un exemple de test `{ToMi/FanToM/NegotiationToM}` et sa réponse.
JA	`{ToMi/FanToM/NegotiationToM}` テストの例とその解答を示します。
EN	You are provided with Sentence 1 from the train split of the `{ToMi/FanToM/NegotiationToM}` dataset. Finish Sentence 2 as appeared in the dataset. Sentence 2 must exactly match the instance in the dataset.Sentence 1: `<Sentence1-of-Example>`
ZH	你将获得来自`{ToMi/FanToM/NegotiationToM}`数据集训练集的句子. 1.请按照数据集中显示的方式完成句子. 2.句子2 必须与数据集中的实例完全匹配.句子1: `<Sentence1-of-Example>`
DE	Sie erhalten Satz 1 aus dem Train-Split des Datensatzes `{ToMi/FanToM/NegotiationToM}`. Vervollständigen Sie Satz 2 wie im Datensatz angegeben. Satz 2 muss exakt mit der Instanz im Datensatz übereinstimmen.Satz 1: `<Sentence1-of-Example>`.
FR	Vous disposez de la phrase 1 issue de la division du train du jeu de données `{ToMi/FanToM/NegotiationToM}`. Complétez la phrase 2 telle qu'elle apparaît dans le jeu de données. La phrase 2 doit correspondre exactement à l'instance du jeu de données.Phrase 1: `<Sentence1-of-Example>`.
JA	文1は`{ToMi/FanToM/NegotiationToM}`データセットの学みデタから提供されています。文2はデタセットに出した通りに完成させてください。文2はデタセット内のインスタンスと完全に一致する必要があります。文1: `<Sentence1-of-Example>`

Table 3: Prompt Template for Verification of Potential Contamination by following Li and Flanigan (2024) and Golchin and Surdeanu (2024).

answer, particularly the translation of containers.

In practice, annotators record the number of errors in each category for each data sample, and the MQM score is calculated using the scoring method outlined in Table 2. Maximum Score Value represents a perfect upper score on a scale, and the default value is one hundred. According to the MQM framework, we set the preset threshold score as 95 to determine whether a translated instance passes or fails. By categorizing translation quality into these dimensions, MQM provides a systematic and adaptable approach to evaluating both human and machine translations. This structured framework enhances the reliability and consistency of translation quality assessment across diverse use cases. To streamline the annotation process, we developed a software interface (shown in Figure 11) to assist annotators. The code will be released alongside the dataset.

A.4 Verification of Potential Contamination

Most of the existing available benchmarks in the NLP field were released prior to the initiation of the LLM training process, indicating that these datasets are likely to have been utilized during the pre-training phase and post-training phase

Dataset	XFANToM	XNegotiationToM	XToMi
Deepseek R1(EN)	0	0	30
Deepseek R1(ZH)	0	0	33
Deepseek R1(DE)	0	0	23
Deepseek R1(FR)	0	0	25
Deepseek R1(JA)	0	0	18
GPT 4o (EN)	0	0	36
GPT 4o (ZH)	0	0	34
GPT 4o (DE)	0	0	25
GPT 4o (FR)	0	0	9
GPT 4o (JA)	0	0	13

Table 4: Verification of Potential Contamination in 100 Independent Trials by following Li and Flanigan (2024).

Dataset	XFANToM	XNegotiationToM	XToMi
Deepseek R1(EN)	0	0	10
Deepseek R1(ZH)	0	0	4
Deepseek R1(DE)	0	0	0
Deepseek R1(FR)	0	0	0
Deepseek R1(JA)	0	0	0
GPT 4o (EN)	0	0	12
GPT 4o (ZH)	0	0	10
GPT 4o (DE)	0	0	2
GPT 4o (FR)	0	0	5
GPT 4o (JA)	0	0	8

Table 5: Verification of Potential Contamination in 100 unique stories by following Golchin and Surdeanu (2024).

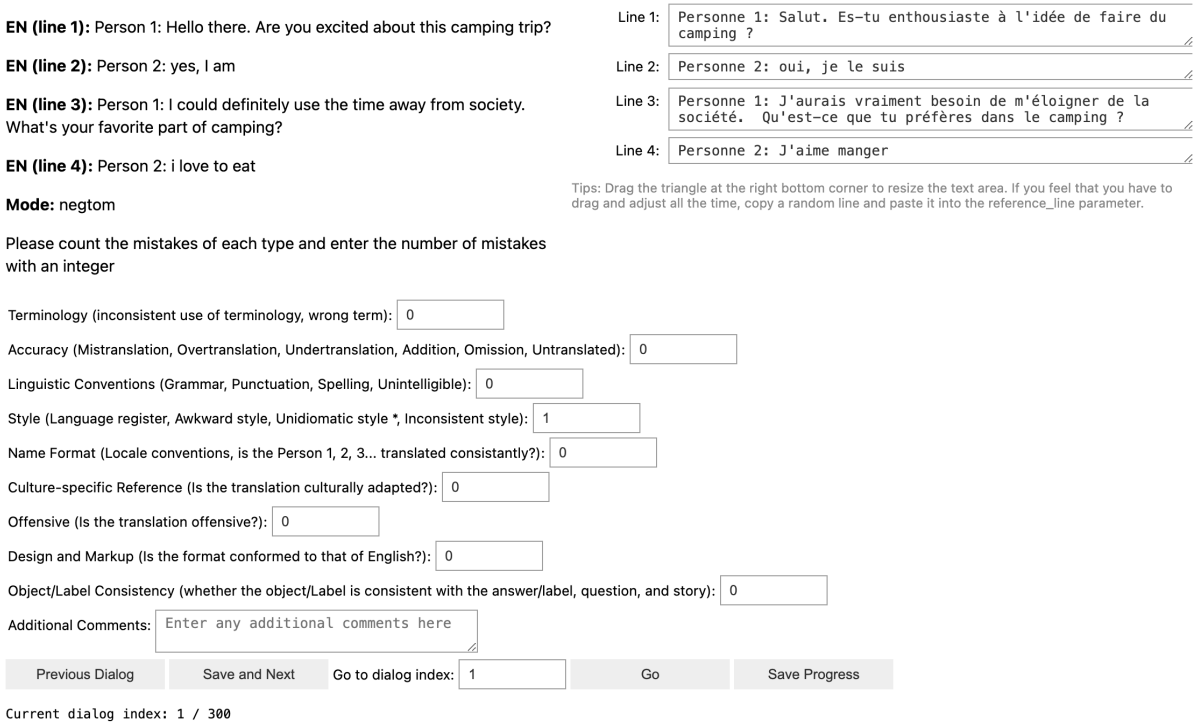


Figure 11: Interface for Human Correction and Validation.

(i.e., SFT (Ouyang et al., 2022) or RLHF (Christiano et al., 2017)) of LLMs (Golchin and Surdeanu, 2024; Li and Flanigan, 2024). Therefore, we follow the established protocols by prior works Golchin and Surdeanu (2024); Li and Flanigan (2024) to tailor two prompting templates in each language to assess the potential contamination issues in the sampled dataset. The prompt template is shown in table 3. For each sampled dataset, we conducted 100 independent trials or 100 unique stories per dataset using the prompting methods in Golchin and Surdeanu (2024) and Li and Flanigan (2024) by utilizing state-of-the-art LLMs (DeepSeek R1 and GPT-4o). These trials aimed to detect overlaps between model-generated outputs (stories, questions, answers) and the benchmark's ground-truth data. All responses underwent a stringent human evaluation to verify contamination issues in existing datasets. The results are reported in the following tables 4 and 5. For FANToM and NegotiationToM, none of the LLMs' generated responses matched with the ground truth by using two prompting methods across various languages, which indicates most instances of the XToM benchmark (i.e., two subtasks XFANToM and XNegotiationToM) are not identified as suffering from the data contamination issue, and the experimental results of the paper are reliable and valuable. Therefore, we believe data contamination may not be a primary concern in XToM. An interesting finding is that some LLM-generated responses match some ToMi story patterns or even data instances. However, to ensure the representativeness of the dataset, we still have to collect the ToMi for reference purposes, as ToMi is a classical ToM benchmark in the ToM field.

Dataset	Total#Questions	BeliefQA	1st.Belief	2nd.Belief	FactQA
XFANToM	3,090	1,590	1,085	505	1,500
Dataset	**Total#Questions**	**Belief**	**1st.Belief**	**2nd.Belief**	**Reality**
XToMi	3,235	1,735	615	1,120	1,500
Dataset	**Total#Questions**	**Belief**	**Desire**	**Intention**	-
XNegotiationToM	8,790	2,930	2,930	2,930	-

Table 6: Statistics of XTOM, which include XFANToM, XToMi, and XNegotiationToM.

Dataset	CH	DE	JA	FR	Avg.Dataset
XFANToM	96.59	93.63	95.11	96.55	96.59
XToMi	94.62	93.98	95.78	92.20	94.15
XNegotiationToM	95.62	97.92	95.61	94.63	96.27
Avg.Language	95.61	97.18	95.78	95.46	95.18

Table 7: Fleiss Kappa of XTOM across various tasks and languages.

B Appendix for Related Works

B.1 Related Works for Theory of Mind

In recent years, benchmarks for evaluating large language models' Theory of Mind capabilities have become a crucial area of artificial intelligence research. The roots of this work can be traced back to classic psychological experiments, such as the Sally-Anne test (Baron-Cohen et al., 1985), and have continually evolved to explore more complex aspects of ToM. Early significant contributions include the ToM-bAbi dataset (Grant et al., 2017), which focused on assessing false beliefs and was later improved by Le et al. (2019) into the more comprehensive ToMi dataset. These early works laid the foundation for subsequent research, prompting the emergence of more advanced benchmarks.

Researchers developed a series of more complex evaluation tools based on this foundation. For instance, T4D (Zhou et al., 2023a) specifically assesses N-ToM capabilities in AI assistants, while Hi-ToM (He et al., 2023) explores higher-order N-ToM concepts. These works significantly expanded our understanding of AI systems' ToM capabilities. Simultaneously, researchers also sought inspiration from classic human ToM tests, such as the Smarties test (Gopnik and Astington, 1988) and the Faux Pas test (Baron-Cohen et al., 1999). These efforts led to the creation of multiple new datasets, including ToMChallenges (Ma et al., 2023a), BigToM (Gandhi et al., 2023b), and FauxPas-EAI (Shapira et al., 2023b), each targeting different aspects of N-ToM assessment. Recognizing the need for more comprehensive evaluation methods, Jones et al. (2024) proposed EPITOME, a comprehensive benchmark that integrates various human ToM tests. Furthermore, researchers focused on assessing N-ToM capabilities in dialogue contexts, leading to benchmarks such as G-DRAGON (Zhou et al., 2022), FANToM (Kim et al., 2023), and SOTOPIA (Zhou et al., 2023b). These works greatly enriched our methods for evaluating AI systems' ToM capabilities.

More recently, several new benchmarks have emerged to address limitations in previous ToM evaluations, offering more realistic and challenging scenarios. TOMBench(Chen et al., 2024) provides a comprehensive framework with 31 social cognition abilities across 8 tasks. At the same time, OpenToM (Xu et al., 2024) focuses on natural narratives featuring characters with distinct personalities and intentions, testing models on psychological and physical mental states. MMToM-QA (Jin et al., 2024) introduces multimodal evaluation, combining video and text inputs to assess ToM reasoning, and NegotiationToM (Chan et al., 2024c) explores multi-dimensional mental states in real-world negotiation dialogues.

B.2 Related Works for Multilingual Capabilities of LLMs

Recent studies have thoroughly evaluated instruction-following large language models (LLMs) (OpenAI, 2023, 2022; Jiang et al., 2023c; Chan et al., 2023a), demonstrating their superior zero-shot performance across numerous tasks (Bubeck et al., 2023; Chan et al., 2024b; Cheng et al., 2023; Wang et al., 2024; Jiayang et al., 2024a; Shi et al., 2025; Jiayang et al., 2024b). However, significant challenges remain unaddressed, including complex mathematical (Frieder et al., 2023), theory of mind reasoning (Lin et al., 2024), uncertainty and confidence calibration (Zong et al., 2025), retrieval-augmented generation (Jiayang et al., 2025; Zhang et al., 2025; Li et al., 2025a), intention reasoning (Yang et al., 2025b), analogical reasoning (Cheng et al., 2023), discourse relation classification (Chan et al., 2023b), text-to-table generation (Deng et al., 2024, 2025), complex game scenarios (Yim et al., 2024; Mo et al., 2025; Liang et al., 2025), argument impact classification (Chan et al., 2024a), and associated ethical and privacy concerns (Li et al., 2023; Susnjak, 2022; Li et al., 2024b; Lukas et al., 2023; Li et al., 2024a). Therefore, it is crucial to investigate whether large language models possess the theory of mind capabilities across languages.

State-of-the-art large language models (LLMs), such as GPT-4o-11-20 (Hurst et al., 2024), LLaMA (Dubey et al., 2024), Mistral (Jiang et al., 2023b), and DeepSeek (DeepSeek-AI et al., 2025), are often described as multilingual. They are pre-trained on a mixture of texts in multiple languages, leveraging linguistic similarities and shared representations (Zeng et al., 2025; Wendler et al., 2024; Dumas et al., 2025) to enhance performance in lower-resource languages. Recent studies have demonstrated that LLMs can develop cross-lingual capabilities, transferring knowledge and skills learned in one language to others, even those with limited training data (Chirkova and Nikoulina, 2024; Pires et al., 2019; Wu and Dredze, 2019). However, a significant performance gap persists across

languages, and researchers are actively exploring methods to quantify and mitigate this disparity (Li et al., 2024c; Kumar et al., 2024; Zeng et al., 2024).

Multilingual models have exhibited strong reasoning abilities across languages, including those that are underrepresented (Shi et al., 2022). For instance, Qin et al. (2023) proposed prompting models with chain-of-thought reasoning to solve tasks in multiple languages, then ensembling reasoning paths across languages to derive more accurate answers. Similarly, She et al. (2024) introduced preference optimization techniques to align reasoning processes in non-dominant languages with those in dominant ones, thereby enhancing multilingual reasoning capabilities. Nevertheless, while LLMs demonstrate promising surface-level cross-lingual abilities in tasks such as machine translation and embedding space alignment, they still face challenges in deeper cross-lingual knowledge transfer (Chua et al., 2024).

C Appendix For Experimental Setting

C.1 Hyperparameter

We use the following hyperparameters for assessing the large language models mentioned in this paper. For ChatGPT (`gpt-3.5-turbo-0125`) and GPT-4o (`gpt-4o-1120`), the default parameters[6] are `temperature`=1 and `top_p`=1. For Llama-3.1-8b-instruct and Llama-3.3-70b-instruct models, we follow the default setting where `temperature`=0.7, `top_p`=0.9. For DeepSeek R1, the parameters are `temperature`=0.5, `top_p`=0.7. For Qwen-3-235b-a22b, Qwen-2.5-7b-instruct and Qwen-2.5-72b-instruct, the parameters are `temperature`=0.5, `top_p`=0.7. For mistral-7b-instruct-v0.3 and mixtral-8x22b-instruct-v0.1, the parameters are `temperature`=0.2, `top_p`=0.7.

C.2 Appendix for Prompt Template

In this study, there are two types of prompting methods. The first one is zero-shot prompting, and we utilize the prompt template in the original paper. Another one is the chain of thought (CoT) prompting method, which follows Wei et al. (2022) and use the prompt "let's think step by step." The zero-shot prompt template and data example for XNegotiationToM are presented in Tables 12, 13, 14, 15, and 16. The zero-shot prompt template and data example of XToMi are illustrated in 17 while XFANToM is shown in Table 18. The COT prompting template is formed by appending the "let's think step by step" to the zero-shot prompt template.

[6]https://platform.openai.com/docs/api-reference/chat/create

C.3 Appendix for Fine-Tuning LLMs

We employ the LLaMA-Factory (Zheng et al., 2024) to fine-tune LLMs using Low-Rank Adaptation (LoRA) (Hu et al., 2022) to achieve parameter-efficient training. The batch size is set to 64, and the learning rate is 5e-5. The LoRA rank is 8, with α set to 32. The maximum length for the input is set to 4,096. We set the warm-up ratio to 0.1, trained the model with 3 epochs, and evaluated the model every 100 steps. A Cosine scheduler is also used.

D Appendix for Experimental Result

D.1 Appendix for Main Experimental Result

Figures 12, 13, 14, 15, and 16 display all LLMs' performance of the XFANToM, XToMi, and XNegotiatioNToM. More detailed performance (i.e., numerical value of performance) of the XFANToM, XToMi, and XNegotiatioNToM are reported in Tables 8, 9, and 10. Table 10 includes the LLMs' performance on belief, desire, and intention dimensions of mental states across five languages in XNegotiatioNToM. For the desire and intention dimension of ToM, we observe a similar trend where Language variation generally impacts model performance. Moreover, Larger models consistently outperform their smaller models across all three sub-tasks. Smaller models, such as Mistral-7b-instruct-v0.3 and Llama-3.1-8b-instruct, exhibit significantly lower performance, suggesting that larger-scale training enhances ToM capabilities. Furthermore, Table 8 provided detailed LLMs' performance on belief questions (including first-order belief and second-order belief) and fact questions in XFANToM, while Table 9 displayed the belief question (including first-order belief and second-order belief) and reality question in XToMi.

D.2 Appendix for Consistency Analysis

The consistency analysis of ToMi is shown in Figure 19, revealing a similar trend to FANToM. As ToMi is a relatively easier task, models achieve higher overall consistency. For fact questions, models demonstrate high consistency, with nearly zero consistently false answers. However, for belief questions, consistency drops significantly, further

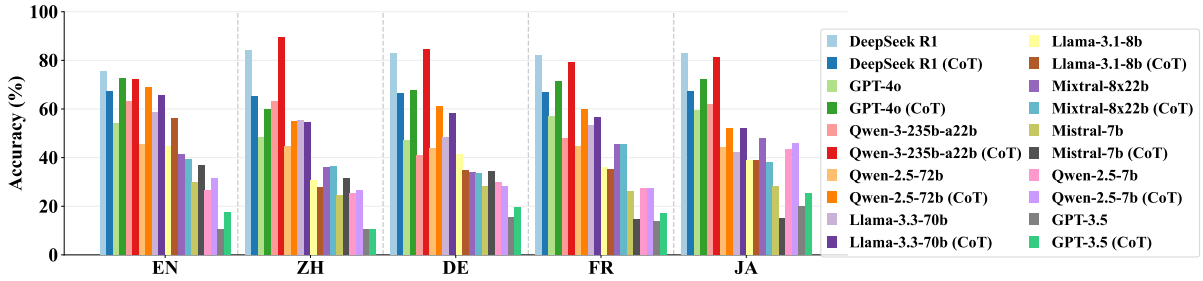


Figure 12: Performance comparison of different models on false belief questions across languages in XFANToM.

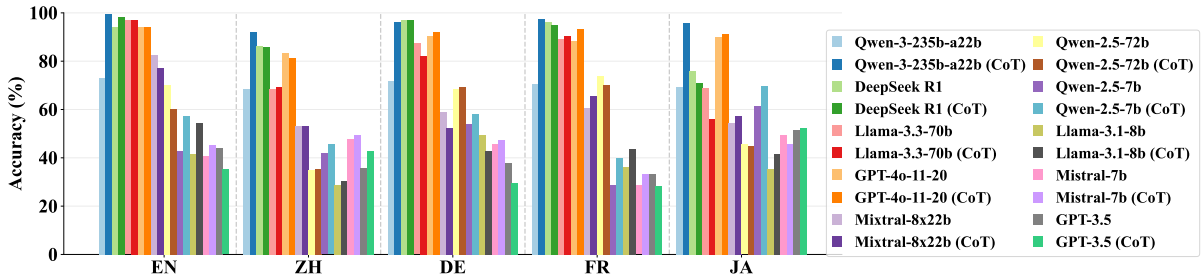


Figure 13: Performance comparison of different models on false belief questions across languages in XToMi.

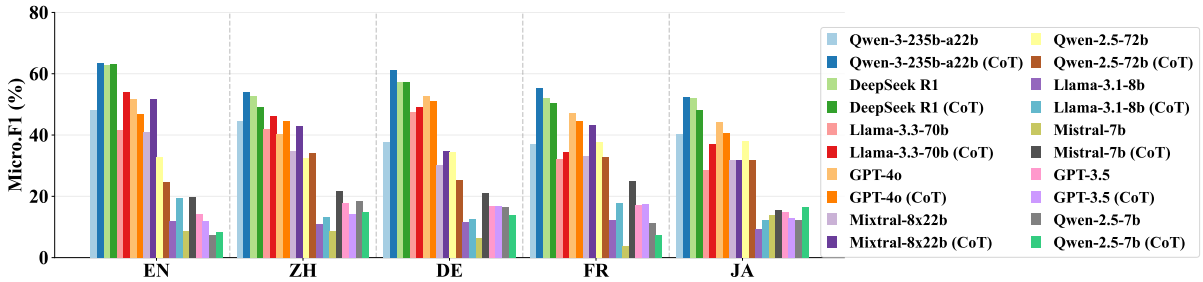


Figure 14: Performance comparison of different models on belief choice across languages in XNegotiationToM.

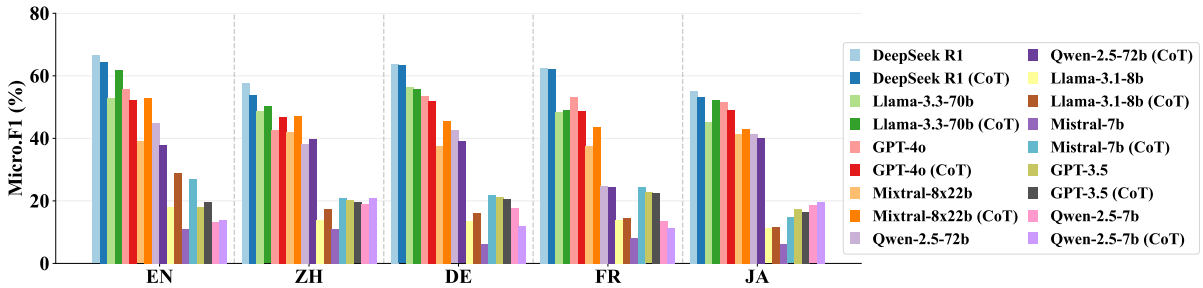


Figure 15: Performance comparison of different models on desire choice across languages in XNegotiationToM.

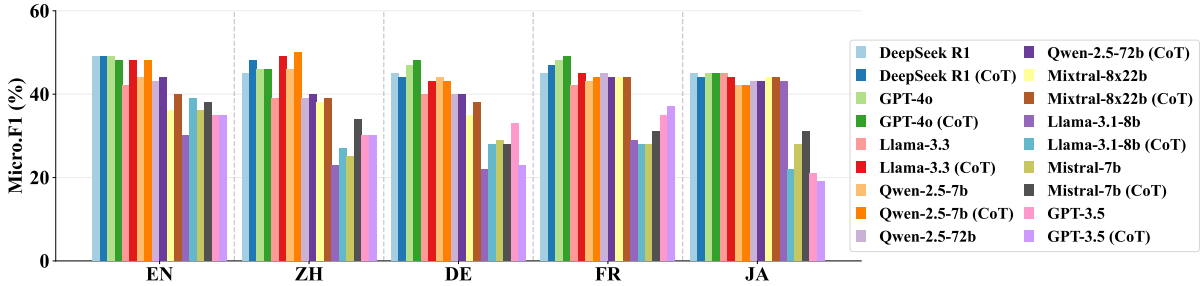


Figure 16: Performance comparison of different models on intention choice across languages in XNegotiationToM.

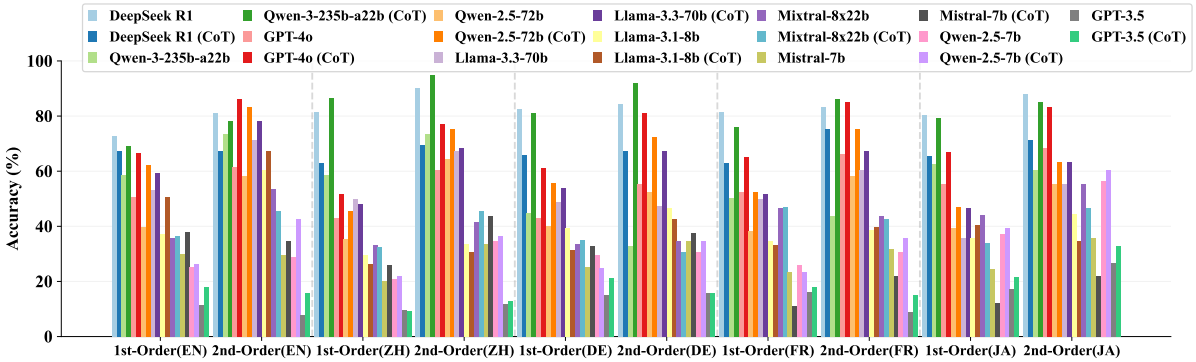


Figure 17: Performance comparison of different models on first order and second order belief choice across languages in XFANToM.

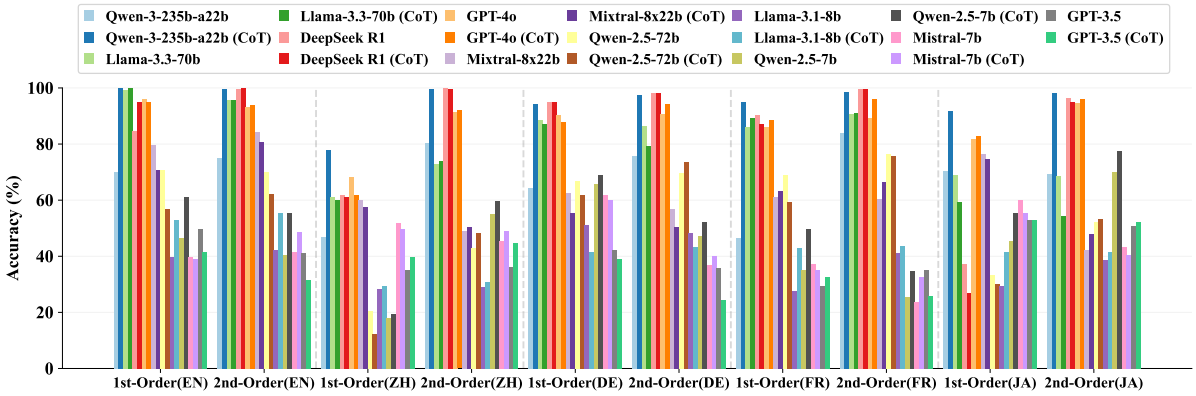


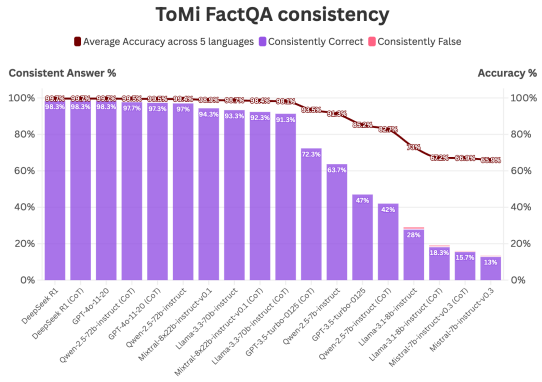
Figure 18: Performance comparison of different models on first order and second order belief choice across languages in XToMi.

emphasizing the models' weaker ToM reasoning compared to fact retrieval.

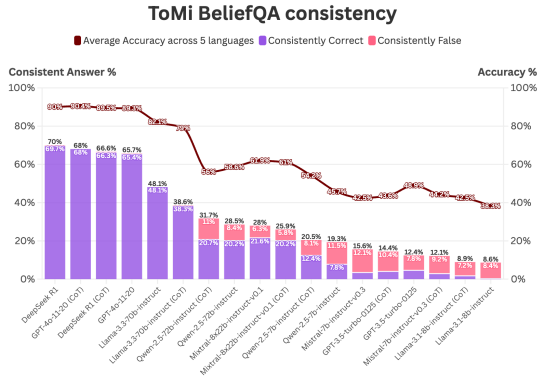
D.3 Error Analysis and Case Study

Error Analysis. The error analysis of different models in XFANToM is presented in Figure 20. We identified several distinct error types and conducted a statistical analysis of their distributions across the models. These error categories are defined as follows:(1) *Theory of Mind Reasoning Error*: The model provides a correct prediction for the fact-based question but an incorrect response for the Theory of Mind (ToM) question, indicating a specific deficit in ToM reasoning. (2) *Language Understanding Error*: The model produces incorrect predictions for both the fact-based question and the ToM question, suggesting broader challenges in language comprehension. (3) *Fact Understanding Error*: The model gives an incorrect prediction for the fact-based question but a correct response for the ToM question, pointing to issues in factual understanding rather than ToM reasoning. (4)*Irrelevant Response Error*: The model generates responses that are unrelated to the posed questions (e.g., repeating the question without providing an answer), reflecting a failure to produce meaningful output. As shown in Figure 20, the results reveal that all models exhibit a significant proportion of ToM reasoning errors. This finding suggests that large language models (LLMs) struggle with Theory of Mind reasoning abilities, reflecting inherent limitations in social cognition rather than deficiencies in language understanding. Consistent with prior research (Shapira et al., 2024), LLMs' proficiency in statistical language patterns does not translate to human-like social reasoning capabilities.

Case Study. Moreover, we also provide a case study of XFANToM in Table 19, investigating the cultural context embedded in language influencing the model's performance. The case study in Table 19 illustrates the phenomenon: when festival culture discussions from the FANToM English version were translated to other languages, models struggled to maintain performance. This degradation likely occurs because cultural references that



(a)



(b)

Figure 19: ToMi consistency analysis across different languages.

were salient in the source language became less accessible in target languages, revealing reporting biases (Gordon and Durme, 2013) in how models process culturally embedded mental state information across languages.

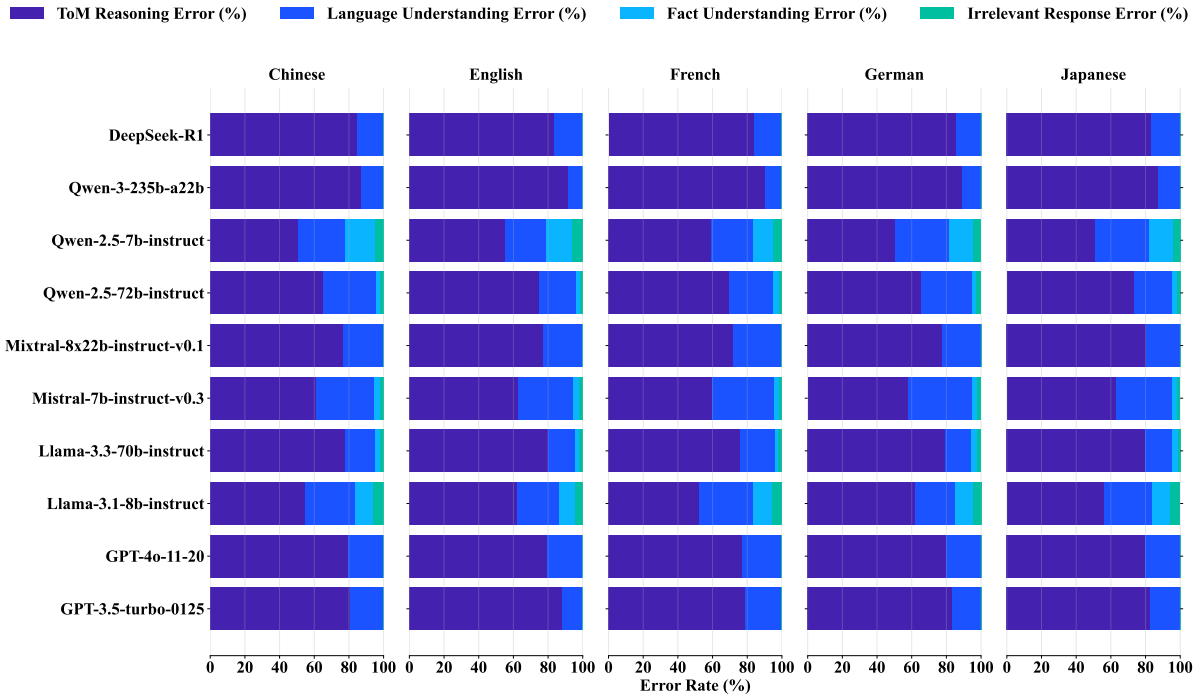


Figure 20: Error analysis of various models on belief questions in XFANToM across languages. Some LLMs made zero Fact Understanding or Irrelevant Response errors and are thus not visible in the corresponding sections of the figure.

	Model	Language	Belief Questions					Fact Questions
			Choice (%)	First-Order (%)	Second-Order (%)	Acyclic (%)	Cyclic (%)	Choice (%)
Short Conversation	GPT-3.5-turbo-0125	EN	10.4	11.5	7.9	7.8	8.0	86.7
	GPT-3.5-turbo-0125 (CoT)		17.3	18.0	15.8	21.6	10.0	86.0
	GPT-4o-11-20		54.1	50.7	61.4	60.8	62.0	89.7
	GPT-4o-11-20 (CoT)		72.6	66.4	**86.1**	**90.2**	82.0	90.7
	DeepSeek R1		**75.5**	**72.8**	81.2	88.2	74.0	**93.3**
	DeepSeek R1 (CoT)		67.3	67.3	67.3	66.7	68.0	84.7
	Qwen-2.5-7b-instruct		26.4	25.3	28.7	31.4	26.0	83.0
	Qwen-2.5-7b-instruct (CoT)		31.4	26.3	42.6	47.1	38.0	85.3
	Qwen-2.5-72b-instruct		45.6	39.6	58.4	58.8	58.0	87.7
	Qwen-2.5-72b-instruct (CoT)		68.9	62.2	83.2	88.2	78.0	85.7
	Qwen-3-235b-a22b		63.2	58.5	73.3	70.6	76.0	92.3
	Qwen-3-235b-a22b (CoT)		72.0	69.1	78.2	80.4	76.0	90.7
	Llama-3.1-8b-instruct		44.7	37.3	60.4	62.7	58.0	83.0
	Llama-3.1-8b-instruct (CoT)		56.0	50.7	67.3	68.6	66.0	85.3
	Llama-3.3-70b-instruct		58.8	53.0	71.3	68.6	74.0	95.0
	Llama-3.3-70b-instruct (CoT)		65.4	59.4	78.2	80.4	76.0	89.7
	Mistral-7b-instruct-v0.3		29.9	30.0	29.7	27.5	32.0	81.7
	Mistral-7b-instruct-v0.3 (CoT)		36.8	37.8	34.7	35.3	34.0	76.3
	Mixtral-8x22b-instruct-v0.1		41.5	35.9	53.5	49.0	58.0	86.7
	Mixtral-8x22b-instruct-v0.1 (CoT)		39.3	36.4	45.5	43.1	48.0	87.7
	GPT-3.5-turbo-0125	ZH	10.4	9.7	11.9	9.8	14.0	80.7
	GPT-3.5-turbo-0125 (CoT)		10.4	9.2	12.9	9.8	16.0	80.3
	GPT-4o-11-20		48.4	42.9	60.4	54.9	66.0	89.7
	GPT-4o-11-20 (CoT)		59.7	51.6	77.2	78.4	76.0	90.0
	DeepSeek R1		**84.3**	81.6	90.1	86.3	94.0	**96.0**
	DeepSeek R1 (CoT)		65.1	63.1	69.3	70.6	68.0	85.0
	Qwen-2.5-7b-instruct		25.2	20.7	34.70	31.4	38.0	81.3
	Qwen-2.5-7b-instruct (CoT)		85.8	82.9	92.1	88.2	96.0	94.3
	Qwen-2.5-72b-instruct		44.7	35.5	64.4	62.7	66.0	85.3
	Qwen-2.5-72b-instruct (CoT)		55.0	45.6	75.2	70.6	80.0	84.0
	Qwen-3-235b-a22b		63.2	58.5	73.3	70.6	76.0	92.3
	Qwen-3-235b-a22b (CoT)		89.3	86.6	95.0	94.1	96.0	96.3
	Llama-3.1-8b-instruct		30.8	29.5	33.7	33.3	34.0	74.0
	Llama-3.1-8b-instruct (CoT)		27.7	26.3	30.7	29.4	32.0	72.3
	Llama-3.3-70b-instruct		55.3	49.8	67.3	64.7	70.0	84.3
	Llama-3.3-70b-instruct (CoT)		54.4	47.9	68.3	66.7	70.0	76.7
	Mistral-7b-instruct-v0.3		24.5	20.3	33.7	31.4	36.0	76.7
	Mistral-7b-instruct-v0.3 (CoT)		31.4	25.8	43.6	35.3	52.0	71.0
	Mixtral-8x22b-instruct-v0.1		35.8	33.2	41.6	31.4	52.0	85.7
	Mixtral-8x22b-instruct-v0.1 (CoT)		36.5	32.3	45.5	35.3	56.0	83.7
	GPT-3.5-turbo-0125	DE	15.4	15.2	15.8	11.8	20.0	85.0
	GPT-3.5-turbo-0125 (CoT)		19.5	21.2	15.8	11.8	20.0	86.3
	GPT-4o-11-20		46.9	42.9	55.4	60.8	50.0	90.3
	GPT-4o-11-20 (CoT)		67.6	61.3	81.2	80.4	82.0	88.3
	DeepSeek R1		**83.0**	82.5	84.2	82.4	86.0	**95.7**
	DeepSeek R1 (CoT)		66.4	65.9	67.3	68.6	66.0	81.3
	Qwen-2.5-7b-instruct		29.9	29.5	30.7	19.6	42.0	86.3
	Qwen-2.5-7b-instruct (CoT)		28.0	24.9	34.7	29.4	40.0	80.7
	Qwen-2.5-72b-instruct		44.0	40.1	52.5	51.0	54.0	87.3
	Qwen-2.5-72b-instruct (CoT)		61.0	55.8	72.3	68.6	76.0	86.3
	Qwen-3-235b-a22b		40.9	44.7	32.7	21.6	44.0	93.3
	Qwen-3-235b-a22b (CoT)		84.6	81.1	92.1	90.2	94.0	97.3
	Llama-3.1-8b-instruct		41.5	39.2	46.5	52.9	40.0	75.7
	Llama-3.1-8b-instruct (CoT)		34.9	31.3	42.6	39.2	46.0	78.3
	Llama-3.3-70b-instruct		48.4	48.8	47.5	49.0	46.0	90.7
	Llama-3.3-70b-instruct (CoT)		58.2	53.9	67.3	68.6	66.0	90.3
	Mistral-7b-instruct-v0.3		28.3	25.3	34.7	33.3	36.0	77.7
	Mistral-7b-instruct-v0.3 (CoT)		34.3	32.7	37.6	37.3	38.0	66.7
	Mixtral-8x22b-instruct-v0.1		34.0	33.6	34.7	31.4	38.0	87.0
	Mixtral-8x22b-instruct-v0.1 (CoT)		33.6	35.0	30.7	27.5	34.0	86.7
	GPT-3.5-turbo-0125	FR	13.8	16.1	8.9	3.9	14.0	85.0
	GPT-3.5-turbo-0125 (CoT)		17.0	18.0	14.9	11.8	18.0	83.0
	GPT-4o-11-20		56.9	52.5	66.3	72.5	60.0	87.7
	GPT-4o-11-20 (CoT)		71.4	65.0	85.1	86.3	84.0	90.3
	DeepSeek R1		**82.1**	81.6	83.2	80.4	86.0	**96.0**
	DeepSeek R1 (CoT)		67.0	63.1	75.2	78.4	72.0	93.0
	Qwen-2.5-7b-instruct		27.4	25.8	30.7	23.5	38.0	82.7
	Qwen-2.5-7b-instruct (CoT)		27.4	23.5	35.6	29.4	42.0	81.7
	Qwen-2.5-72b-instruct		44.7	38.2	58.4	62.7	54.0	86.3
	Qwen-2.5-72b-instruct (CoT)		59.7	52.5	75.2	80.4	70.0	85.0
	Qwen-3-235b-a22b		48.1	50.2	43.6	37.3	50.0	93.7
	Qwen-3-235b-a22b (CoT)		79.2	76.0	86.1	82.4	90.0	95.7
	Llama-3.1-8b-instruct		35.8	34.6	38.6	45.1	32.0	76.7
	Llama-3.1-8b-instruct (CoT)		35.2	33.2	39.6	35.3	44.0	75.0
	Llama-3.3-70b-instruct		53.1	49.8	60.4	58.8	62.0	89.0
	Llama-3.3-70b-instruct (CoT)		56.6	51.6	67.3	66.7	68.0	88.3
	Mistral-7b-instruct-v0.3		26.1	23.5	31.7	31.4	32.0	76.7
	Mistral-7b-instruct-v0.3 (CoT)		14.5	11.1	21.8	21.6	22.0	30.7
	Mixtral-8x22b-instruct-v0.1		45.6	46.5	43.6	39.2	48.0	86.7
	Mixtral-8x22b-instruct-v0.1 (CoT)		45.6	47.0	42.6	41.2	44.0	83.7
	GPT-3.5-turbo-0125	JA	20.1	17.1	26.7	27.5	26.0	81.0
	GPT-3.5-turbo-0125 (CoT)		25.2	21.7	32.7	27.5	38.0	55.7
	GPT-4o-11-20		59.4	55.3	68.3	72.5	64.0	91.7
	GPT-4o-11-20 (CoT)		72.0	66.8	83.2	86.3	80.0	90.7
	DeepSeek R1		**82.7**	80.2	88.1	90.2	86.0	**97.3**
	DeepSeek R1 (CoT)		67.3	65.4	71.3	70.6	72.0	83.3
	Qwen-2.5-7b-instruct		43.4	37.3	56.4	51.0	62.0	80.3
	Qwen-2.5-7b-instruct (CoT)		45.9	39.2	60.4	56.9	64.0	76.3
	Qwen-2.5-72b-instruct		44.3	39.2	55.4	47.1	64.0	86.7
	Qwen-2.5-72b-instruct (CoT)		52.2	47.0	63.4	58.8	68.0	86.3
	Qwen-3-235b-a22b		61.9	62.7	60.4	54.9	66.0	95.0
	Qwen-3-235b-a22b (CoT)		81.1	79.3	85.1	84.3	86.0	94.3
	Llama-3.1-8b-instruct		38.7	35.9	44.6	47.1	42.0	74.7
	Llama-3.1-8b-instruct (CoT)		38.7	40.6	34.7	37.3	32.0	73.7
	Llama-3.3-70b-instruct		42.1	35.9	55.4	56.9	54.0	91.7
	Llama-3.3-70b-instruct (CoT)		51.9	46.5	63.4	66.7	60.0	87.0
	Mistral-7b-instruct-v0.3		28.0	24.4	35.6	35.3	36.0	75.3
	Mistral-7b-instruct-v0.3 (CoT)		15.1	12.0	21.8	21.6	22.0	41.0
	Mixtral-8x22b-instruct-v0.1		47.8	44.2	55.4	47.1	64.0	86.0
	Mixtral-8x22b-instruct-v0.1 (CoT)		38.1	34.1	46.5	47.1	46.0	70.7

Table 8: Results of models for XFANToM (short conversation).

Model	Language	1st Belief Accuracy(%)	2nd Belief Accuracy(%)	Average Accuracy(%)	Reality Accuracy(%)
GPT-3.5-turbo-0125	EN	49.59	41.07	44.09	92.00
GPT-3.5-turbo-0125 (CoT)		41.46	31.70	35.16	85.33
GPT-4o-11-20		96.02	93.30	94.24	**100.00**
GPT-4o-11-20 (CoT)		95.12	93.75	94.24	**100.00**
DeepSeek R1		84.55	99.55	94.24	99.67
DeepSeek R1 (CoT)		95.12	**100.00**	**98.27**	**100.00**
Qwen-2.5-7b-instrcut		46.34	40.63	42.65	99.33
Qwen-2.5-7b-instrcut (CoT)		60.98	55.36	57.35	96.67
Qwen-2.5-72b-instrcut		70.73	70.09	70.32	**100.00**
Qwen-2.5-72b-instrcut (CoT)		56.91	62.05	60.23	**100.00**
Qwen-3-235b-a22b		70.05	75.00	73.20	**100.00**
Qwen-3-235b-a22b (CoT)		100.00	99.55	99.71	**100.00**
Llama-3.1-8b-instruct		39.84	42.41	41.50	77.00
Llama-3.1-8b-instruct (CoT)		52.85	55.36	54.47	67.00
Llama-3.3-70b-instruct		99.19	95.54	96.83	**100.00**
Llama-3.3-70b-instruct (CoT)		**100.00**	95.54	97.12	**100.00**
Mistral-7b-instruct-v0.3		39.84	41.52	40.92	93.00
Mistral-7b-instruct-v0.3 (CoT)		39.02	48.66	45.24	93.00
Mixtral-8x22b-instruct-v0.1		79.67	84.38	82.71	**100.00**
Mixtral-8x22b-instruct-v0.1 (CoT)		70.73	80.80	77.23	**100.00**
GPT-3.5-turbo-0125	ZH	34.96	36.16	35.73	85.67
GPT-3.5-turbo-0125 (CoT)		39.84	44.64	42.94	78.67
GPT-4o-11-20		**68.29**	91.52	83.29	**100.00**
GPT-4o-11-20 (CoT)		61.79	91.96	81.27	**100.00**
DeepSeek R1		61.79	**100.00**	**86.46**	**100.00**
DeepSeek R1 (CoT)		60.98	99.55	85.88	**100.00**
Qwen-2.5-7b-instruct		17.89	54.91	41.79	97.00
Qwen-2.5-7b-instruct (CoT)		19.51	59.82	45.53	87.00
Qwen-2.5-72b-instruct		20.33	42.86	34.87	**100.00**
Qwen-2.5-72b-instruct (CoT)		12.20	48.21	35.45	**100.00**
Qwen-3-235b-a22b		46.72	80.36	68.30	**100.00**
Qwen-3-235b-a22b (CoT)		78.02	99.55	91.93	**100.00**
Llama-3.1-8b-instruct		28.46	29.02	28.82	83.00
Llama-3.1-8b-instruct (CoT)		29.27	30.80	30.26	73.67
Llama-3.3-70b-instruct		60.98	72.77	68.59	99.67
Llama-3.3-70b-instruct (CoT)		60.16	74.11	69.16	**100.00**
Mistral-7b-instruct-v0.3		52.03	45.54	47.84	57.33
Mistral-7b-instruct-v0.3 (CoT)		49.59	49.11	49.28	64.00
Mixtral-8x22b-instruct-v0.1		60.16	49.11	53.03	**100.00**
Mixtral-8x22b-instruct-v0.1 (CoT)		57.72	50.45	53.03	**100.00**
GPT-3.5-turbo-0125	DE	42.28	35.71	38.04	86.33
GPT-3.5-turbo-0125 (CoT)		39.02	24.55	29.68	83.00
GPT-4o-11-20		90.24	90.63	90.49	**98.67**
GPT-4o-11-20 (CoT)		87.80	94.20	91.93	97.67
DeepSeek R1		**95.12**	**98.21**	**97.12**	**98.67**
DeepSeek R1 (CoT)		**95.12**	**98.21**	**97.12**	98.33
Qwen-2.5-7b-instruct		65.85	47.32	53.89	87.00
Qwen-2.5-7b-instruct (CoT)		69.11	52.23	58.21	71.00
Qwen-2.5-72b-instruct		66.67	69.64	68.59	98.00
Qwen-2.5-72b-instruct (CoT)		61.79	73.66	69.45	98.00
Qwen-3-235b-a22b		64.30	75.89	71.76	97.00
Qwen-3-235b-a22b (CoT)		94.34	97.32	96.25	98.00
Llama-3.1-8b-instruct		51.22	48.21	49.28	71.67
Llama-3.1-8b-instruct (CoT)		41.46	43.30	42.65	68.00
Llama-3.3-70b-instruct		88.62	86.61	87.32	96.33
Llama-3.3-70b-instruct (CoT)		87.00	79.46	82.13	93.33
Mistral-7b-instruct-v0.3		61.79	37.05	45.82	57.33
Mistral-7b-instruct-v0.3 (CoT)		60.16	40.18	47.26	58.67
Mixtral-8x22b-instruct-v0.1		62.60	56.70	58.79	97.67
Mixtral-8x22b-instruct-v0.1 (CoT)		55.28	50.45	52.16	96.67
GPT-3.5-turbo-0125	FR	29.27	35.27	33.14	96.00
GPT-3.5-turbo-0125 (CoT)		32.52	25.89	28.24	93.00
GPT-4o-11-20		86.18	89.29	88.18	99.67
GPT-4o-11-20 (CoT)		88.62	95.98	93.37	99.67
DeepSeek R1		**90.24**	**99.55**	**96.25**	**100.00**
DeepSeek R1 (CoT)		86.99	**99.55**	95.10	**100.00**
Qwen-2.5-7b-instruct		34.96	25.45	28.82	99.00
Qwen-2.5-7b-instruct (CoT)		49.59	34.82	40.06	97.00
Qwen-2.5-72b-instruct		69.11	76.34	73.78	99.33
Qwen-2.5-72b-instruct (CoT)		59.35	75.89	70.03	99.67
Qwen-3-235b-a22b		46.54	83.93	70.61	98.33
Qwen-3-235b-a22b (CoT)		95.01	98.66	97.41	98.67
Llama-3.1-8b-instruct		27.64	41.07	36.31	68.33
Llama-3.1-8b-instruct (CoT)		43.09	43.75	43.52	70.67
Llama-3.3-70b-instruct		86.18	90.63	89.05	99.33
Llama-3.3-70b-instruct (CoT)		89.43	91.07	90.49	99.33
Mistral-7b-instruct-v0.3		37.40	23.66	28.53	89.67
Mistral-7b-instruct-v0.3 (CoT)		34.96	32.59	33.43	84.00
Mixtral-8x22b-instruct-v0.1		60.98	60.27	60.52	99.33
Mixtral-8x22b-instruct-v0.1 (CoT)		63.41	66.52	65.42	99.67
GPT-3.5-turbo-0125	JA	52.85	50.89	51.59	65.00
GPT-3.5-turbo-0125 (CoT)		52.85	52.23	52.45	66.00
GPT-4o-11-20		81.71	94.64	90.20	**100.00**
GPT-4o-11-20 (CoT)		82.93	95.98	91.35	**100.00**
DeepSeek R1		37.2	96.40	76.08	**100.00**
DeepSeek R1 (CoT)		26.83	95.09	70.89	**100.00**
Qwen-2.5-7b-instruct		45.53	70.09	61.38	74.00
Qwen-2.5-7b-instruct (CoT)		55.28	77.68	69.74	61.67
Qwen-2.5-72b-instruct		33.33	52.23	45.53	99.67
Qwen-2.5-72b-instruct (CoT)		30.08	53.13	44.96	**100.00**
Qwen-3-235b-a22b		70.52	69.20	69.45	99.00
Qwen-3-235b-a22b (CoT)		**91.84**	**98.21**	**95.97**	98.67
Llama-3.1-8b-instruct		29.27	38.84	35.45	65.00
Llama-3.1-8b-instruct (CoT)		41.46	41.52	41.50	56.67
Llama-3.3-70b-instruct		69.11	68.75	68.88	98.00
Llama-3.3-70b-instruct (CoT)		59.35	54.46	56.20	98.00
Mistral-7b-instruct-v0.3		60.16	43.30	49.28	32.00
Mistral-7b-instruct-v0.3 (CoT)		55.28	40.63	45.82	35.00
Mixtral-8x22b-instruct-v0.1		76.42	42.41	54.47	97.33
Mixtral-8x22b-instruct-v0.1 (CoT)		74.80	47.77	57.35	95.67

Table 9: Results of models for XToMi.

Model	Language	Belief	Desire	Intention	
		Exact.Match.(%)	Exact.Match.(%)	Micro.F1(%)	Macro.F1(%)
GPT-3.5-turbo-0125	EN	14.33	18.17	35.00	27.00
GPT-3.5-turbo-0125 (CoT)		11.83	19.67	35.00	24.00
GPT-4o-11-20		51.70	55.80	**49.00**	**42.00**
GPT-4o-11-20 (CoT)		46.70	52.30	48.00	41.00
DeepSeek R1		62.83	**66.50**	**49.00**	41.00
DeepSeek R1 (CoT)		**63.00**	64.50	**49.00**	41.00
Qwen-2.5-7b-instrcut		7.34	13.33	44.00	35.00
Qwen-2.5-7b-instrcut(CoT)		8.34	14.00	48.00	38.00
Qwen-2.5-72b-instrcut		32.84	44.84	43.00	35.00
Qwen-2.5-72b-instrcut(CoT)		24.67	37.84	44.00	37.00
Qwen-3-235b-a22b		48.17	54.17	44.00	37.00
Qwen-3-235b-a22b (CoT)		63.33	66.67	50.00	43.00
Llama-3.1-8b-instruct		12.00	18.17	30.00	24.00
Llama-3.1-8b-instruct (CoT)		19.50	28.84	39.00	29.00
Llama-3.3-70b-instruct		41.50	52.84	42.00	35.00
Llama-3.3-70b-instruct (CoT)		54.00	61.84	48.00	40.00
Mistral-7b-instruct-v0.3		8.67	11.00	36.00	28.00
Mistral-7b-instruct-v0.3 (CoT)		19.84	27.00	38.00	28.00
Mixtral-8x22b-instruct-v0.1		40.84	39.00	36.00	30.00
Mixtral-8x22b-instruct-v0.1 (CoT)		51.67	53.00	40.00	33.00
GPT-3.5-turbo-0125	ZH	17.76	20.34	30.00	22.00
GPT-3.5-turbo-0125 (CoT)		14.14	19.48	30.00	22.00
GPT-4o-11-20		40.35	42.59	46.00	39.00
GPT-4o-11-20 (CoT)		44.48	46.73	46.00	40.00
DeepSeek R1		**52.59**	**57.76**	45.00	**43.00**
DeepSeek R1 (CoT)		48.97	53.97	48.00	40.00
Qwen-2.5-7b-instrcut		18.28	18.97	46.00	34.00
Qwen-2.5-7b-instrcut(CoT)		15.00	21.03	50.00	36.00
Qwen-2.5-72b-instrcut		32.59	38.28	39.00	34.00
Qwen-2.5-72b-instrcut(CoT)		34.14	39.83	40.00	34.00
Qwen-3-235b-a22b		44.48	47.24	45.00	38.00
Qwen-3-235b-a22b (CoT)		54.14	58.79	45.00	38.00
Llama-3.1-8b-instruct		11.03	13.97	23.00	18.00
Llama-3.1-8b-instruct (CoT)		13.28	17.42	27.00	21.00
Llama-3.3-70b-instruct		42.00	48.67	39.00	35.00
Llama-3.3-70b-instruct (CoT)		46.21	50.17	**49.00**	41.00
Mistral-7b-instruct-v0.3		8.67	11.00	25.00	18.00
Mistral-7b-instruct-v0.3 (CoT)		21.72	21.03	34.00	25.00
Mixtral-8x22b-instruct-v0.1		34.67	42.00	38.00	28.00
Mixtral-8x22b-instruct-v0.1 (CoT)		42.76	47.25	39.00	32.00
GPT-3.5-turbo-0125	DE	16.67	21.33	33.00	25.00
GPT-3.5-turbo-0125 (CoT)		16.83	20.50	23.00	15.00
GPT-4o-11-20		52.84	53.50	47.00	39.00
GPT-4o-11-20 (CoT)		51.17	51.84	**48.00**	**41.00**
DeepSeek R1		**57.33**	**63.67**	45.00	37.00
DeepSeek R1 (CoT)		**57.33**	63.33	44.00	36.00
Qwen-2.5-7b-instrcut		16.50	17.84	44.00	31.00
Qwen-2.5-7b-instrcut(CoT)		14.00	12.00	43.00	29.00
Qwen-2.5-72b-instrcut		34.33	42.50	40.00	34.00
Qwen-2.5-72b-instrcut(CoT)		25.33	39.00	40.00	33.00
Qwen-3-235b-a22b		37.67	51.50	44.00	37.00
Qwen-3-235b-a22b (CoT)		61.33	64.17	41.00	34.00
Llama-3.1-8b-instruct		11.67	13.67	22.00	15.00
Llama-3.1-8b-instruct (CoT)		12.67	16.17	28.00	19.00
Llama-3.3-70b-instruct		47.50	56.34	40.00	33.00
Llama-3.3-70b-instruct (CoT)		49.00	55.84	43.00	36.00
Mistral-7b-instruct-v0.3		6.50	6.34	29.00	23.00
Mistral-7b-instruct-v0.3 (CoT)		21.17	21.83	28.00	22.00
Mixtral-8x22b-instruct-v0.1		30.34	37.50	35.00	28.00
Mixtral-8x22b-instruct-v0.1 (CoT)		34.84	45.67	38.00	30.00
GPT-3.5-turbo-0125	FR	17.17	22.67	35.00	26.00
GPT-3.5-turbo-0125 (CoT)		17.33	22.50	37.00	29.00
GPT-4o-11-20		47.00	53.33	48.00	40.00
GPT-4o-11-20 (CoT)		44.50	48.67	**49.00**	41.00
DeepSeek R1		**52.17**	**62.50**	45.00	**44.00**
DeepSeek R1 (CoT)		50.34	62.17	47.00	40.00
Qwen-2.5-7b-instrcut		11.34	13.67	43.00	32.00
Qwen-2.5-7b-instrcut (CoT)		7.50	11.34	44.00	32.00
Qwen-2.5-72b-instrcut		37.84	24.67	45.00	37.00
Qwen-2.5-72b-instrcut (CoT)		32.83	24.33	44.00	37.00
Qwen-3-235b-a22b		37.17	35.33	38.00	34.00
Qwen-3-235b-a22b (CoT)		55.17	56.83	38.00	34.00
Llama-3.1-8b-instruct		12.33	13.83	29.00	23.00
Llama-3.1-8b-instruct (CoT)		17.67	14.50	28.00	21.00
Llama-3.3-70b-instruct		32.00	48.50	42.00	36.00
Llama-3.3-70b-instruct (CoT)		34.50	49.17	45.00	37.00
Mistral-7b-instruct-v0.3		3.67	8.17	28.00	23.00
Mistral-7b-instruct-v0.3 (CoT)		24.83	24.50	31.00	24.00
Mixtral-8x22b-instruct-v0.1		33.00	37.50	44.00	34.00
Mixtral-8x22b-instruct-v0.1 (CoT)		43.17	43.67	44.00	36.00
GPT-3.5-turbo-0125	JA	15.00	17.50	21.00	17.00
GPT-3.5-turbo-0125 (CoT)		12.83	16.50	19.00	14.00
GPT-4o-11-20		44.17	51.50	**45.00**	37.00
GPT-4o-11-20 (CoT)		40.50	49.16	**45.00**	**39.00**
DeepSeek R1		**52.00**	**55.17**	**45.00**	36.00
DeepSeek R1 (CoT)		48.17	53.34	44.00	35.00
Qwen-2.5-7b-instrcut		12.17	18.50	42.00	34.00
Qwen-2.5-7b-instrcut (CoT)		16.50	19.67	42.00	34.00
Qwen-2.5-72b-instrcut		38.17	41.33	43.00	36.00
Qwen-2.5-72b-instrcut (CoT)		31.84	40.17	43.00	35.00
Qwen-3-235b-a22b		40.17	47.00	40.00	33.00
Qwen-3-235b-a22b (CoT)		52.33	56.50	40.00	34.00
Llama-3.1-8b-instruct		9.17	11.17	43.00	32.00
Llama-3.1-8b-instruct (CoT)		12.34	11.67	22.00	15.00
Llama-3.3-70b-instruct		28.50	45.33	**45.00**	37.00
Llama-3.3-70b-instruct (CoT)		37.00	52.17	44.00	37.00
Mistral-7b-instruct-v0.3		13.83	6.33	28.00	23.00
Mistral-7b-instruct-v0.3 (CoT)		15.50	14.67	31.00	24.00
Mixtral-8x22b-instruct-v0.1		31.83	41.33	44.00	34.00
Mixtral-8x22b-instruct-v0.1 (CoT)		31.83	43.00	44.00	36.00

Table 10: Results of models for XNegotiationToM.

Model	Language	1st Belief Accuracy(%)	2nd Belief Accuracy(%)	Average Accuracy(%)	Reality Accuracy(%)
GPT-4o-11-20	ZH	69.92	93.75	85.30	100.00
GPT-4o-11-20 (CoT)		65.85	95.09	84.73	100.00
Llama-3.3-70b-instruct		65.85	83.48	77.23	99.67
Llama-3.3-70b-instruct (CoT)		68.29	84.82	78.96	100.00
Mixtral-8x22b-instruct-v0.1		56.91	56.25	56.48	100.00
Mixtral-8x22b-instruct-v0.1 (CoT)		51.22	51.79	51.59	100.00
GPT-4o-11-20	DE	85.37	89.29	87.90	98.67
GPT-4o-11-20 (CoT)		73.98	89.73	84.15	98.33
Llama-3.3-70b-instruct		90.24	87.50	88.47	96.33
Llama-3.3-70b-instruct (CoT)		88.62	87.50	87.90	95.67
Mixtral-8x22b-instruct-v0.1		67.48	45.98	53.60	96.33
Mixtral-8x22b-instruct-v0.1 (CoT)		65.85	46.88	53.60	96.00
GPT-4o-11-20	FR	79.67	94.20	89.05	99.67
GPT-4o-11-20 (CoT)		86.18	98.21	93.95	100.00
Llama-3.3-70b-instruct		88.62	95.98	93.37	100.00
Llama-3.3-70b-instruct (CoT)		91.06	94.64	93.37	99.33
Mixtral-8x22b-instruct-v0.1		65.04	56.70	59.65	100.00
Mixtral-8x22b-instruct-v0.1 (CoT)		57.72	56.25	56.77	100.00
GPT-4o-11-20	JA	73.17	93.75	86.46	100.00
GPT-4o-11-20 (CoT)		78.05	95.98	89.63	100.00
Llama-3.3-70b-instruct		70.73	87.05	81.27	98.67
Llama-3.3-70b-instruct (CoT)		67.48	75.00	72.33	99.33
Mixtral-8x22b-instruct-v0.1		61.79	51.34	55.04	91.00
Mixtral-8x22b-instruct-v0.1 (CoT)		60.16	49.11	53.03	92.00

Table 11: Results of models for other languages in EN template (XToMi benchmark).

Dimension	Example
Belief	**Background:** Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. **Dialogue History:** **{Context}** **Question1:** Based on the dialogue, what is the high preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood **Question2:** Based on the dialogue, what is the medium preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood **Question3:** Based on the dialogue, what is the low preference for items Agent 1 thinks Agent 2 is? A.Not given B.Water C.Food D.Firewood **Answer:**
Desire	**Background:** Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. Please answer the following three questions using "A", "B", "C", "D" without any explanation. **Dialogue History:** **{Context}** **Question1:** What is agent 1's high preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood **Question2:** What is agent 1's medium preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood **Question3:** What is agent 1's low preference for items based on the dialogue history? A.Not given B.Water C.Food D.Firewood **Answer:**
Intention	**Background:** Here is a negotiation conversation for a camping trip. There are two agents who own some basic supplies and negotiate with each other to split the additional food packages, water bottles, and firewood to make their camping trip even better. Each of these items will be of either High, Medium or Low priority for these two agents. Each of the additional items only has an available quantity of 3. **Dialogue History:** **{Context}** **Question:** What are the plausible intentions of {agent} expressed in '{utterance}'? Based on the dialogue history, select one or more strategies (i.e., 'A', 'B', 'C', . . . , 'I') from the following choices and their definition. Please select 'A', 'B', 'C', . . . , 'I' without any explanation. A. Build-Rapport: Participants discussing topics apart from the negotiation, in an attempt to build a rapport with the partner. B. Show-Empathy: An utterance depicts empathy when there is evidence of positive acknowledgments or empathetic behavior towards a personal context of the partner. C. Promote-Coordination: Used when a participant promotes coordination among the two partners. D. Callout-Fairness: A callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them. E. Undermine-Requirements: Refers to the scenario where a participant undermines the requirements of their opponent. F. Discover-Preference: An attempt to discover the preference order of the opponent. G. Describe-Need: Refers to arguments for creating a personal need for an item in the negotiation. H. No-Need: When a participant points out that they do not need an item based on personal context. I. No-Intention: If no strategy is evident, the utterance is labeled as No-Intention. **Answer:**

Table 12: Baseline prompt template. (XNegotiationToM - EN).

Dimension	Example (ZH)
Belief	**Background:** 背景：以下是一次关于露营旅行的谈判对话。两位参与者拥有一些基本物资，并相互协商如何分配额外的食物、水和火柴，以使他们的露营旅行更加愉快。这些物品对每位参与者的重要性优先级可以是高、中或低。每种额外物品的最大可用数量为3。请仅使用"A"、"B"、"C"、"D"作答，不需要解释。 **Dialogue History:** **{Context}** 问题**1**：根据对话，人物1认为人物2高优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 问题**2**：根据对话，人物1认为人物2中优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 问题**3**：根据对话，人物1认为人物2低优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 答案：
Desire	**Background:** 背景：以下是一次关于露营旅行的谈判对话。两位参与者拥有一些基本物资，并相互协商如何分配额外的食物、水和火柴，以使他们的露营旅行更加愉快。这些物品对每位参与者的重要性优先级可以是高、中或低。每种额外物品的最大可用数量为3。请仅使用"A"、"B"、"C"、"D"作答，不需要解释。 **Dialogue History:** **{Context}** 问题**1**：根据对话历史，人物1高优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 问题**2**：根据对话历史，人物1中优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 问题**3**：根据对话历史，人物1低优先级的物品是什么？ A.未提供 B.水 C.食物 D.火柴 答案：
Intention	**Background:** 背景：以下是一次关于露营旅行的谈判对话。两位参与者拥有一些基本物资，并相互协商如何分配额外的食物、水和火柴，以使他们的露营旅行更加愉快。这些物品对每位参与者的重要性优先级可以是高、中或低。每种额外物品的最大可用数量为3。 **Dialogue History:** **{Context}** 问题：{agent}在"{utterance}"中表达的可能意图是什么？基于对话历史，从以下选项（即"A"、"B"、"C"、...、"I"）及其定义中选择一个或多个策略。请仅选择"A"、"B"、"C"、...、"I"，无需解释。 A. 建立融洽关系：参与者讨论与谈判无关的主题，试图与对方建立融洽关系。 B. 表达同情：当对方提到个人背景时，表现出积极的认可或同情行为的语句。 C. 促进协调：当参与者促进双方之间的协调时使用。 D. 呼吁公平：为了个人利益而呼吁公平，包括承认公平交易或对方提出有利于自己的交易时。 E. 破坏要求：指参与者破坏对方需求的情境。 F. 发现偏好：试图发现对方偏好顺序的行为。 G. 描述需求：为某一物品的个人需求提供论据。 H. 没有需求：根据个人背景指出他们不需要某一物品。 I. 没有意图：如果没有明显的策略，则将语句标记为"没有意图"。 答案：

Table 13: Baseline prompt template (XNegotiationToM - ZH).

Dimension	Example (DE)
Belief	**Background:** Hintergrund: Hier ist ein Verhandlungsgespräch für einen Campingausflug. Es gibt zwei Agenten, die einige grundlegende Vorräte besitzen und miteinander verhandeln, um die zusätzlichen Lebensmittelpakete, Wasserflaschen und Brennholz aufzuteilen, um ihren Campingausflug noch besser zu machen. Jeder dieser Gegenstände hat für diese beiden Agenten entweder eine hohe, mittlere oder niedrige Priorität. Für jeden der zusätzlichen Gegenstände ist nur eine Menge von 3 verfügbar. Bitte beantworten Sie die folgenden drei Fragen mit "A", "B", "C", "D" ohne Begründung. **Dialogverlauf:** **{Context}** **Frage1:** Nach dem Dialog, was ist die hohe Präferenz für Gegenstände von Person 2 laut Person 1? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Frage2:** Nach dem Dialog, was ist die mittlere Präferenz für Gegenstände von Person 2 laut Person 1? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Frage3:** Nach dem Dialog, was ist die niedrige Präferenz für Gegenstände von Person 2 laut Person 1? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Antwort:**
Desire	**Background:** Hintergrund: Hier ist ein Verhandlungsgespräch für einen Campingausflug. Es gibt zwei Agenten, die einige grundlegende Vorräte besitzen und miteinander verhandeln, um die zusätzlichen Lebensmittelpakete, Wasserflaschen und Brennholz aufzuteilen, um ihren Campingausflug noch besser zu machen. Jeder dieser Gegenstände hat für diese beiden Agenten entweder eine hohe, mittlere oder niedrige Priorität. Für jeden der zusätzlichen Gegenstände ist nur eine Menge von 3 verfügbar. Bitte beantworten Sie die folgenden drei Fragen mit "A", "B", "C", "D" ohne Begründung. **Dialogverlauf:** **{Context}** **Frage1:** Was ist die hohe Präferenz für Gegenstände von Person 1 nach dem Dialogverlauf? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Frage2:** Was ist die mittlere Präferenz für Gegenstände von Person 1 nach dem Dialogverlauf? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Frage3:** Was ist die niedrige Präferenz für Gegenstände von Person 1 nach dem Dialogverlauf? A. Nicht angegeben B. Wasser C. Essen D. Brennholz **Antwort:**
Intention	**Background:** Hintergrund: Hier ist ein Verhandlungsgespräch für einen Campingausflug. Es gibt zwei Agenten, die einige grundlegende Vorräte besitzen und miteinander verhandeln, um die zusätzlichen Lebensmittelpakete, Wasserflaschen und Brennholz aufzuteilen, um ihren Campingausflug noch besser zu machen. Jeder dieser Gegenstände hat für diese beiden Agenten entweder eine hohe, mittlere oder niedrige Priorität. Für jeden der zusätzlichen Gegenstände ist nur eine Menge von 3 verfügbar. **Dialogverlauf:** **{Context}** **Frage:** Was sind die plausiblen Absichten von {agent}, die in "{utterance}" ausgedrückt werden? Wählen Sie nach dem Dialogverlauf eine oder mehrere Strategien (d. h. "A", "B", "C", . . . , "I") aus den folgenden Optionen und deren Definition aus. Wählen Sie "A", "B", "C", . . . , "I" ohne Begründung aus. A. Beziehung aufbauen: Teilnehmer diskutieren Themen abseits der Verhandlung, um ein Vertrauensverhältnis zum Partner aufzubauen. B. Empathie zeigen: Eine Äußerung zeigt Empathie, wenn es Anzeichen für positive Anerkennungen oder empathisches Verhalten gegenüber einem persönlichen Kontext des Partners gibt. C. Koordination fördern: Wird verwendet, wenn ein Teilnehmer die Koordination zwischen den beiden Partnern fördert. D. Fairness einfordern: Ein Aufruf zur Fairness für persönlichen Vorteil, entweder wenn ein fairer Deal anerkannt wird oder wenn der Gegner einen Deal anbietet, der ihm Vorteile bringt. E. Anforderungen untergraben: Bezieht sich auf das Szenario, in dem ein Teilnehmer die Anforderungen seines Gegners untergräbt. F. Präferenz herausfinden: Ein Versuch, die Präferenzreihenfolge des Gegners herauszufinden. G. Bedarf beschreiben: Bezieht sich auf Argumente für die Schaffung eines persönlichen Bedarfs für einen Gegenstand in der Verhandlung. H. Kein Bedarf: Wenn ein Teilnehmer darauf hinweist, dass er einen Gegenstand aufgrund des persönlichen Kontexts nicht benötigt. I. Keine Absicht: Wenn keine Strategie erkennbar ist, wird die Äußerung als Keine Absicht gekennzeichnet. **Antwort:**

Table 14: Baseline prompt template (XNegotiationToM - DE).

Dimension	Example (FR)
Belief	**Background:** Contexte : Voici une conversation de négociation pour un voyage de camping. Il y a deux agents qui possèdent quelques fournitures de base et négocient entre eux pour répartir les paquets de nourriture supplémentaires, les bouteilles d'eau et les bois de chauffage afin d'améliorer leur voyage de camping. Chacun de ces éléments sera d'une priorité Haute, Moyenne ou Faible pour ces deux agents. Chacun des articles supplémentaires n'a qu'une quantité disponible de 3. Veuillez répondre aux trois questions suivantes en utilisant "A", "B", "C", "D" sans aucune explication. **Historique de la conversation:** **{Context}** **Question 1:** D'après le dialogue, quels sont les articles que la Personne 1 considère comme étant de haute priorité pour la Personne 2? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Question 2:** D'après le dialogue, quels sont les articles que la Personne 1 considère comme étant de priorité moyenne pour la Personne 2? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Question 3:** D'après le dialogue, quels sont les articles que la Personne 1 considère comme étant de faible priorité pour la Personne 2? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Réponse:**
Desire	**Background:** Contexte : Voici une conversation de négociation pour un voyage de camping. Il y a deux agents qui possèdent quelques fournitures de base et négocient entre eux pour répartir les paquets de nourriture supplémentaires, les bouteilles d'eau et les bois de chauffage afin d'améliorer leur voyage de camping. Chacun de ces éléments sera d'une priorité Haute, Moyenne ou Faible pour ces deux agents. Chacun des articles supplémentaires n'a qu'une quantité disponible de 3. Veuillez répondre aux trois questions suivantes en utilisant "A", "B", "C", "D" sans aucune explication. **Historique de la conversation:** **{Context}** **Question 1:** Quels sont les articles de haute priorité pour la Personne 1? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Question 2:** Quels sont les articles de priorité moyenne pour la Personne 1? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Question 3:** Quels sont les articles de priorité faible pour la Personne 1? A. Pas donné B. Eau C. Nourriture D. Bois de chauffage **Réponse:**
Intention	**Background:** Contexte : Voici une conversation de négociation pour un voyage de camping. Il y a deux agents qui possèdent quelques fournitures de base et négocient entre eux pour répartir les paquets de nourriture supplémentaires, les bouteilles d'eau et les bois de chauffage afin d'améliorer leur voyage de camping. Chacun de ces éléments sera d'une priorité Haute, Moyenne ou Faible pour ces deux agents. Chacun des articles supplémentaires n'a qu'une quantité disponible de 3. **Historique de la conversation:** **{Context}** **Question:** Quelles sont les intentions plausibles de {agent} exprimées dans "{utterance}"? Sur la base de l'historique de la conversation, sélectionnez une ou plusieurs stratégies (c.-à-d., "A", "B", "C", ..., "I") parmi les choix suivants et leur définition. Veuillez sélectionner "A", "B", "C", ..., "I" sans aucune explication. A. Établir des relations: Les participants discutent de sujets autres que la négociation, dans le but de créer une relation avec le partenaire. B. Faire preuve d'empathie: Une énonciation montre de l'empathie lorsqu'il y a des preuves de reconnaissance positive ou de comportement empathique envers un contexte personnel du partenaire. C. Promouvoir la coordination: Utilisé lorsqu'un participant favorise la coordination entre les deux partenaires. D. Revendiquer l'équité: Un appel à l'équité pour un avantage personnel, soit en reconnaissant un accord équitable, soit lorsque l'adversaire propose un accord qui lui profite. E. Saper les exigences: Désigne le cas où un participant sape les exigences de son adversaire. F. Découvrir la préférence: Une tentative de découvrir l'ordre de préférence de l'adversaire. G. Décrire le besoin: Arguments visant à créer un besoin personnel pour un article dans la négociation. H. Aucun besoin: Lorsqu'un participant indique qu'il n'a pas besoin d'un article selon son contexte personnel. I. Aucune intention: Si aucune stratégie n'est évidente, l'énoncé est étiqueté comme Aucune-Intention. **Réponse:**

Table 15: Baseline prompt template (XNegotiationToM - FR).

Dimension	Example (JA)
Belief	**Background:** 背景：以下はキャンプ旅行に関する交渉の会話です。二人の登場人物が基本的な用品を所有しており、追加の食べ物、水、薪を分配して、キャンプ旅行をより良くするために交渉しています。これらの各アイテムは、二人にとって「高」、「中」、「低」のいずれかの優先度を持ちます。追加アイテムは各々最大3個までしか利用できません。 以下の3つの質問に"A"、"B"、"C"、"D"を使って説明なしで答えてください。 対話履歴**:** **{Context}** 質問**1:** 対話に基づき、人物1が人物2について考える「高」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 質問**2:** 対話に基づき、人物1が人物2について考える「中」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 質問**3:** 対話に基づき、人物1が人物2について考える「低」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 回答**:**
Desire	**Background:** 背景：以下はキャンプ旅行に関する交渉の会話です。二人の登場人物が基本的な用品を所有しており、追加の食べ物、水、薪を分配して、キャンプ旅行をより良くするために交渉しています。これらの各アイテムは、二人にとって「高」、「中」、「低」のいずれかの優先度を持ちます。追加アイテムは各々最大3個までしか利用できません。 以下の3つの質問に"A"、"B"、"C"、"D"を使って説明なしで答えてください。 対話履歴**:** **{Context}** 質問**1:** 対話履歴に基づいて、人物1の「高」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 質問**2:** 対話履歴に基づいて、人物1の「中」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 質問**3:** 対話履歴に基づいて、人物1の「低」優先度のアイテムは何ですか？ A. 未提供 B. 水 C. 食べ物 D. 薪 回答**:**
Intention	**Background:** 背景：以下はキャンプ旅行に関する交渉の会話です。二人の登場人物が基本的な用品を所有しており、追加の食べ物、水、薪を分配して、キャンプ旅行をより良くするために交渉しています。これらの各アイテムは、二人にとって「高」、「中」、「低」のいずれかの優先度を持ちます。追加アイテムは各々最大3個までしか利用できません。 対話履歴**:** **{Context}** 質問**:** {agent}が発した「{utterance}」において、考えられる意図は何ですか？対話履歴に基づいて、以下の選択肢から1つ以上の戦略（"A"、"B"、"C"、...、"I"）を選択してください。説明なしで"A"、"B"、"C"、...、"I"を選択してください。 A. 信頼関係を築く：交渉とは別の話題を議論し、相手との信頼関係を築こうとする発言。 B. 共感を示す：相手の個人的な文脈に対して、肯定的な応答や共感的な行動が見られる発言。 C. 協調を促進する：両者間の協調を促進しようとする発言。 D. 公平性を求める：自分に有利な条件を認める、または相手の提案が自分に利益をもたらすことを指摘する発言。 E. 要件を損なう：相手の要件を軽視または否定する発言。 F. 好みを見つける：相手の優先順位を探ろうとする発言。 G. 需要を説明する：自分のアイテム需要を論じる発言。 H. 要求なし：個人的な文脈に基づき、アイテムが必要ないことを指摘する発言。 I. 意図なし：特定の戦略が明らかでない場合、このラベルが適用される。 回答**:**

Table 16: Baseline prompt template (XNegotiationToM - JA).

Language	Template
EN	{story} {question} Choose from the following: (a) {containers_0} (b) {containers_1} Keep your answer concise. Answer with a single word.
ZH	{story} {question} 请从以下选项中选择: (a) {containers_0} (b) {containers_1} 请保持答案简洁。使用单个词回答。
DE	{story} {question} Wählen Sie aus den folgenden Möglichkeiten: (a) {containers_0} (b) {containers_1} Halten Sie Ihre Antwort kurz. Antworten Sie mit einem einzigen Wort.
FR	{story} {question} Choisissez parmi les options suivantes : (a) {containers_0} (b) {containers_1} Gardez votre réponse concise. Répondez avec un seul mot.
JA	{story} {question} 以下から選択してください： (a) {containers_0} (b) {containers_1} 簡潔に答えてください。一語で回答してください。

Table 17: Multi-language prompt templates (XToMi).

Language	Template
EN	{Context} Question: {Question} (a) {Answer_0} (b) {Answer_1} Choose an answer from above:
ZH	{Context} 问题: {Question} (a) {Answer_0} (b) {Answer_1} 从上面选择一个答案:
DE	{Context} Frage: {Question} (a) {Answer_0} (b) {Answer_1} Wähle eine Antwort von oben:
FR	{Context} Question : {Question} (a) {Answer_0} (b) {Answer_1} Choisissez une réponse ci-dessus :
JA	{Context} 質問: {Question} (a) {Answer_0} (b) {Answer_1} 上記の中から答えを選んでください:

Table 18: Multi-language prompt templates (XFANToM).

FANToM Conversation:
... <The-Beginning-Sentences-of-Example>...

Person 1: Definitely the vibe. The charm of Paris during Christmas is unmatched. The beautiful decorations, the bustling markets, and the lovely smell of fresh baked goods - it all was so different and vibrant. Have you ever had a chance to explore Paris yourself?
Person 2: No, not yet, but I would definitely love to. I've heard the Eiffel Tower looks phenomenal during Christmas. You know, the holiday which I found interesting was **Thanksgiving**. It was during one instance when I was in college.
Person 1: That's interesting. What was so memorable about it?
Person 2: We didn't return home during the break because we were preparing for our finals. My friends and I decided to cook a **Thanksgiving meal** together; none of us had any cooking experience prior to that. The **turkey**, a total disaster, but the cornbread and pie were amazing. That was the occasion when I realized I was quite a decent baker!
Person 1: That sounds like a fun experience, especially the part where you discovered your hidden talent. Isn't it amazing how holidays often lead us to such happy surprises? Person 2: Absolutely. Holidays indeed have a charm of their own. They bring us closer, build strong bonds and generate memories that last a lifetime.
Person 1: Couldn't agree more, Person 2. Let's hope we create some fantastic memories here today as well!
Person 2: Certainly, Person 1. I'm looking forward to it.
Person 3: Hello Person 1, Person 2, I hope I'm not interrupting anything. This place is really coming alive with preparations for the iday.
Person 1: Hello Person 3, not at all! We were just sharing our cherished holiday experiences. Person 2 here discovered he could bake uring a college **Thanksgiving**!
Person 3: Baking, huh? That sounds a lot like my New Year's experience. I had to whip up a dessert at the last minute.
Person 2: Really? What was it? I love hearing about impromptu cooking ordeals!
Person 3: It was a chocolate cheesecake. I'm personally more of a main course guy, but that was the only thing we were missing. Turned out well, surprisingly!
Person 1: Haha, both of you men of hidden talents! You know, I've had my own fair share of kitchen successes and failures during Easter. Quite the adventure it is.
Person 3: Then I guess we can all agree that holidays are for discovering new facets about ourselves, right?
Person 2: Couldn't have put it better, Person 3. To holidays and their endless surprises! Person 4: Hello everyone, what an exciting conversation you're all having!
Person 2: Hi Person 4! Good to see you. We've been sharing our holiday stories and about the discoveries we made about ourselves during those times.
Person 4: Absolutely love the idea. We all have something to learn from our holidays.

... <The-Middle-Sentences-of-Example>...

Person 5: Hello everyone, I see we have a lively discussion happening over here!
Person 1: Hi Person 5, good to see you. We were just reminiscing about our holiday experiences and how they've led to surprising discoveries. Now, if we're talking about holidays, we surely can't ignore the fun and sometimes stress of holiday gifting and shopping, right?
Person 5: Oh, absolutely. I usually take the role of Santa in my family, deciding and buying gifts for everyone. It's a challenge, but I love the joy it brings.
Person 2: Funny you mentioned Santa, Person 5. One Christmas, I decided to handcraft all my gifts. It showed me I have a decent knack for arts and crafts.
Person 4: Crafting your own gifts - that's lovely, Person 2. I remember one year, I ended up forgetting about buying gifts until the last minute. I ended up learning that I'm pretty good at making quick decisions!
Person 3: Seems like we've all learned something new about ourselves from holiday shopping. For me, it's probably the fact that I'm surprisingly frugal. I tend to find great gifts without breaking the bank!

... <The-Rest-Sentences-of-Example>...

Belief Question: What does Person 5 believe about how these holiday experiences led to surprising discoveries about their cooking and baking abilities?
A : "Person 5 is unaware of how these holiday experiences led to surprising discoveries about their cooking and baking abilities. This is due to the fact that he was not involved in the conversation when this topic was discussed."
B : "Person 5 believes that these holiday experiences led Person 2, Person 3, and Person 4 to surprising discoveries about their cooking and baking abilities. Person 2 discovered his baking skills when he made an amazing cornbread and pie during a Thanksgiving in college. Person 3 found out about his knack for dessert-making when he made a chocolate cheesecake for New Year's unexpectedly. Person 4 found her passion for cooking during a family spring holiday where she enjoyed preparing the main meal."
Ground Truth: A

English Version Answer: A
Chinese Version Answer: B
German Version Answer: A
French Version Answer: A
Japanese Version Answer: B

Table 19: An case study on the cultural context embedded in language in XFANToM.