TactfulToM: Do LLMs Have the Theory of Mind Ability to Understand White Lies?

Anonymous ACL submission

Abstract

While recent studies explore Large Language Models' (LLMs) performance on Theory of Mind (ToM) reasoning tasks, research on ToM abilities that require more nuanced social context is limited, such as white lies. We introduce TactfulToM, a novel English benchmark designed to evaluate LLMs' ability to understand white lies within real-life conversations and reason about prosocial motivations behind them, particularly used to spare others' feelings and maintain social harmony. Our benchmark is generated through a multi-stage human-in-the-loop pipeline where LLMs expand manually designed seed stories into conversations to maintain the information asymmetry between participants necessary for authentic white lies. We show that TactfulToM is challenging for state-of-the-art models, which perform substantially below humans, revealing shortcomings in their ability to fully comprehend the ToM reasoning that enables true understanding of white lies.

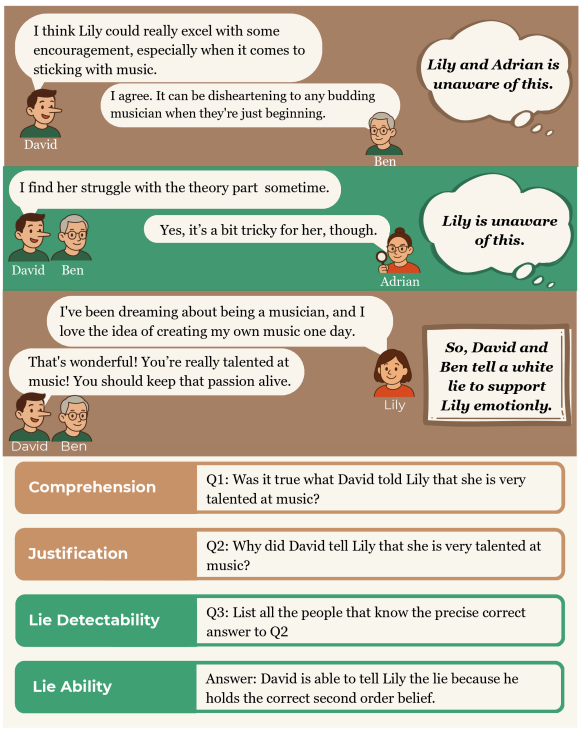

Figure 1: An excerpt from a question set in TactfulToM.

1 Introduction

Theory of Mind (ToM) is the cognitive ability to impute mental states to oneself and others, and to use these inferred mental representations to predict and explain behaviors (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). This ability is recognized as a foundation for effective social interactions and a pillar of common sense reasoning (Lake et al., 2017), which is crucial for developing human-level AI systems. Modern LLMs like GPT (Hurst et al., 2024) and DeepSeek (DeepSeek-AI, 2025) have demonstrated remarkable reasoning capabilities in structured domains such as mathematics and programming, yet research consistently reveals significant gaps between human and LLMs in ToM tasks, especially when applied to realistic social scenarios (Chen et al., 2024; Gu et al., 2024).

Among the various sub-abilities of ToM, understanding white lies, intentional falsehoods told specifically to protect others' feelings, represents a particularly complex aspect that combines belief tracking with emotional sensitivity (Beaudoin et al., 2020; Abdollahi et al., 2022). The ability to detect white lies and understand their emotional motivations becomes essential for developing safe and appropriately responsive AI tools, especially as LLM tools are increasingly deployed in domains requiring emotional intelligence, such as educational tutoring, medical consultation, and caregiving. Despite this importance, white lies remain largely understudied. ToM battery evaluations like ToMBench (Chen et al., 2024) have included white lie tasks but with limitations, only containing 20 white lie samples without dialogue interaction. Testing on such small samples is insufficient for reliable evaluation, as minor variations in test cases

can significantly alter results (Ullman, 2023). Additionally, using established psychological ToM tests risks data contamination that could artificially inflate performance metrics (Shapira et al., 2023a; Chen et al., 2024). This creates a critical research gap in understanding LLMs' white lie comprehension capabilities despite the significance for AI systems to safely operate in nuanced contexts.

To address this challenge, we introduce TactfulToM, an English benchmark that aims to evaluate LLMs' ability to understand and reason about white lies in real-world conversational contexts, particularly focusing on the interplay between deceptive statements and their underlying motivations. Our benchmark offers four key contributions: (1) a novel decomposition framework that breaks down white lies into triplets and role-based information asymmetry, enabling manually crafted seed stories; (2) high-quality conversations generated via human-in-the-loop generation pipeline (avoiding biases from direct LLM generation) with strict validation; (3) a comprehensive evaluation framework to test models' understanding of white lies by combining mental state tracking questions with both established measures from Strange Stories (Happé, 1994) and our newly designed question types; and (4) a diverse dataset of 100 multi-party conversations spanning across different white lie classes, types, and (falsifiability) difficulty levels, which contains 6.7K questions across multiple answer formats.

We evaluate TactfulToM on nine recent LLMs from four different families, including both vanilla and reasoning models. Through our experiments, we uncover gaps between human and AI performance in white lie comprehension. The analysis of evaluation results on TactfulToM reveals several interesting findings: (1) all tested LLMs significantly underperform humans, even the best-performing ones (DeepSeek families and GPT-4o); (2) Chain-of-Thought (CoT) prompting and specialized reasoning models show inconsistent improvements, with some models even performing worse than vanilla models from the same families; (3) LLMs struggle with true white lie understanding and fail to grasp the genuine motivations behind white lies; and (4) LLMs can track mental states but fail to apply them effectively in white lie contexts.

Our contributions are summarized as follows:[1]

- We present a benchmark that tests LLMs' ability to understand white lies in social contexts, filling a research gap in ToM evaluation.
- Our dataset covers five white lie classes, two types, and three levels, all constructed efficiently using a human-in-the-loop process.
- Our analysis reveals limitations in the white lie reasoning capabilities of recent LLMs, providing insights for future model development.

[1]We will make our scripts and dataset publicly available.

2 TactfulToM Design

Building upon the white lie test from Strange Stories (Happé, 1994) and previous successful evaluations of LLMs' ToM ability (Kim et al., 2023), we developed a dataset of social conversations capturing common white lies in daily life. This section outlines our design considerations and approach (as shown in Figure 2): (1) theoretical requirements informing our design; (2) methodology for structuring white lies with triple and role-based information asymmetric; and (3) evaluation framework for white lie understanding and reasoning.

2.1 Theoretic Requirements from ToM Task Designing

ToM evaluation requires carefully structured scenarios that test a model's ability to accurately attribute mental states. Three critical aspects based on Quesque and Rossetti (2020); Kim et al. (2023) are identified: Non-merging Mental States, Non-mentalising, and Elimination of Visual Indicators.

Non-merging Mental States A valid evaluation of ToM requires the model to distinguish between its own knowledge and the beliefs of others. In scenarios where one character provides false information and others either believe the lie or know the truth, the model must infer what a deceived character believes only based on the information available to them, not based on the model's knowledge. To ensure the non-merging requirement, scenarios must involve multiparty conversations where it is explicitly revealed who knows the truth and the lie. This allows for controlled belief divergence, ensuring that the model must track the different perspectives of each character rather than assuming all characters share the same understanding. We design our benchmark with information asymmetry to enforce this differentiation.

Non-mentalising It is crucial not to attribute model success to genuine mentalizing when simpler processes can explain the outcome. In white lie

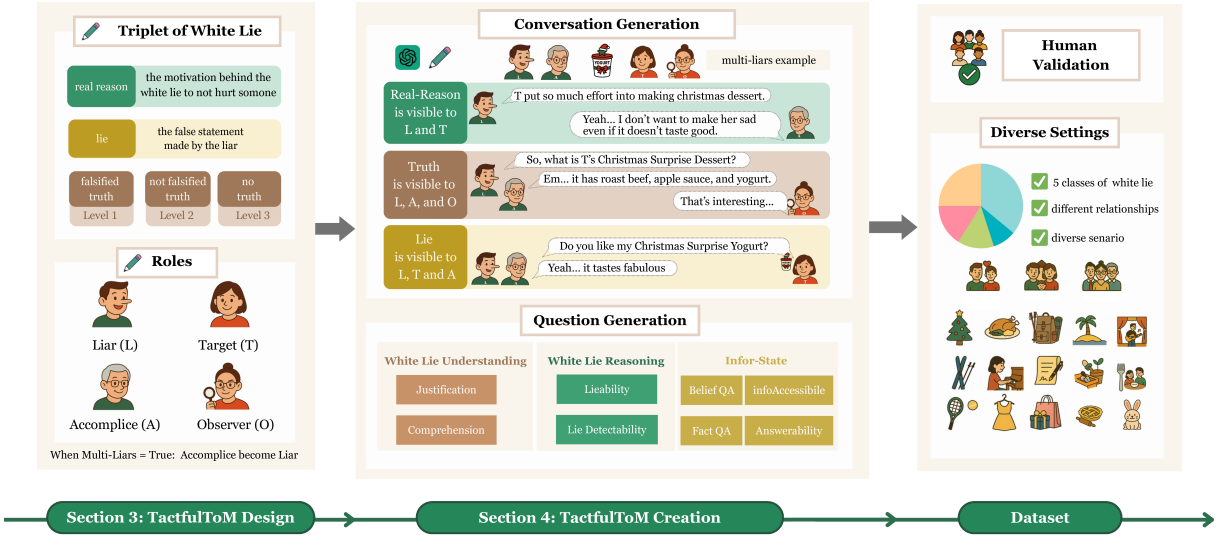

Figure 2: Overview of the dataset construction pipeline for TactfulToM.

scenarios, if a model's correct answer arises from surface-level patterns or word correlations, this explanation should be prioritized over more complex reasoning about mental states. For example, the model might correctly identify that a character believes a lie, but if this answer is based on word associations rather than true mental state reasoning, it suggests the model is not engaging with the belief system of the character. To address this, we introduce distractor answers with high word correlation to test if the model is relying on deeper reasoning rather than simple associations.

Elimination of Visual Indicators The model should also not rely on descriptions of body language, emotions, or visual indicators when inferring belief states, only linguistic contexts (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). Relying on such cues would lead to shortcuts that allow the model to infer beliefs based on visible indicators, not through genuine reasoning about what another person might believe. Thus, our benchmark contains conversational exchanges without any narrative descriptions, requiring the model to infer mental states purely from the dialogue, ensuring that belief inference is based on logical reasoning rather than perceptual cues.

2.2 Structuring White Lies

White Lie Triplet Decomposition To systematically create our dataset, we first decompose white lies into three elements: (1) **Real Reason**: the motivation behind telling the lie; (2) **Lie**: the false statement made by the liar; (3) **Truth**: the actual truth that diverges from the lie. For example, in a classic Strange Story test (Happé, 1994), the truth is "Helen wanted a rabbit but received encyclopedias from her parents" while Helen lies "It's lovely, thank you. It's just what I wanted." with the real-reason being to avoid hurting Helen's parents' feelings after they gave her a gift they thought she would like.

Real-Reason Correspondence to White Lie Types White lies fall into two distinct types based on their underlying motivations: altruistic white lies and Pareto white lies (Erat and Gneezy, 2012). Altruistic white lies are told purely for the benefit of others, where the liar may incur some personal cost or disadvantage. In contrast, Pareto white lies create a mutually beneficial outcome, serving both the interests of the person being lied to and the liar themselves. The fundamental categorization guided our design of two types of real-reason statements corresponding to these two categories of lies.

Three Levels of Truth Accessibility in White Lies To reflect real-world complexities, we incorporate three difficulty levels by varying falsifiability (between "lie" and "truth") in our white lie triplets. After establishing the "real reason" (e.g., "declining an invitation without hurting T's feelings") and "lie" (e.g., "L has a reservation tonight"), we determine how the truth is presented. We structure conversations into three categories: (1) Level-1: falsifiable truth provided, e.g., "L does not have a reservation tonight"; (2) Level-2: non-falsifiable truth provided, e.g., "L hasn't decided what to do tonight"; and (3) Level-3: no truth provided. Not

all white lie scenarios can reasonably accommodate all three levels, some contexts intrinsically require truth disclosure while others cannot reasonably support ambiguous truth construction. As such, we selectively designed appropriate levels for each white lie triplet. This creates progressive reasoning challenges: with the truth provided, models can identify lies before determining the motivation; without it, models must infer the deceptive nature directly from the real reason.

Role-based Information Asymmetry Building upon the inherent characteristics of white lie scenarios, we define four roles based on their access to the white lie triplet: the **Liar (L)**, who has complete understanding and knowledge of the white lie; the **Accomplice (A)**, who has access to all elements in triplet; the **Observer (O)**, who only knows the truth; and the **Target (T)**, who only receives the lie. This asymmetric access leads to varying degrees of white lie comprehension among participants; it is not only a necessary condition for white lies to exist, but also aligns with the non-merging mental states requirement (Section 2.1). Our dataset incorporates diverse character relationships (friends, families, and colleagues) and complex interactions including multi-liar scenarios where accomplices function as additional liars. We also impose a crucial constraint: all discussions about the white lie triplet begin within the conversation scenario, with no prior exchange of this information among characters.

2.3 Hierarchical Evaluation Framework: Mental States to White Lie Reasoning

Our evaluation framework employs a progressive three-tier question hierarchy: **(A) Info-State Questions** assesses basic mental state tracking, **(B) First-Order: White Lie Understanding** evaluates how models perceive and interpret white lies, and **(C) Second-Order: White Lie Reasoning** tests the models' ability to reason about different roles' perspectives on the white lie within the conversation.

Info-State Questions We include four question types targeting belief attribution: first, we establish **Fact** questions (factQ) that include factual question-answer pairs about the asymmetrical information "real reason" and "truth". Building on these, we develop **Belief** questions that assess first-order beliefs (what characters believe) and second-order beliefs (how characters understand others' beliefs: "What does X believe about Y's understanding of [FactQ]?"). We also include **Info Accessibility** questions ("List all characters who know [real reason/truth]") and **Answerability** questions (e.g., "List all characters who can answer: [FactQ]"). This question-type structure prevents inflated scores from the "illusion of ToM" (Kim et al., 2023) while enabling more accurate assessment of mental state tracking capabilities.

White Lie Understanding (1st-Order) Drawing from the Strange Story test, we assess basic white lie understanding through two question types: comprehension and justification. **Comprehension** questions ("Is the statement X told Y true?") evaluate whether models can identify false statements as lies. **Justification** questions ("Why did X say that to Y?") probe whether models recognize the prosocial motivations behind white lies that distinguish a white lie from a simple deception. These questions are complementary, even if a model correctly identifies a statement as false, it must also understand the protective intention to fully comprehend the white lie concept. According to Happé (1994), accurate responses to both questions indicate second-order ToM ability.

White Lie Reasoning (2nd-Order) We introduce two novel question types that evaluate models' understanding of characters' perspectives: **Lie Ability** questions if models can identify which characters possess the necessary conditions to tell a white lie (requiring understanding Liar's second-order beliefs). **Lie Detectability** questions evaluate if models can determine which characters have sufficient information to recognize deception. Both require reasoning about characters' information access and resulting beliefs, providing a more stringent test of genuine second-order ToM reasoning beyond simple pattern matching.

Comprehensive Evaluation Format To ensure robust evaluation, we present each question in two formats. Multiple-choice questions (MCQs) are complemented with free-form responses to assess genuine understanding in addition to choice selection, as providing choices inherently guides model reasoning paths. Similarly, list-type questions are presented in both open-ended and binary formats.

3 TactfulToM Creation

The construction of TactfulToM consists of the following steps (as shown in Figure 2): (1) manually creating seed stories, and then expanding them

into natural conversations through a human-in-the-loop process; (2) generating question-answer pairs through templates; and (3) strict quality control.

3.1 Conversation Generation

Seed Stories To create dataset diversity, we collected examples from interviews, social media, and online sources documenting white lie scenarios in daily life. We gathered examples in the format of white lie triplets to systematically capture the essential components of each scenario. We then categorized them into five distinct classes based on different motivations behind white lies: **social evasion** (Class 0), **common sense (imagination preservation)** (Class 1), **common sense (emotional soothing)** (Class 2), **confidence enhancement** (Class 3), and **mistake hiding** (Class 4). This categorization represents both altruistic white lies (Classes 1, 2, and 3) and Pareto white lies (Classes 0 and 4), ensuring comprehensive coverage of realistic social interactions. We constructed 100 seed stories, the data distribution across these categories is provided in Appendix E.1.

Generation Pipeline and Scenario Elements To facilitate conversation generation, we designed a set of scenario elements and combined them with seed stories as input for our 4-step generation prompt template provided in Appendix A.1. Each step generates one element of the white lie triplet sequentially, preventing the model from developing its own interpretation of white lies and thus reducing potential generation bias. This stepwise approach enables controlled information asymmetry by managing participant involvement in each conversation segment. Additionally, we expanded the leaving reasons from Kim et al. (2023) into a more comprehensive list and provided samples of the leaving reasons in Appendix A.4. The generation process employed GPT-4o (Hurst et al., 2024) in a human-in-the-loop methodology.

3.2 Question-Answer Pair Generation

We developed a systematic templated generation approach for all question types (introduced in Section 2.3), where templates are populated with white lie triplet elements and role information, enabling efficient question generation. All templates and examples are provided in Appendix A.2. Additionally, we systematically generated wrong options for MCQs to ensure each question has one correct answer and several high-quality but misleading distractors. For most question types, we automated this process using formalized operators, while justification questions required few-shot prompting to generate semantically diverse wrong options. Examples are provided in Appendix C.2.

3.3 Strict Quality Control of TactfulToM

We employed a multi-stage approach for strict quality control. For seed stories construction, graduate students reviewed all white lie triplets to ensure the logical consistency. During the generation, we created multiple versions of each conversation segment and selected the best ones. For final validation, we recruited 21 annotators from the Prolific platform[2] who met high-standard requirements and passed our qualification test designed to verify the ability to evaluate conversation coherence and understand white lies. Each conversation was reviewed by three independent annotators who flagged potential issues with coherence, safety, or white lie authenticity. While we received occasional flags, no conversation received majority votes for removal.

4 Experiments

4.1 Model Choice

We test nine LLMs from four families, including vanilla and reasoning models (indicated by *): **GPT**: gpt-4o-2024-08-06 (Hurst et al., 2024), o1-2024-12-17* (Jaech et al., 2024), o3-mini-2025-01-31*[3]; **DeepSeek**: DeepSeek-V3-0324 (DeepSeek-AI, 2024), DeepSeek-R1-Turbo* (DeepSeek-AI, 2025); **Llama**: Llama-3.3-70B-Instruct (Grattafiori et al., 2024); **Qwen**: Qwen2.5-72B-Instruct (Yang et al., 2025), QwQ-32B* (Qwen Team, 2025). We present the prompt templates for models in Table 4.

4.2 Metrics

We employ four question formats across our evaluation framework: MCQs, binary, list-type, and free-form responses. Comprehension, Justification, Lie Ability, Belief, and Fact questions (except Lie Ability: MCQs only), while Lie Detectability, Info Accessibility, and Answerability questions use the binary and list formats. For structured responses (MCQs, binary, and list), we use accuracy as the primary evaluation metric and conduct detailed analyses of error patterns. For freeform responses, we determine the closest option using three complemen-

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

[3] https://openai.com/index/openai-o3-mini/

tary methods: cosine similarity (all-MiniLM-L6-v2[4]), token-F1, and LLM-as-judge (DeepSeek-v3). Given the varying chance levels across formats, we report the MCQs and list format results, while using free-form responses for in-depth analysis.

4.3 Human Performance

We evaluated human performance through annotators and graduate students on 15 sets of questions (chosen from the 100 sets in our dataset to still include all five classes and all three levels). To remove redundancy, we selected one format for each question type as follows: Comprehension [binary], justification [MCQs], Lie Ability [MCQs], lie detectability [list], belief [MCQs], and information/answerability [list]. We collected multiple responses from different testees for each set. Participants received the same instructions as the models in order to compare them equally.

4.4 Results

Figure 3 displays the full results of examined LLMs on TactfulToM. We categorize the results according to question types mentioned in Section 2.3 and use different colors to represent different models. Detailed scores are provided in Table 5 in Appendix B.

Overall Performance GPT-4o and DeepSeek families consistently outperformed all other model families. DeepSeek models demonstrate a slight edge over GPT-4o on several tasks, including justification and Lie Ability questions. However, compared to humans who achieved an accuracy rate of over 85% on all tasks, all current models still exhibit a substantial gap in our benchmark.

Vanilla vs. CoT Prompting vs. Reasoning Models CoT prompting shows inconsistent benefits across model families. GPT models show minimal improvements or even degrade performance with CoT prompting, particularly on lie detectability tasks. GPT reasoning models also unexpectedly underperformed their regular models. DeepSeek models exhibited a different pattern, with reasoning variants outperforming both vanilla models and CoT-prompted versions across most question categories. Llama and Qwen families demonstrated no consistent pattern in response to either CoT prompting or reasoning-specialized models. These findings suggest that current reasoning enhancement techniques provide inconsistent benefits for ToM reasoning involving white lies, indicating the need to improve performance in this domain.

LLMs Struggle with True White Lie Understanding As described in Section 2.3, true white lie understanding requires models to identify falsity while recognizing prosocial motivation. However, as shown in Figure 4, model performance drops significantly on this combined task, with even the best models achieving $< 50\%$ accuracy. This suggests that models may succeed on individual dimensions by chance or through pattern matching, without integrating the complementary aspects required for genuine understanding. DeepSeek-v3 performs best but remains far from human-level competence. Given that psychological research shows second-order ToM reasoning as a necessary condition for white lie understanding (Happé, 1994), this result encourages further investigation into the second-order ToM reasoning capabilities of current LLMs.

LLMs Can Track Mental States But Fail to Apply Them in White Lie Contexts Our analysis reveals a performance gap between Info-State questions and White Lie Reasoning questions. While models track beliefs reasonably well, they struggle with questions requiring the application of these representations, particularly lie detectability where accuracy drops significantly. This pattern is consistent across all model families. This suggests two possibilities: either current LLMs possess mental state tracking abilities but cannot integrate these states to understand behavioral capabilities in white lie scenarios, or their apparent success in belief tracking may be superficial, lacking genuine second-order ToM reasoning needed to determine conditions for detecting deception.

4.5 In-depth Analysis

Common Sense Falsehoods Are Easier for Models Our analysis reveals performance differences across different white lie classes as shown in Figure 5. While Info-State questions show consistent performance, White Lie Understanding and Reasoning questions vary significantly. Models perform exceptionally well in Classes 1 and 2, this pattern suggests models use common sense knowledge as a shortcut rather than engaging in genuine contextual reasoning. For Class 1 scenarios involving globally recognized falsehoods (e.g., "Santa is real"), models can directly identify the statement

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

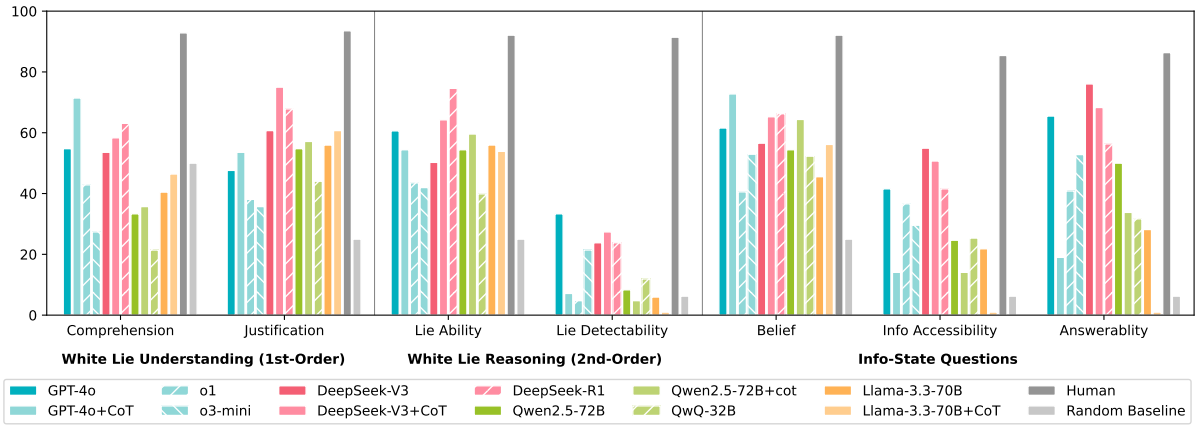

Figure 3: The answer accuracy of different LLM families on our benchmark: Comprehension[MCQs], Justification[MCQs], Lie Ability[MCQs], Lie Detectablity[list], Belief[MCQs], Info Accessibility[list], and Answerablity[list]

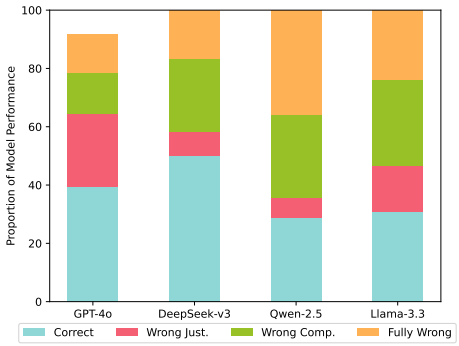

Figure 4: The proportion of model performance types in Justification questions.

Model	Level	Comp	Justi	B-2	LieAb	LieDe
GPT-4o	L-1	79.41	55.88	69.02	51.06	8.82
	L-2	70.37	59.26	73.58	63.08	7.41
	L-3	60.87	43.48	71.50	47.06	4.35
DeepSeek	L-1	79.41	88.24	62.63	65.96	26.47
	L-2	51.85	66.67	69.81	64.62	18.52
	L-3	34.78	65.22	55.56	58.82	39.13

Table 1: Performance (%) of GPT-4o and DeepSeek-v3 across levels. Abbreviations: Comp=Comprehension, Justi=Justification, B-2=2-order Belief, LieAb=Lie Ability, LieDe=Lie Detectability.

as false without complex belief reasoning. Similarly, Class 2 scenarios featuring symbolic explanations of sensitive topics (e.g., death) are recognizable through common patterns in the data. In contrast, scenarios requiring situation-specific reasoning without obvious common sense cues pose significantly greater challenges, highlighting that models still largely rely on statistical regularities rather than sophisticated ToM capabilities when navigating white lie understanding.

Surface-Level Detection vs. Motivation Understanding Table 1 shows a clear drop in model performance across three falsifiability levels (described in Section 2.2). DeepSeek-v3's comprehension, for instance, falls from Level-1 (79.41%) to Level-3 (34.78%). This reveals a critical insight: models excel at detecting lies through explicit contradictions but struggle to infer deception directly from motivations. A manual examination of DeepSeek's reasoning reveals the model primarily identifies lie detectability by checking which characters have access to what information. This pattern explains DeepSeek-v3's counterintuitive improvement in lie detectability for Level-3 (39.13%) compared to Level-1 (26.47%): without explicitly stated truths, the model faces less confusion about characters' information access but fails to recognize that genuine white lie detection requires understanding protective intentions, not merely contradiction recognition.

Models Struggle with Genuine Motivation Understanding Without Guidance To assess models' true comprehension of white lie motivations, we examined Justification question's free-form responses where no options provide hints. As shown in Table 2, models' performance drops significantly from MCQs to free-form responses. DeepSeek's falls from 75% to approximately 30% across different metrics. This gap suggests multiple-choice accuracy is inflated by provided options, as models struggle to independently infer the prosocial intentions behind white lies. Even the best-performing

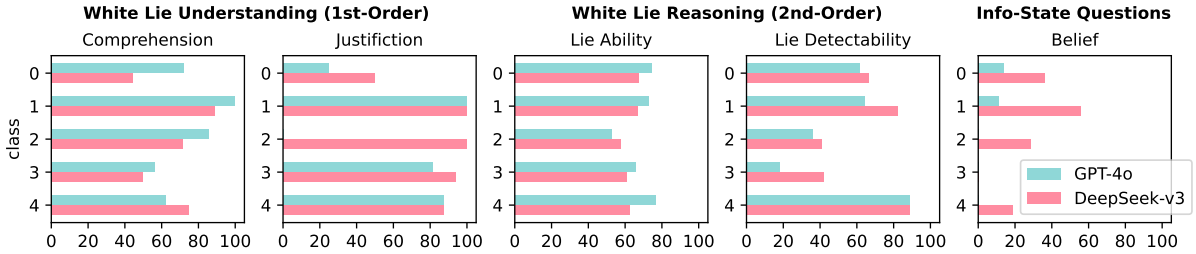

Figure 5: The performance of models across different classes.

Model	MCQs	FreeForm		
		Cos. Sim.	Token-F1	LLM-Judge
GPT-4o	53.57	22.62	27.38	16.67
DeepSeek	75.00	29.76	35.71	26.19
Qwen	57.14	19.05	9.52	25.00
Llama	46.43	20.24	10.71	23.81

Table 2: The accuracy of the model's CoT performance in Justification tasks under different task formats and evaluation methods.

models fail to identify emotional protection motivations in most free-form responses, highlighting significant limitations in their unprompted emotional reasoning.

5 Related Work

ToM in Psychology Second-order ToM is typically assessed through false-belief tasks and nested belief attribution (Wimmer and Perner, 1983; Quesque and Rossetti, 2020). Beyond belief tracking, having second-order ToM ability enables the interpretation of non-literal language, such as irony, sarcasm, and white lies, where the intended meaning diverges from the literal. Beaudoin et al. (2020) reviewed findings showing that accurate interpretation of such expressions depends on the listener's capacity to infer communicative intent and consider the speaker's emotional motivations. These forms of pragmatic inference are especially relevant in white lies, where the goal may be to avoid harm or maintain relationships (Erat and Gneezy, 2012). This reflects a broader understanding of ToM as a key mechanism for navigating complex social communication, supported by evidence from developmental, clinical, and neurocognitive studies (Baron-Cohen et al., 1985; Langley et al., 2022).

ToM in LLMs Most existing ToM evaluations focus on false-belief tests, such as the benchmarks ToMi (Nematzadeh et al., 2018), ToM-QA (Le et al., 2019), and FANToM (Kim et al., 2023), primarily testing whether models can track belief states when objects are moved or information changes. Other ToM-related benchmarks address narrative emotions and mental states (Rashkin et al., 2018; Sap et al., 2019), and some work has explored ToM in applied contexts (Chan et al., 2024; Bara et al., 2021). Within the framework of non-literal communication understanding, faux pas detection has been studied (Shapira et al., 2023b), but white lies remain largely understudied. While ToMBench (Chen et al., 2024) included white lie tests, it offers only 20 non-conversational samples, too limited for comprehensive evaluation. This limited understanding of white lie capabilities poses risks as LLMs are increasingly deployed in emotional support and caregiving applications where such skills are essential.

6 Conclusion

We present TactfulToM, an English ToM benchmark designed to evaluate LLMs' understanding of white lies through complex social scenarios. Our comprehensive evaluation reveals that even state-of-the-art LLMs underperform compared to humans in white lie understanding and reasoning, particularly in understanding the emotional motivation behind it. This performance gap raises ethical questions about LLMs' development: should LLMs understand white lies merely to interpret human behavior, or also to potentially generate them? The dilemma lies in choosing between strict truthfulness and social grace that might involve benign deception. TactfulToM provides a foundation for improving LLMs' social reasoning of white lie understanding, but we must carefully consider whether aligning LLMs completely with human social behaviors, including prosocially-motivated deception, is truly desirable for human-AI interaction.

Limitations

The main limitations of this paper are:

Limited to White Lies This dataset is primarily focused on white lie scenarios in order to analyze LLMs' ToM capabilities in such contexts. We do not extensively explore LLMs' other second-order ToM abilities; however, we hope that the methodology proposed in this paper can provide insights for future researchers seeking to construct related datasets.

Lack of Prior Impression In real-life situations, people typically possess prior knowledge and impressions of others. In our dataset, we deliberately constrained the scenarios such that the white lie triplets are not previously known to any of the involved roles, with the exception of the liar who initiates the deception. While this design choice helps isolate the ToM reasoning process, it does not fully capture the complexity of real-world social interactions. We consider incorporating this aspect of human cognition in our future work.

Limited Culture and Language Our benchmark includes only English-language data. However, in some other languages and cultures, communication tends to be more indirect, which may lead to different patterns of ToM reasoning in white lie scenarios.

Societal and Ethical Considerations

We acknowledge that our focus on white lies and Theory of Mind may raise concerns about anthropomorphizing AI systems. However, our research does not advocate for developing AI systems capable of telling white lies. Rather, we aim to systematically evaluate LLMs' social reasoning capabilities within specific informational contexts. Our results demonstrate that current models fall significantly short of human-like understanding in these scenarios, primarily relying on pattern matching rather than genuine understanding of mental states or intentions. We recognize the ethical complexities surrounding deception, even when prosocially motivated, and the particular sensitivity of developing AI systems with capabilities that could involve any form of misrepresentation.

All annotators participating in our data collection and validation were recruited through Prolific. We established fair compensation standards based on estimated task duration, ensuring payment rates above minimum wage requirements. We maintained transparent communication channels with annotators, promptly addressing questions and incorporating feedback to improve task instructions. All annotator data was anonymized, with only minimal identifiers stored securely and not included in the released dataset. We were careful to design our task instructions clearly, providing sufficient context without biasing responses. Annotators were informed about the academic research nature of the task and how their contributions would be used. When selecting annotators, we sought diversity across demographic factors to minimize potential biases in our data collection process, though we acknowledge that online recruitment platforms have inherent demographic limitations.

Our dataset is intended for research purposes only. While we have taken measures to ensure the conversations do not contain offensive content, research using generative models always carries a risk of unexpected outputs, particularly in free-form reasoning contexts. We encourage responsible use of our benchmark and dataset for advancing understanding of social reasoning in AI systems while remaining mindful of potential misapplications.

References

Hojjat Abdollahi, Mohammad H Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H Qualls. 2022. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. *IEEE Transactions on Affective Computing*, 14(3):2020–2032.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

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

DeepSeek-AI. 2024. Deepseek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

DeepSeek-AI. 2025. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sanjiv Erat and Uri Gneezy. 2012. White lies. *Management science*, 58(4):723–733.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. *arXiv preprint arXiv:2410.13648*.

Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in artificial intelligence*, 5:778852.

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

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. 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.

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023a. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

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

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Prompt Templates

A.1 Prompt Templates for Conversation Generation

Prompt templates for generating the conversation are listed below as Step_1, Step_2, Step_3, and Step_4.

Step_1

In this conversation, four characters are participating: A: the liar name, B: the target name, C: the accomplice name and D: the observer name.
Start the conversation with all participants discussing Topic for the scenario.
They are relationship descriptor, A: the liar name and C: the accomplice name are parents, B: the target name is kid, and D: the observer name is aunt/uncle.

Step_2

Now, B: the target name leaves the conversation because of the reason: leave reason B, while D: the observer name leaves the conversation because of the reason: leave reason D_1.
With B: the target name and D: the observer name absent, A: the liar name and C: the accomplice name continue their conversation, shifting the topic to 'real_reason_c'. Ensure the dialogue flows naturally without directly repeating these exact words.
Remember to indicate that B: the target name and D: the observer name are leaving the conversation.

Step_3

Now, only D: the observer name returns and rejoins the conversation after previously leaving. Note that B: the target name has not returned to the conversation yet.
Then the conversation continues and naturally shifts to situation_topic. Do NOT catch up or recap details from the previous conversation.
And then 'truth_c' is naturally revealed in the conversation.
After engaging briefly, D: the observer name leaves the conversation again because of the reason: leave reason D_2.
IMPORTANT:

- Clearly indicate through dialogue only (e.g., "I'm back," or "I have to leave again now.") that D: the observer name first returns and later departs again.
- Do NOT use third-person narrative or stage directions.
- Avoid mentioning details from previous conversations.

Step_4

Now B: the target name returns to the conversation after leaving the conversation.
First have B: the target name explicitly indicate the return through dialogue. Do NOT catch up or recap details from the previous conversation.
And then situation naturally unfolds. Make sure the dialogue flows naturally without directly repeating these exact words.
In response, A: the liar name and C: the accomplice name tells B: the target name that 'the lie'.
IMPORTANT:

- Do NOT use third-person narrative or stage directions.
- Avoid mentioning details from previous conversations.

A.2 Question Generation Templates

The question generation templates we used are provided below as examples in Table (**?**).

A.3 Prompt Templates for Model Evaluation

Prompt templates used for model evaluation are listed in Table 4.

A.4 Leaving Reason List

1. have to return a borrowed item
2. have unexpected visitor
3. need to quickly tidy the room before another meeting
4. have to refill my water bottle
5. remembered to submit some papers

6. must respond to a phone call
7. forgot to run errands
8. coffee break
9. remembered to take care of something urgent
10. need to grab a quick snack

B Model Performance on All Tasks

Detailed scores of the model performance on all tasks are provided in Table 5.

C Wrong Option Design Details

C.1 Belief Statement Options

For second-order belief statements, we formalized four logically distinct cases using belief operators:

- $Bel_Z(\varphi)$: Person Z believes proposition φ
- $\neg Bel_Z(\varphi)$: Person Z is unaware of (or does not believe) φ

A second-order belief statement takes the form $Bel_X(\cdot)$, where the inner argument concerns Y's epistemic state about proposition p:

$Bel_X(Bel_Y(p))$	(X thinks Y thinks p)
$Bel_X(\neg Bel_Y(p))$	(X thinks Y is unaware of p)
$\neg Bel_X(Bel_Y(p))$	(X is unaware that Y thinks p)
$\neg Bel_X(\neg Bel_Y(p))$	(X is unaware that Y is unaware of p)

When "X thinks Y thinks p" is supported by the dialogue, we use $Bel_X(Bel_Y(p))$ as the correct answer. The remaining expressions serve as distractors representing three error types:

- Wrong attribution of Y's first-order belief
- Wrong attribution of X's meta-belief
- Simultaneous error in both belief layers

C.2 Justification Options

For justification questions, we employed few-shot prompting with the following criteria:

- Correct answer must reflect the genuine prosocial motivation (e.g., sparing feelings, maintaining harmony)
- Wrong options:
  - Mutually exclusive
  - Plausible alternative explanations
  - Consistent with the dialogue context

Example:

> Correct: "Because Jamey wants to politely decline without making Pearl feel bad about choosing an expensive restaurant."
>
> Wrong options:
>
> - "Because Jamey actually has to work this weekend."
> - "Because Jamey dislikes Pearl and doesn't want to spend time with him."
> - "Because Jamey already has dinner plans with someone else."

D More Analysis

We conducted detailed error analyses by tracking the specific wrong options selected by models across different question types. These analyses provide deeper insights into the reasoning patterns and failure modes of various LLMs when handling white lie scenarios. The distribution of error types for Lie Ability questions (Figure 6), Belief Understanding (Figure 7), and Role-Specific Performance in Lie Detection (Figure 9) reveal systematic patterns in how models misunderstand white lie contexts. These visualizations complement our main findings by illustrating specific misconceptions about mental state attribution and prosocial motivations.

D.1 Across Class Performance with All Models

E Dataset Details

E.1 Dataset Distribution

The proportion distribution of different classes within the TacfulToM dataset is shown in Figure 11.

E.2 An Example from TactfulToM

We provided a full conversation sample from TactfulToM below for reference:

Pearl: So, I was thinking about food and I'm curious, what's everyone's favorite cuisine? I absolutely love Italian, especially a good risotto. It just feels like a warm hug in a bowl!

Jamey: Oh, Italian is great! But for me, it's definitely Thai food. I love the bold flavors and the perfect balance of sweet, sour, and spicy. Pad Thai is my absolute favorite.

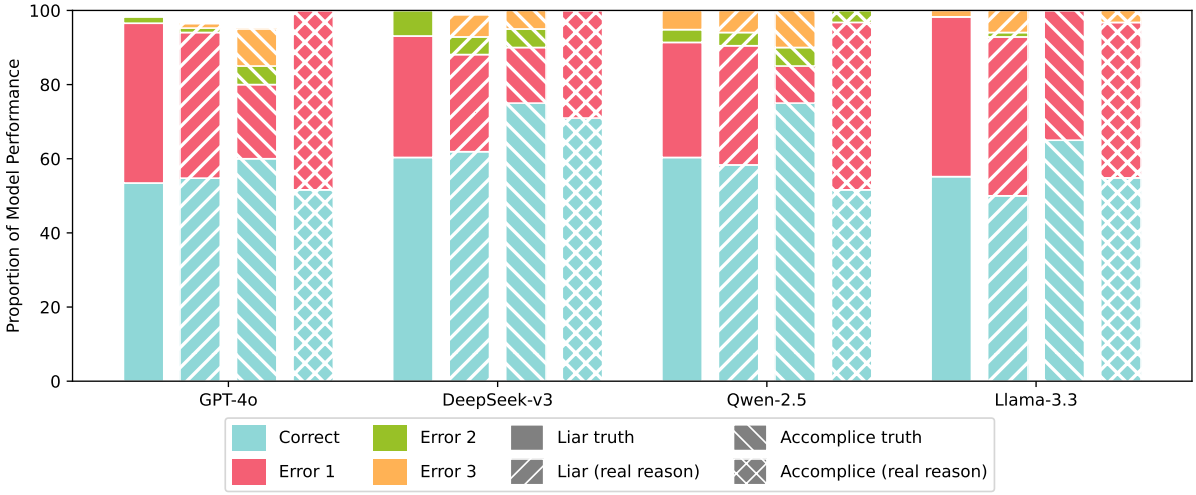

Figure 6: The proportion of model performance types in Lie Ability questions.

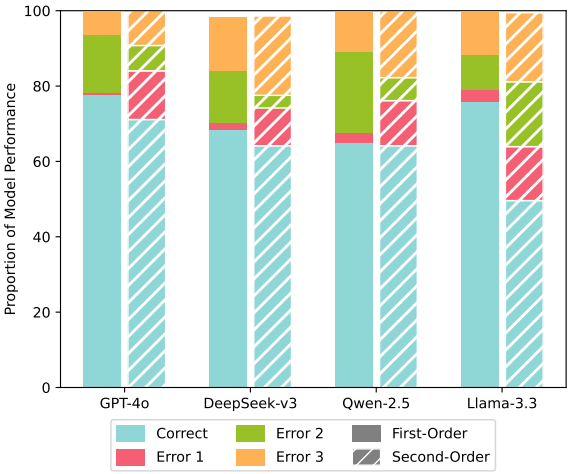

Figure 7: The proportion of model performance types in BeliefQA questions.

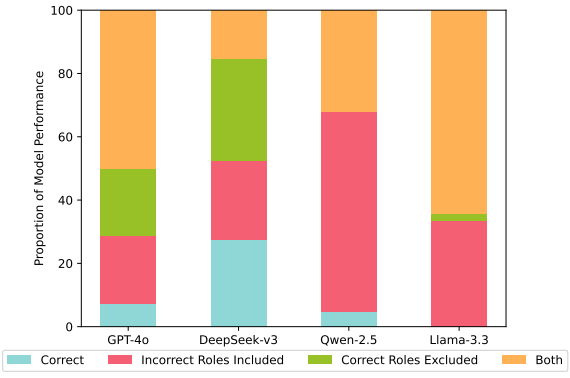

Figure 8: The proportion of model performance types in Lie Detectability questions of list format.

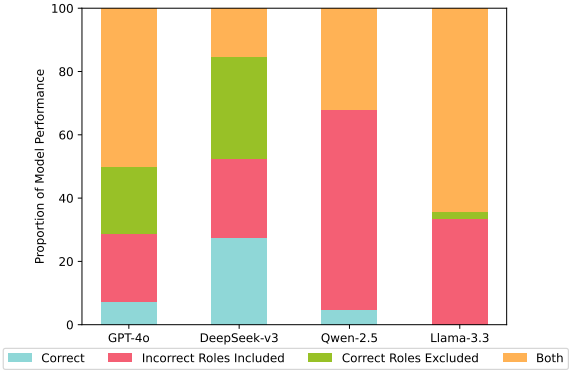

Figure 9: The proportion of model performance types in Lie Detectability questions of list format.

Marithza: Thai is amazing, Jamey! For me, it's all about Japanese cuisine. I can't get enough of sushi and sashimi. It's so fresh and delicate, plus there's something exciting about eating with chopsticks.

Donyae: Oh, all great choices! I'm all in for Mexican food. Tacos, enchiladas, and don't even get me started on the guacamole. It's just so satisfying.

Pearl: Yum, I could eat some good guac every day. Has anyone tried making it from scratch? It's surprisingly easy and so much better than store-bought.

Jamey: I have! I actually make a mean guacamole—and you're right, Pearl, fresh ingredients make all the difference. It's a hit whenever I have friends over.

Marithza: I bet! Guacamole's a crowd pleaser for sure. On another note, I tried making homemade sushi once. Let's say, rolling those little vinegared rice gems isn't as easy as it looks!

Donyae: I can imagine! But doing it yourself must be fun. Maybe we should have a cooking

Figure 10: The performance of models across different classes.

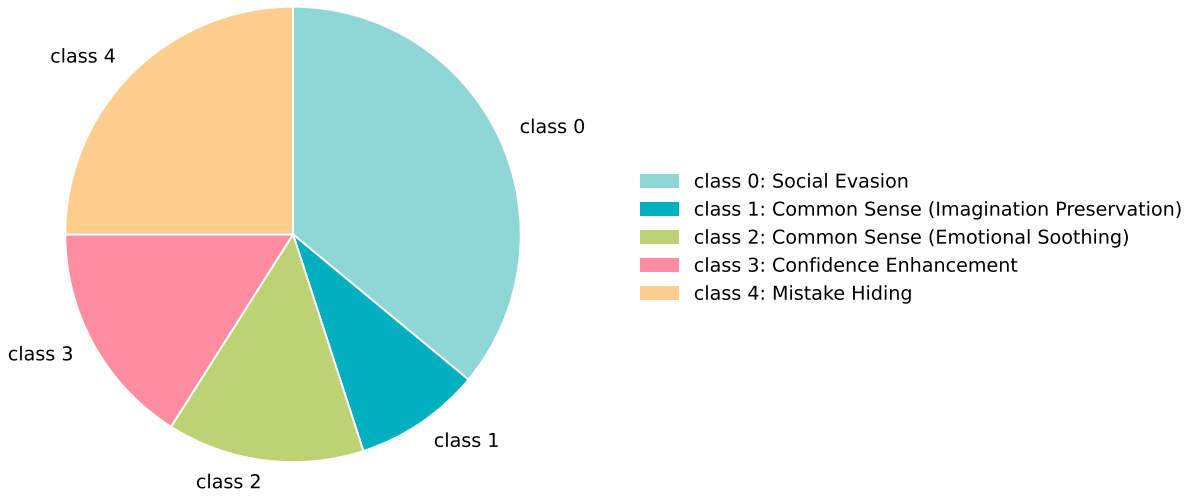

Figure 11: Proportion of different classes in TacfulToM.

night sometime, where everyone brings a dish from their favorite cuisine.

Pearl: That's a fantastic idea, Donyae! We could even make it like a friendly mini competition. Each dish gets a few points for taste and presentation. What do you guys think?

Jamey: I'm totally in! It'll be a fun way to try different cuisines and maybe even learn a few new recipes from each other.

Donyae: Awesome! I just remembered, I need to submit some papers before the end of the day. I'll catch you guys later, though. This cooking night sounds like such a blast—keep me posted!

Pearl: Oh, I need to make a quick phone call. Let's definitely plan this out soon! Bye for now, everyone.

Marithza: Bye, Donyae and Pearl! We'll keep you in the loop. So, Jamey, what kind of place were you thinking about for our cooking night? We could do it at someone's house or rent out a small venue.

Jamey: Honestly, I love the idea of just doing it at someone's house. It keeps things cozy and informal. Kind of like why I prefer casual dining

places over the upscale spots. There's something special about a laid-back atmosphere.

Marithza: I know what you mean. Sometimes those upscale places can feel a bit too… stiff, maybe? I guess it's nice to just unwind and not worry about which fork to use—chat over some good food without all the fancy stuff.

Jamey: Exactly! Plus, it can be pretty expensive going out to those places too often. I'd much rather hit up a cozy little diner or a local joint where you can really taste the love put into the cooking.

Marithza: For sure! And there's something so genuine about finding that hole-in-the-wall spot with the best food. It kind of feels like discovering a secret everyone should know about.

Donyae: Hey everyone, I'm back! I finally submitted all the papers. What did I miss?

Marithza: Welcome back, Donyae! We were just talking about where to hold our cooking night. Jamey and I thought doing it at someone's house might be fun and cozy.

Donyae: That sounds perfect! I think the home setting adds a nice personal touch. Plus, it's easier to relax and enjoy each other's company.

Jamey: Absolutely, home-cooked vibes really can't be beat. Do you have a dish in mind for when we do this?

Donyae: I'm thinking something classic like enchiladas. They are always a hit and not too complicated. By the way, I have to leave again now—I have a meeting starting soon that I need to prepare for.

Marithza: No worries, Donyae. Thanks for popping back in! Let us know how the meeting goes.

Jamey: Yeah, hope all goes well, Donyae! We'll continue planning and keep you updated.

Pearl: Hey everyone, I'm back from my call! What did I miss in the world of culinary planning?

Jamey: Welcome back, Pearl! We were chatting about turning our houses into five-star restaurants for our cooking night.

Pearl: Sounds exciting! But speaking of five-star, there's this new Italian restaurant that just opened downtown. Would you all be interested in checking it out this weekend?

Jamey: That sounds amazing, Pearl, but unfortunately, I have to work this weekend, so I can't make it. Maybe next time?

Pearl: Oh no, that's a bummer! We'll definitely catch up soon then. Maybe for that cooking night we talked about—we can even bring some Italian-inspired dishes to you instead.

Marithza: I'm up for the restaurant visit if it's still on. I've been dying to try their truffle pasta from what I've heard.

Pearl: Awesome, Marithza! Let's make it a date then. We'll let Donyae know and hopefully, she can join us too.

Jamey: You guys enjoy it! Be sure to save me a slice of that truffle pasta, at least in spirit.

Marithza: We will! And we'll definitely share all the delicious details with you. Catch up soon, Jamey!

Pearl: For sure, Jamey. Good luck with work, hope the weekend goes smoothly!

F Instructions Given To Participants

Our participants were recruited on the Prolific platform. They met the following criteria: English as their first and primary language, fluency in English, and completion of an undergraduate degree (BA/BSc/other). Additionally, participants had an approval rate between 95–100%. We paid participants 9 pounds per hour, which is considered appropriate according to Prolific's standards. This payment level ensures fair compensation given the demographic of participants, predominantly located in English-speaking regions.

G AI usage

In this project, we used LLMs for assistance. During paper writing, we used models from GPT and Claude families to help us refine and enhance our expressions. For programming, we also relied on models from GPT family to generate reference code, which we subsequently modified to complete our tasks. We also used recraft and GPT for icons generation used in Figure 1 and Figure 2

Thank you for considering participating in our study. ***In this study, you will be presented with a conversation followed by ~15 questions. Your task is to answer each question based on the content of the conversation. Please read both the conversation and the questions carefully. Take your time and do your best to provide accurate answers.***

Please read the following information carefully before providing your consent.

1. Data Collection:

- We will collect your responses to the questions provided.
- **No personal data will be collected.** All responses are anonymous.

2. Data Usage:

- Your annotations will be used for research purposes to improve the quality of our white lies conversation dataset.

3. Data Storage:

- Annotations will be securely stored in our database.
- Data will be retained for research purposes for up to 5 years.

4. Anonymity&Data Sharing:

- All responses are anonymous, and no identifying information will be linked to your annotations.
- Anonymized data may be made available to other researchers in the future.

5. Withdrawal of Consent:

- You may withdraw from the study at any time before submission by closing the form.
- Once submitted, your anonymous annotations cannot be traced back for removal.

6. Legal Framework:

- Data collection and storage comply with GDPR and all applicable privacy laws.

Important Notice: There is a small risk that some conversations may contain non-safe content, such as offensive or sensitive material.

Figure 12: The Instruction Provided for annotators.

Question Type	Template
Comprehension	f"Was what {liar} told {target}, {lie}', true?"
Justification	fWhy did {liar} and {accomplice} tell {target} {lie}'?"
Lie Ability	f"Please choose the correct answer from the options below:" Correct: f"{liar} tell {target} that {lie}' because {liar} believes that {target} is unaware that {real reason}." Wrong 1: f{liar} tell {target} that {lie}' because {liar} believes that {target} believes {real reason}." Wrong 2: f"{liar} tell {target} that {lie}' because {liar} is unaware that {target} is unaware that {real reason}." Wrong 3: f{liar} tell {target} that {lie}' because {liar} is unaware that {target} believes {real reason}."
Lie Detectability	f"Was what {liar} told {target}, {lie}', true?" fList all the characters who know the precise correct answer to this question."
Fact	fWho discussed {real reason}'?" f"Who discussed {truth}'?"
Belief	fWhat does {character} believe about {real reason}'?" f"What does {character} believe about {truth question}'?" fWhat does {liar} believe about {target}'s belief regarding{real reason}'?"
Info Accessibility (List)	f"List all the characters who know this information: {real reason}"
Info Accessibility (Binary)	f"Does {character} know this information: {real reason}?"
Answerability (List)	f"Who discussed {real reason}'?" fList all the characters who know the precise correct answer to this question."
Answerability (Binary)	fWho discussed {real reason}'?" f"Does {character} know the precise correct answer to this question?"

Table 3: Question generation templates Examples for different question types in TactfulToM.

Answer Type	Prompt Template
Binary	You are an expert in social reasoning. Answer the following question with 'Yes' or 'No'. Remember: Your answer should ONLY include 'Yes' or 'No' with nothing else. # Context: {} # Question: {} (Let's think step by step:)
MCQs	You are an expert in social reasoning. Answer the following question with the option number of the most appropriate answer. Remember: Your answer should ONLY include the option number with nothing else. # Context: {} # Question: {} # Options: {} (Let's think step by step:)
List	You are an expert in social reasoning. List the required items and split them with commas. Remember: Your answer should ONLY include the required items splited by commas with nothing else. # Context: {} # Question: {} (Let's think step by step:)
Freeform	You are an expert in social reasoning. Answer the following question with a single sentence. # Context: {} # Question: {} (Let's think step by step:)

Table 4: The prompt templates for model evaluation. The CoT prompt template additionally includes the instruction "Let's think step by step: ".

Model	Class	Comp		Justification		LieAb.	Lie Detectability		Belief		Info Accessibility		Answerablity		FactReason		FactTruth	
		MCQs	Free	MCQs	Free	MCQs	List	Binary	MCQs	Free	List	Binary	List	Binary	MCQs	Free	MCQs	Free
Human		92.85	-	93.54	-	-	-	-	-	90.48	-	85.42	-	86.36	-	92.1	91.43	-
GPT-4o	0	41.67	19.44	11.11	22.22	60.26	36.11	51.85	62.36	41.95	34.48	75.97	62.07	62.07	100.0	63.89	86.36	63.64
+*CoT*		72.22	25.0	25.0	16.67	61.54	13.89	62.04	74.86	49.71	20.69	88.96	15.52	15.52	100.0	58.33	63.64	68.18
o1		30.56	75.0	5.56	33.33	34.62	8.33	36.11	31.18	32.76	36.21	48.7	39.66	39.66	50.0	41.67	31.82	27.27
o3-mini		16.67	11.11	2.78	5.56	38.46	30.56	71.3	51.87	36.49	27.59	68.83	51.72	51.72	97.22	55.56	59.09	50.0
DeepSeek-V3		36.11	13.89	25.0	19.44	51.28	27.78	57.41	57.9	40.23	41.38	65.58	82.76	82.76	100.0	63.89	68.18	54.55
+*CoT*		44.44	38.89	50.0	27.78	66.67	36.11	65.74	67.53	54.17	60.34	88.31	75.86	75.86	94.44	69.44	50.0	72.73
DeepSeek-R1		52.78	25.0	44.44	19.44	73.08	27.78	70.37	66.95	49.43	36.21	87.66	56.9	56.9	100.0	58.33	77.27	50.0
Qwen2.5		8.33	27.78	8.33	25.0	55.13	5.56	73.15	47.7	36.35	24.14	64.29	44.83	44.83	100.0	58.33	54.55	54.55
+*CoT*		16.67	33.33	11.11	11.11	62.82	5.56	65.74	63.36	39.37	18.97	61.69	24.14	24.14	97.22	83.33	50.0	45.45
QwQ		2.78	61.11	8.33	33.33	14.1	5.56	27.78	35.92	28.74	12.07	44.16	20.69	20.69	47.22	25.0	27.27	13.64
Llama-3.3		19.44	19.44	8.33	5.56	60.26	11.11	75.0	45.55	48.28	18.97	74.03	39.66	39.66	100.0	55.56	59.09	63.64
+*CoT*		41.67	38.89	16.67	16.67	48.72	0.0	63.89	59.63	44.97	0.0	76.62	0.0	0.0	100.0	58.33	77.27	59.09
GPT-4o	1	100.0	55.56	100.0	11.11	82.14	66.67	55.56	67.86	36.31	50.0	75.0	78.57	78.57	100.0	77.78	80.0	60.0
+*CoT*		100.0	77.78	100.0	0.0	64.29	11.11	62.96	75.0	40.48	7.14	78.57	28.57	28.57	88.89	77.78	60.0	60.0
o1		100.0	77.78	100.0	33.33	92.86	0.0	100.0	67.26	39.29	42.86	71.43	50.0	50.0	100.0	66.67	80.0	60.0
o3-mini		88.89	77.78	100.0	44.44	75.0	44.44	55.56	58.93	30.95	35.71	64.29	64.29	64.29	100.0	77.78	40.0	40.0
DeepSeek-V3		88.89	77.78	100.0	44.44	75.0	55.56	59.26	64.88	33.93	50.0	67.86	50.0	50.0	100.0	33.33	60.0	20.0
+*CoT*		88.89	77.78	100.0	22.22	82.14	55.56	55.56	68.45	45.24	35.71	75.0	64.29	64.29	88.89	88.89	60.0	60.0
DeepSeek-R1		100.0	66.67	100.0	11.11	100.0	44.44	59.26	74.4	40.48	35.71	82.14	50.0	50.0	100.0	55.56	80.0	40.0
Qwen2.5		100.0	77.78	100.0	22.22	71.43	22.22	81.48	70.24	34.52	21.43	53.57	42.86	42.86	100.0	88.89	40.0	40.0
+*CoT*		88.89	77.78	100.0	0.0	75.0	22.22	62.96	72.62	42.26	0.0	75.0	35.71	35.71	100.0	88.89	20.0	60.0
QwQ		100.0	88.89	100.0	11.11	92.86	22.22	92.59	73.21	35.12	50.0	71.43	42.86	42.86	100.0	77.78	80.0	40.0
Llama-3.3		100.0	88.89	100.0	0.0	64.29	0.0	88.89	55.36	35.12	14.29	71.43	21.43	21.43	88.89	55.56	60.0	60.0
+*CoT*		100.0	77.78	100.0	0.0	67.86	0.0	77.78	61.31	37.5	0.0	42.86	0.0	0.0	100.0	77.78	80.0	60.0
GPT-4o	2	85.71	28.57	0.0	14.29	36.36	28.57	38.1	56.06	43.18	54.55	68.18	72.73	72.73	100.0	14.29	50.0	50.0
+*CoT*		85.71	57.14	0.0	14.29	36.36	0.0	38.1	59.09	46.97	18.18	68.18	9.09	9.09	85.71	0.0	50.0	50.0
o1		100.0	57.14	0.0	14.29	36.36	0.0	66.67	52.27	48.48	36.36	59.09	45.45	45.45	85.71	14.29	75.0	50.0
o3-mini		85.71	28.57	0.0	14.29	45.45	14.29	33.33	50.76	31.06	18.18	45.45	36.36	36.36	57.14	28.57	50.0	50.0
DeepSeek-V3		85.71	42.86	100.0	57.14	54.55	14.29	33.33	49.24	45.45	54.55	72.73	81.82	81.82	100.0	42.86	50.0	25.0
+*CoT*		71.43	71.43	100.0	14.29	40.91	28.57	47.62	59.85	41.67	45.45	68.18	54.55	54.55	85.71	28.57	50.0	50.0
DeepSeek-R1		71.43	85.71	100.0	28.57	68.18	28.57	61.9	59.09	50.0	63.64	81.82	72.73	72.73	85.71	28.57	75.0	50.0
Qwen2.5		57.14	14.29	100.0	28.57	40.91	0.0	71.43	56.06	41.67	18.18	59.09	54.55	54.55	100.0	14.29	75.0	50.0
+*CoT*		57.14	42.86	100.0	14.29	27.27	0.0	61.9	59.85	44.7	9.09	63.64	27.27	27.27	100.0	28.57	75.0	50.0
QwQ		14.29	100.0	100.0	42.86	4.55	0.0	4.76	1.52	22.73	0.0	9.09	0.0	0.0	0.0	14.29	0.0	0.0
Llama-3.3		100.0	42.86	100.0	14.29	63.64	14.29	47.62	41.67	40.15	18.18	77.27	45.45	45.45	85.71	28.57	75.0	50.0
+*CoT*		100.0	42.86	100.0	14.29	50.0	0.0	85.71	53.79	43.18	0.0	36.36	0.0	0.0	100.0	28.57	50.0	50.0
GPT-4o	3	43.75	12.5	87.5	31.25	31.58	25.0	33.33	57.03	38.28	65.62	73.33	65.62	65.62	100.0	18.75	68.75	0.0
+*CoT*		56.25	12.5	81.25	31.25	18.42	0.0	41.67	70.05	38.28	3.12	72.22	18.75	18.75	100.0	12.5	62.5	0.0
o1		25.0	62.5	81.25	50.0	7.89	0.0	16.67	23.96	34.9	12.5	34.44	15.62	15.62	50.0	12.5	37.5	0.0
o3-mini		6.25	6.25	93.75	43.75	34.21	12.5	33.33	53.65	29.69	43.75	70.0	53.12	53.12	100.0	12.5	56.25	6.25
DeepSeek-V3		43.75	12.5	93.75	43.75	21.05	12.5	33.33	56.51	36.46	78.12	68.89	68.75	68.75	100.0	12.5	56.25	0.0
+*CoT*		50.0	25.0	93.75	37.5	42.11	0.0	35.42	64.84	46.35	40.62	81.11	65.62	65.62	100.0	50.0	50.0	0.0
DeepSeek-R1		56.25	31.25	81.25	18.75	57.89	6.25	41.67	64.58	42.71	65.62	75.56	62.5	62.5	100.0	6.25	75.0	0.0
Qwen2.5		31.25	12.5	93.75	31.25	36.84	12.5	41.67	60.42	36.2	46.88	65.56	71.88	71.88	100.0	12.5	43.75	0.0
+*CoT*		37.5	31.25	93.75	37.5	55.26	0.0	39.58	65.1	37.24	21.88	71.11	43.75	43.75	100.0	87.5	43.75	6.25
QwQ		31.25	12.5	81.25	50.0	55.26	31.25	45.83	72.4	38.02	46.88	83.33	53.12	53.12	100.0	6.25	56.25	6.25
Llama-3.3		31.25	18.75	87.5	31.25	31.58	0.0	62.5	41.15	38.02	37.5	73.33	15.62	15.62	93.75	6.25	50.0	0.0
+*CoT*		18.75	37.5	100.0	25.0	36.84	0.0	45.83	50.78	33.59	0.0	62.22	3.12	3.12	100.0	6.25	68.75	0.0
GPT-4o	4	56.25	6.25	81.25	31.25	100.0	18.75	39.58	64.2	47.53	18.52	72.84	62.96	62.96	93.75	62.5	100.0	18.18
+*CoT*		62.5	12.5	87.5	43.75	88.89	0.0	52.08	75.93	54.01	14.81	82.72	25.93	25.93	93.75	68.75	81.82	36.36
o1		31.25	18.75	50.0	37.5	74.07	6.25	70.83	61.73	53.4	62.96	90.12	66.67	66.67	93.75	75.0	90.91	63.64
o3-mini		12.5	6.25	31.25	43.75	25.93	0.0	52.08	52.16	35.8	18.52	71.6	55.56	55.56	93.75	50.0	45.45	36.36
DeepSeek-V3		68.75	6.25	68.75	37.5	59.26	12.5	37.5	52.47	43.83	59.26	62.96	81.48	81.48	93.75	68.75	81.82	45.45
+*CoT*		75.0	18.75	87.5	37.5	88.89	18.75	50.0	61.42	48.46	51.85	79.01	62.96	62.96	93.75	68.75	81.82	36.36
DeepSeek-R1		68.75	6.25	75.0	37.5	81.48	18.75	62.5	65.43	48.15	18.52	79.01	44.44	44.44	93.75	68.75	100.0	54.55
Qwen2.5		43.75	6.25	75.0	31.25	70.37	6.25	52.08	52.78	44.44	3.7	76.54	37.04	37.04	93.75	68.75	63.64	36.36
+*CoT*		37.5	31.25	81.25	31.25	66.67	0.0	60.42	63.27	46.3	3.7	74.07	44.44	44.44	87.5	87.5	63.64	36.36
QwQ		12.5	0.0	31.25	31.25	66.67	6.25	64.58	73.46	46.6	25.93	85.19	37.04	37.04	87.5	68.75	81.82	54.55
Llama-3.3		37.5	18.75	87.5	31.25	62.96	0.0	58.33	47.22	51.85	14.81	76.54	14.81	14.81	93.75	62.5	72.73	54.55
+*CoT*		31.25	12.5	81.25	37.5	81.48	0.0	66.67	53.4	45.37	3.7	70.37	0.0	0.0	93.75	68.75	100.0	54.55

Table 5: The performance of different LLM families on our benchmark dataset.