DIG-DIS: Transformer-Based Models vs State Space Models on Complex Theory of Mind Reasoning Tasks

Aayushya Patel* **Maximilian Prince** **Sahasra Kalakonda** **Adithri Manda**
Hannah You **Aslihan Akalin** **Kevin Zhu** **Sean O'Brien** **Vasu Sharma**
Algoverse AI Research
kevin@algoverse.us

Abstract

Transformer architectures have become the backbone of domains such as language modeling but often suffer in many inference settings due to their quadratic self-attention. Recently proposed subquadratic architectures, such as State Space Models (SSMs), reveal promising results ranging from more efficient compute to outperforming transformer-based models on certain benchmarks. Theory of Mind (ToM)–the ability to reason about one's own and other's mental states–is a complex task that poses significant challenges to state-of-the-art models. While previous works have focused on ToM reasoning solely in transformer-based models, we evaluate a leading transformer-based model and a leading SSM on various long context ToM tasks. Through our evaluation, we reveal that although the SSM generally performs worse on ToM tasks compared to the transformer-based model, the SSM exhibits less performance degradation on more difficult tasks, indicating potential advantages for SSMs in difficult and long-context ToM reasoning.

1 Introduction

Transformers have been considered the state-of-the-art architecture for Large Language Models (LLMs) (Wolf et al., 2020) in tasks such as language, image, and video processing (Raffel et al., 2023; Dosovitskiy et al., 2021; Chen et al., 2023). Notably, GPT-4o, as one of the most advanced LLMs to date (Bubeck et al., 2023), has shown high proficiency in tasks requiring few-shot learning (Shahriar et al., 2024), part of this success stems from the self-attention mechanisms that transformers benefit from (Vaswani et al., 2023). However, this self-attention mechanism has led to a limitation of the transformer architecture itself (Keles et al., 2022). This mechanism requires computations that scale quadratically with the length of the input sequence, creating challenges in computational efficiency and higher demands with long sequence processing (Tay et al., 2020). Additionally, transformer architectures have performed poorly on abstract reasoning and fluid intelligence tasks, such as the ConceptARC Benchmark (Moskvichev et al., 2023; Kumar et al., 2023).

State Space Models (SSMs) are sequence models that are designed to process long-range dependencies (Gu et al., 2021). Unlike transformers, SSMs maintain structured latent states that continuously update, allowing them to process long sequences with reduced computational demands (Patro and Agneeswaran, 2024). (Iapăscurtă et al., 2024)

Theory of Mind (ToM) is the cognitive ability to attribute mental states, enhancing fundamental capability for human-like reasoning (Premack and Woodruff, 1978; Apperly, 2010; Baker et al., 2017). The ability of language models to interpret ToM is crucial for improving a model's ability to anticipate intentions, reduce misinterpretations, and enhance safety in applications like self-driving cars and virtual assistants (Kosinski, 2024). Despite the evolving improvements in ToM capabilities of transformer models, they remain significantly far behind compared to human proficiency (Xu et al., 2024). This gap has opened up many possibilities for alternatives for long-context reasoning, where beliefs must dynamically update based on new information (Rabinowitz et al., 2018b; Peskin, 1992).

Belief modeling, part of human-like reasoning, is still an area with significant gaps in dealing with embedded irrelevant information (distractors), ensuring temporal consistency, as well as changing from formatting to realistic conversation dialogues (Zhang et al., 2021, 2024).

1.1 Contributions

In this paper, we contribute:

- New evaluations on the OpenToM dataset for

*Lead Author

the models GPT-4o Mini and Phi-Mamba.

- Evaluations for new alterations of the OpenToM dataset, introducing the Dig-Dis benchmark formed by two datasets and a framework for a third:
  - Di-ToM (Dialogue-Theory of Mind) – dialogue formatting.
  - Dis-ToM (Distractor-Theory of Mind) – interjected irrelevant information.
  - TC-ToM (Temporal Consistency-Theory of Mind) - changes in temporal consistency.
- These datasets and code for replicating the evaluations will be made public. The benchmark contains 596 unique stories across Di-ToM and Dis-ToM datasets.
- Analysis of results and inferences on the effect of architectural differences on evaluation metrics.

2 Related Works

Long-Context Processing and Belief Stability: Transformers consist of self-attention mechanisms enabling them to process long-range dependencies (Brown et al., 2020; Dasgupta et al., 2022; Bubeck et al., 2023). However, as context length increases, transformers degrade due to attention-collapsing (Liu et al., 2021; Press et al., 2022). Various adaptations have been proposed, such as sparse attention, but their effectiveness remains an active area of research (Zaheer et al., 2020; Ainslie et al., 2023; Bulatov et al., 2023). In contrast, State Space Models (SSMs) operate sequentially, but their ability to maintain belief consistency over extended reasoning chains is less understood (Gu et al., 2021; Goel et al., 2022). This study assesses the limitations of transformers and SSMs retaining belief consistency across long-context reasoning tasks.

Resilience to Distractors: In real-world interactions, agents must distinguish between relevant and irrelevant information, filtering out distractions while preserving core belief representations (Rabinowitz et al., 2018a; Baker et al., 2017). Transformers are resilient to distractors, yet remain vulnerable to adversarial inputs that manipulate contextual understanding (Zellers et al., 2019; Lin et al., 2021). Research has shown that even state-of-the-art transformer models can be misled by subtle perturbations (Geva et al., 2023; He et al., 2023a). SSMs process information incrementally, making them more susceptible to distractors that disrupt belief trajectories over time (Gu et al., 2021; Mitchell et al., 2022). This study systematically evaluates the susceptibility of transformers and SSMs to distractors in belief tracking tasks.

Multiagent Interactions and Conversational ToM: Conversational ToM requires reasoning about nested mental states across multiple turns,where beliefs evolve through dialogue exchanges (Dinan et al., 2020; Zhou et al., 2021). Although transformers have been used in dialogue systems and belief tracking, their ability to handle multi-agent recursive beliefs remains underexplored (Shuster et al., 2022; Sun et al., 2023). This study assesses how well transformers and SSMs maintain recursive beliefs within real-world conversational contexts, by framing belief-tracking tasks as interactive dialogues.

Evaluating Model Scale in ToM and Long-Context Reasoning: Large-scale transformers, such as GPT-4o, require significant computational resources for long-context tasks (Brown et al., 2020; Tay et al., 2022). Smaller models, such as GPT-4o mini, face challenges in recursive belief tracking due to reduced parameter capacity (Liu et al., 2021; Gururangan et al., 2023). While SSMs, like Phi-Mamba, offer promising scalability advantages, their ability to model deep recursive ToM reasoning is limited (Gu et al., 2021; Peng et al., 2023). This study examines the trade-offs between model efficiency, long-context reasoning, and belief-tracking fidelity.

3 Methodology

SSM's ability to model long-range dependencies matches the recursive nature of reasoning in ToM tasks (He et al., 2023b). Motivated by these advancements, we explored the SSM framework, in particular, to address local problems requiring contextual depth and nuanced reasoning in the ToM domain. We use Phi-Mamba, an SSM distilled from the transformer-based model Phi-1.5K.(Bick et al., 2024)

To compare the performance of Phi-Mamba, we chose Gpt-4o mini as a representative transformer-based model due to its similar token size to Phi-

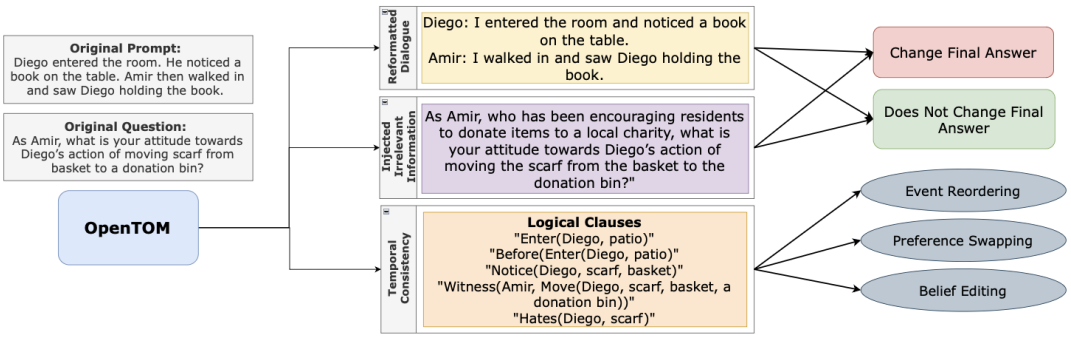

Figure 1: Process utilized for OpenToM dataset

Mamba.

3.1 Distractors Construction

An entry from the original OpenToM consists of two characters, an entity-of-interest (EOI), and several locations and containers. One of the characters will be a "mover" who may shift the EOI. The other is stated to be the witness who may or may not view the mover take action. Each character has a set of beliefs and values which they follow.

Distractors consist of information that may or may not be relevant to the text. The task was then divided into 2 sections, one that would contain distractors distorting the final answer, and the other with distractors that do not change the final answer. We use the OpenToM dataset as the baseline to then add distractors within the prompt, to test how the addition of distractors within the prompt would effect the models response.

Original Question Format	**Reformatted Question with Distractor**
As Amir, what is your attitude towards Diego's action of moving scarf from basket to a donation bin?	As Amir, who has been encouraging residents to donate items to a local charity, what is your attitude towards Diego's action of moving the scarf from the basket to the donation bin?"

Table 1: Adding distractor information into prompting

3.2 Dialogue Format

In order to test how different models respond to a dialogue format, the dialogue structure was used. The OpenToM data set was manipulated to reformat each scenario as a conversation.

Original Prompt	**Reformatted Dialogue**
Diego entered the room. He noticed a book on the table. Amir then walked in and saw Diego holding the book.	**Diego:** I entered the room and noticed a book on the table. **Amir:** I walked in and saw Diego holding the book.

Table 2: Shift from narrative to dialogue format.

This was done by assigning each person to either be an "observer" or a "mover". Then we had them form their own set of dialogues based on the prompt. Next, we reorganized each scenario into a dialogue format, where the name of the observer and mover would be replaced with the names of the characters.

3.3 Temporal Consistency

We constructed a systematic evaluation method which analyzed how well the models utilized their capabilities to deal with narrative temporal coherence which underlies Theory of Mind (ToM). The model needs to reconstruct chronological dependencies including event order and causal chains together with evolving character perceptions from prompts which present unresolved inconsistencies.

We applied temporal perturbations to the stories present within the OpenToM dataset using patterns and searches to identify the relevant information given the parameters of OpenToM prompts. These perturbations fell into three categories:

- Event Reordering: Researchers modified the event order of the plot by placing "Amir left the room" before "Diego entered.".
- Preference Swapping: The system generates

incongruence between stated opinions and surrounding circumstantial details.

- Belief Editing: A change made to belief content reduced clarity regarding a belief's temporal relationship to other events.

Our methodology produced consistent datasets that incorporate matching original entries with their corresponding perturbed versions. The dataset would include OpenToM data such as narrative of what has occurred in the story– including character beliefs– plot info, and plot. We include separate sections along side the dataset for the rest attaching new distracted question and answers, logical and flipped clauses, or dialogue-formatted entries. Due to the nature of the OpenToM dataset, this allows for retrieving and working with the data functional and easy since most extractions from the situation are pre-saved.

4 Experiments

4.1 Data Collection and Processing

Dataset. We utilized a publicly available dataset in the Theory of Mind. OpenToM contains a large library of rich hand-written and machine-generated stories, comprising of 696 stories designed specifically for prompting ToM understanding. The dataset was accessed from the project's public Github. In this piece we have used 254 entries from the short context subset. The exact replication of the experiment will be uploaded to github soon. (Xu et al., 2024; He et al., 2023b)

The aim is to assess their effectiveness in reasoning over dynamic external knowledge and modeling recursive cognitive processes. SSM models are quite superficial in their current state though we use Phi-mamba, as discussed before, to evaluate claims. The model brings granular control to the SSM model's functionality while also bringing transformer level generations. We build off the large OpenToM datasets and test how SSMs perform on altered data. We also distill the OpenToM model into logical clauses and prepare them for activities regarding temporal consistency. Finally we evaluate a series of practicals on both phi-mamba and the similarly sized GPT-4o Mini.

We proceed with creating distinct additions to existing entries and updated the answer values as required. We create a dataset that includes insightful additions to modify subtle details visible in higher ToM levels by targeting the sentences with high influence of the situation. ToM levels can be thought of as: "He thinks, that I think, that she knows..." in which each theoretical belief is considered a ToM level; each level adds more uncertainty. Such distractions are tuned to change beliefs, preferences, or unstated inferences that require pulling from the input data in question rather than relying on training data which cannot match the novel context.

We report results for both Phi-Mamba and GPT-4o-mini evaluation using Chain-of-Thought Prompting on long context lengths, extra long context lengths, altered dialogues, and inserted irrelevant clauses. Each of these categories was tested on different question types in the OpenToM dataset. Factual Location Reasoning ($\text{Loc}_c(F)$) tracks the explicit and factual location of the objects in the narrative. Social Location Reasoning ($\text{Loc}_c(s)$) infers the socially relevant information of the object's location based soley on beliefs. Temporal Factual Location Reasoning ($\text{Loc}_f(F)$) tracks the factual object locations overtime as they sequentially change. Temporal Social Location Reasoning ($\text{Loc}_f(S)$) tracks the socially relevant object's locations overtime including beliefs of past events. Multihop Factual Reasoning (MHop(F)) connects multiple pieces of factual evidence to deduce conclusions. Multihop Social Reasoning (MHop(S)) performs multi-step reasoning in social contexts in order to infer beliefs. Attribution Based Reasoning (Att) interprets the attitudes or intentions of characters. F1 in the charts displays a model's ability to make correct predictions in complex scenarios. The ΔF1 demonstrates the change in the model's performance from the model's performance on OpenToM Long context lengths.

5 Results and Discussion

5.1 Interpretation

In Factual Reasoning ($\text{Loc}_c(F)$ and (MHop(F))) GPT-4o Mini unwaveringly outperforms Phi-Mamba. In Table 3 GPT-4o Mini scores 0.892 compared to Phi-Mamba's 0.571. In Table 4 GPT-4o Mini remains relatively stable (ΔF1 = -0.045) while Phi-Mamba shows a slight decrease (ΔF1 = -0.007). Even in distractors, GPT-4o mini maintains its performance with ΔF1 = 0.000.

In Social Reasoning ($\text{Loc}_c(S)$ and (MHop(S))) both models are challenged but GPT-4o mini outperforms Phi-Mamba. In table 3 GPT-4o Mini is able to achieve 0.754 on $\text{Loc}_c(S)$ while Phi-

Mamba scores 0.482. On MHop(S), GPT-4o Mini remains ahead in all scenarios but shows minor declines with altered dialogues and distractors (ΔF1 = -0.012 in Table 6). This might be due to the fact that social reasoning needs multi-level belief tracking, which GPT-4o Mini handles better because of its global attention mechanism.

In Temporal Reasoning ($\text{Loc}_f(F)$ and $\text{Loc}_f(S)$) both models are significantly challenged. In table 4 Phi-Mamba improves slightly on temporal tasks showing a positive ΔF1 for $\text{Loc}_f(F)$ (+0.077) and $\text{Loc}_f(S)$ (+0.064) while GPT-4o Mini drops significantly. Despite temporal consistency being an issue for both models, Phi-Mamba's structured latent state allows it to be better suited for tracing sequences better over time-thus its slight improvement in longer contexts.

In handling Distractors GPT-4o Mini demonstrates better robustness especially in Table 4. GPT-4o mini maintains stable performance on factual tasks and improves slightly in MHop(F) and Att tasks as well. Meanwhile Phi-Mamba sees a slight drop in most tasks such as $\text{Loc}_c(F)$ (ΔF1 = -0.044). GPT's ability to selectively pick out relevant information probably makes it better prepped to deal with distractors.

In handling the altered formatting to dialogues in table 5 GPT-4o Mini was more successful with stronger adaptability to the altered formatting with minimal performance drops compared to Phi-Mamba which demostrated higher sensitivity to the changes.

5.1.1 Overall Trends

GPT-4o Mini excels in tasks that use robust contextual reasoning as well as distractor filtering as well as when in a dialogue format. Phi-Mamba shows similar potential for long-sequence tasks, but struggles with distractors and altered formatting indicating limited flexibility. Despite Phi-Mamba's overall worse performance than GPT-4o mini, its ΔF1 across longer contexts and irrelevant clauses were on average less than Gpt-4o mini's change in performance. This indicates that state space models like Phi-Mamba experience less performance degradation on more difficult theory of mind tasks relating to longer context lengths and irrelevant clauses.

6 Conclusion

In this paper, we introduce 3 new ToM benchmarking methods derived from the OpenToM dataset: interjected irrelevant information(distractors), reformatted dialogue from narrative to first person, and temporal consistency. We also provide results from novel evaluations using GPT-4o mini and Phi-Mamba on the OpenToM dataset Long and ExtraLong context lengths, and two of the new benchmarking methods derived from the OpenToM dataset(distractors and reformatted dialogue). In these evaluations we find that although Phi-Mamba had overall worse performance than GPT-4o mini, in longer context prompts and prompts containing distractors there was a significant gap in performance degradation between Phi-Mamba and GPT-4o mini in which Phi-Mamba experiences less degredation than GPT-4o mini. This gap suggests that SSMs could have an inherently greater ability to reason more difficult theory of mind tasks, even if it is currently lagging behind, as there still remains a large gap in research between transformer based models and state-space models.

7 Limitations

While our controlled experimental setup ensures consistency, it also introduces several constraints. The OpenToM dataset, while useful, does not fully capture the complexity of real-world Theory of Mind (ToM) reasoning, limiting external validity. There is also a risk of overfitting, as certain architectures may perform better due to dataset biases.

Additionally, our evaluation methodology, which relies on structured question-answering and consistency tests, does not fully account for implicit belief inference or multi-agent interactions, both crucial for real-world ToM. Furthermore, comparing GPT-4o Mini (a fixed pre-trained model) to Phi-Mamba, which requires task-specific tuning, is not entirely equivalent.

Lastly, since the OpenToM dataset was generated using GPT-4o, we acknowledge the potential risks of LLM-generated biases or unsuitable content, though we did not observe any harmful material. Future research should carefully consider dataset construction methods to minimize biases introduced by human annotation and automated generation.

Acknowledgments

The authors extend their sincere gratitude to Kevin Zhu, Aslihan Akalin, Sean O'Brien, and Vasu Sharma for insightful lectures on ML and research skills, high-level guidance, and for offering respon-

sive feedback during the manuscript's review process.

References

J. Ainslie, Y. Jernite, A. Roberts, S. Narang, J. Lee-Thorp, and Y. Tay. 2023. Infinite context: Scaling transformers to 1m token context windows and beyond. *arXiv preprint arXiv:2306.15595*.

Ian A. Apperly. 2010. *Mindreading: An integrated account of pretense, self-awareness, and understanding other minds*. Psychology Press.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10.

Aviv Bick, Kevin Y. Li, Eric P. Xing, J. Zico Kolter, and Albert Gu. 2024. Transformers to ssms: Distilling quadratic knowledge to subquadratic models. *arXiv preprint arXiv:2408.10189*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Jonas Gehring, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yann LeCun, and Sergey Tulyakov. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Igor Bulatov, Shivanshu Singh, Yi Tay, and Donald Metzler. 2023. Scaling laws and interpretability of sparse mixture of experts. *arXiv preprint arXiv:2302.06634*.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. 2023. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*.

Ishita Dasgupta, Felix Hill, Robert Hawkins, Brenden M. Lake, and Matthew Botvinick. 2022. Language models show cognitive biases. *Advances in Neural Information Processing Systems (NeurIPS)*.

Emily Dinan, Alexander H. Miller, Kurt Shuster, Jason Weston, and Douwe Kiela. 2020. Multi-agent emotion understanding in dialogues. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mor Geva, Tal Schuster, Hila Gonen, Luke Zettlemoyer, and Omer Levy. 2023. Disentangling reasoning capabilities from knowledge in language models. *Transactions of the Association for Computational Linguistics*, 11:361–378.

Apoorv Goel, Albert Gu, Karan Goel, and Christopher Ré. 2022. Ssms for physiological time-series data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Albert Gu, Isys Johnson, Karan Goel, and Christopher Ré. 2021. Efficient sequence modeling with structured state spaces. *Advances in Neural Information Processing Systems (NeurIPS)*.

Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2023. Scaling expert language models with unsupervised domain discovery. *arXiv preprint arXiv:2303.14177*.

Xinyuan He, Ming Zhu, Hong He, and Zhou Yu. 2023a. Hallucination in large language models: A survey. *arXiv preprint arXiv:2311.01668*.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023b. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *Preprint*, arXiv:2310.16755.

Victor Iapăscurtă, Sergey Kronin, and Ion Fiodorov. 2024. Retrieval-augmented generation using domain-specific text: A pilot study. *JOURNAL OF ENGINEERING SCIENCE*.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2022. On the computational complexity of self-attention. *arXiv preprint arXiv:2209.04881*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences (PNAS)*, 121(45):e2405460121.

Sreejan Kumar, Ishita Dasgupta, Nathaniel D. Daw, Jonathan D. Cohen, and Thomas L. Griffiths. 2023. Disentangling abstraction from statistical pattern matching in human and machine learning. *PLoS Computational Biology*, 19(8):e1011316.

Bill Yuchen Lin, Zexuan Zhong, Danqi Chen, Seyeon Lee, and Xiang Ren. 2021. Adversarial contexts for robust language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Pengfei Liu, Weizhe Yuan, and Graham Neubig. 2021. Pre-training size and downstream task generalization. *Advances in Neural Information Processing Systems (NeurIPS)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Belief editing in neural language models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. 2023. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*.

Badri Narayana Patro and Vijay Srinivas Agneeswaran. 2024. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *Preprint*, arXiv:2404.16112.

Hongyu Peng, Yi Tay, and Albert Gu. 2023. Scaling structured state space models. *arXiv preprint arXiv:2304.02101*.

Joan Peskin. 1992. Rethinking the role of executive function in the development of theory of mind. *Psychological Review*, 99(3):543–576.

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

Ofir Press, Noah A. Smith, and Omer Levy. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2208.07630*.

Neil C. Rabinowitz, Feryal Perbet, H. Francis Song, Chiyuan Zhang, S. M. Eslami, and Matthew Botvinick. 2018a. Machine theory of mind. *Proceedings of the National Academy of Sciences*, 115(37):9888–9893.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018b. Machine theory of mind. *Preprint*, arXiv:1802.07740.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Sakib Shahriar, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *arXiv preprint arXiv:2407.09519*.

Kurt Shuster, Y-Lan Boureau, and Jason Weston. 2022. Multi-agent dialogue modeling with belief tracking. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 547–561. MIT Press.

Jie Sun, Tian Gao, Xiaojun Wang, and Jianfeng Liu. 2023. Generative models for conversational theory of mind reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1234–1245. Association for Computational Linguistics.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Jai Gupta, Dara Bahri, and Donald Metzler. 2022. Transformers are sample-efficient world models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. 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. *Preprint*, arXiv:2402.06044.

Manzil Zaheer, Guru Guruganesh, Kumar A. Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Defending against neural fake news. *Advances in Neural Information Processing Systems (NeurIPS)*.

H. Zhang et al. 2021. Challenges in dialogue state tracking: Handling distracting contexts and belief state errors. *arXiv preprint arXiv:2103.00109*.

H. Zhang et al. 2024. Improving temporal consistency and belief tracking in conversational ai. *arXiv preprint arXiv:2401.10353*.

Hao Zhou, Chunyuan Li, Mingyuan Zhou, Weinan Zhang, and Yong Yu. 2021. Commonsense-aware dialogue systems. In *Association for Computational Linguistics (ACL)*.

8 Appendix

Table 3: Macro F1 score of LLMs GPT-4o mini and Phi-Mamba evaluated on the OpenToM Long split using CoT prompting.

Question	GPT-4o-Mini	Phi-Mamba
$Loc_c(F)$	0.892	0.571
$Loc_c(S)$	0.754	0.482
$Loc_f(F)$	0.438	0.301
$Loc_f(S)$	0.150	0.179
MHop(F)	0.767	0.632
MHop(S)	0.699	0.575
Att	0.558	0.285

Table 4: Macro F1 score of LLMs GPT-4o mini and Phi-Mamba evaluated on the OpenTom ExtraLong split. The relevant performances compared to the OpenToM Long split are shown as relative increase, relative decrease, or approximately equal (ΔF1 < 0.010).

Question	GPT-4o-Mini		Phi-Mamba	
	F1	ΔF1	F1	ΔF1
$Loc_c(F)$	0.501	-0.391	0.311	-0.260
$Loc_c(S)$	0.358	-0.396	0.402	-0.080
$Loc_f(F)$	0.505	+0.067	0.378	+0.077
$Loc_f(S)$	0.252	+0.102	0.243	+0.064
MHop(F)	0.429	-0.338	0.406	-0.226
MHop(S)	0.353	-0.346	0.414	-0.161
Att	0.423	-0.135	0.284	-0.001

Table 5: Macro F1 score of LLMs GPT-4o mini and Phi-Mamba evaluated on the OpenTom Long split dataset with altered dialogues. The relevant performances are shown as relative increase, relative decrease, or approximately equal (ΔF1 < 0.010).

Question	GPT-4o-Mini		Phi-Mamba	
	F1	ΔF1	F1	ΔF1
$Loc_c(F)$	0.847	-0.045	0.564	-0.007
$Loc_c(S)$	0.704	-0.050	0.447	-0.035
$Loc_f(F)$	0.412	-0.026	0.293	-0.008
$Loc_f(S)$	0.136	-0.014	0.160	-0.019
MHop(F)	0.776	+0.009	0.626	-0.006
MHop(S)	0.687	-0.012	0.565	-0.010
Att	0.578	+0.020	0.292	+0.007

Table 6: Macro F1 score of LLMs GPT-4o mini and Phi-Mamba evaluated on the OpenTom Long split dataset with inserted irrelevant clauses. The relevant performances are shown as relative increase, relative decrease, or approximately equal (ΔF1 < 0.010).

Question	GPT-4o-Mini		Phi-Mamba	
	F1	ΔF1	F1	ΔF1
$Loc_c(F)$	0.892	0.000	0.527	-0.044
$Loc_c(S)$	0.628	-0.126	0.423	-0.059
$Loc_f(F)$	0.359	-0.079	0.247	-0.054
$Loc_f(S)$	0.143	-0.007	0.147	-0.032
MHop(F)	0.649	-0.118	0.581	-0.051
MHop(S)	0.583	-0.116	0.516	-0.059
Att	0.464	-0.094	0.252	-0.033