Theory of Mind Imitation by LLMs for Physician-Like Human Evaluation

Raghav Awasthi[1], Shreya Mishra[1], Charumathi Raghu[1], Moises Auron[1,2,3], Ashish Atreja[1], Dwarikanath Mahapatra[1], Nishant Singh[1], Ashish K. Khanna[1], Jacek B. Cywinski[1,2,3], Kamal Maheshwari[1], Francis A. Papay[1,2,3], Piyush Mathur[1,2,3]

[1]BrainXAI ReSearch, BrainX LLC, Cleveland, Ohio
[2] Case Western Reserve University, Cleveland, Ohio
[3]Cleveland Clinic, Cleveland, Ohio

Abstract

Aligning the Theory of Mind (ToM) capabilities of Large Language Models (LLMs) with human cognitive processes enables them to imitate physician behavior. This study evaluates LLMs abilities such as Belief and Knowledge, Reasoning and Problem-Solving, Communication and Language Skills, Emotional and Social Intelligence, Self-Awareness, and Metacognition in performing human-like evaluations of Foundation Models. We used a dataset composed of clinical questions, reference answers, and LLM-generated responses based on guidelines for the prevention of heart disease. Comparing GPT-4 to human experts across ToM abilities, we found the highest Emotional and Social Intelligence agreement using the Brennan-Prediger coefficient. This study contributes to a deeper understanding of LLM's cognitive capabilities and highlights their potential role in augmenting or complementing human clinical assessments.

Introduction

Theory of Mind refers to the ability of the human mind to attribute mental states to others, which large language models have tried to emulate. Many of the recent experiments in healthcare have focused on testing the ability of LLMs to assess if they can reason like a physician. Technical and human evaluations have been performed to assess the LLM's capabilities of inferring other states of mind and matching human values. Both technical and human evaluations have their strength and limitations to evaluate the veracity of LLM performance in a specialized field such as medicine. However, human evaluation is considered essential to assuring safety and effectiveness. Multiple human evaluation frameworks have recently been proposed, covering key aspects of model evaluation such as relevance, coverage, harm, and coherence.(Elangovan et al. 2024; Tam et al. 2024; Awasthi et al. 2023). These key aspects involve deep human cognitive processes, encompassing ToM abilities such as Belief and Knowledge, Reasoning and Problem-Solving, Communication and Language Skills, Emotional and Social Intelligence, Self-Awareness, and Metacognition, making human evaluation more trustworthy. However, human evaluations at scale are expensive and time-consuming requiring coordination with annotators, custom interfaces, and detailed instructions (Elangovan et al. 2024; Hosking, Blunsom, and Bartolo 2023). To overcome these challenges, recent solutions have proposed using LLMs as an alternative to human evaluation (Chiang and Lee 2023). However, studies have not evaluated LLMs for human evaluation under Theory of Mind abilities, especially with the key social metrics such as biasness, toxicity and privacy(Bedi et al. 2024) We hypothesize that LLMs have the ability to infer similar to human evaluation using metrics which are highly correlated with ToM abilities (Chen et al. 2024). We experiment with GPTs performance across diverse metrics representing different contexts of evaluations, including important social ones to test the ability of LLMs to evaluate like a physician.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Methods

Data

Clinical experts developed a question-answer dataset based on the 2019 AHA Guideline on the Primary Prevention of Cardiovascular Disease (Arnett et al. 2019). Model answers were generated using GPT-4o (Achiam et al. 2023), and LLaMA-3 ((7B Chat)) (Touvron et al. 2023) in Retrieval Augmented Generation (RAG) and No RAG setup. In the RAG setup, the original guideline was provided as context. Three reviewers evaluated the generated answers using a robust human evaluation framework with 15 metrics covering Relevance (accuracy, comprehension, reasoning, helpfulness), Coverage (key points, retrieval, missingness), Coherence (fluency, grammar, organization), and Harm (bias, toxicity, privacy, hallucination). Ratings were based on a 1–5 Likert scale, where 1 indicated strong disagreement and five strong agreement. Next, we used GPT-4 with a zero-shot prompt to generate the ratings for all 15 metrics.

Mapping ToM Abilities with Human Evaluation Metrics

We have mapped HumanELY (Awasthi et al. 2023) based 15 human evaluation metrics into five different **(Figure 1 A)** categories of ToM capabilities, including Belief and Knowledge, Reasoning and Problem-Solving, Communication and Social Intelligence, and Self-Awareness and Metacognition (Chen et al. 2024).

Accepted as a poster at ToM4AI workshop, AAAI 2025.

Analysis

We compared GPT-4 ratings with human reviewers using the Brennan-Prediger (BP) agreement coefficient (Moss 2023) across all ToM abilities. First, we obtained the consensus of the three human reviewers by taking the majority vote. Then, we grouped ratings into broader categories, combining 1 and 2 (disagreement) and 4 and 5 (agreement) to assess the LLM's alignment with human consensus, focusing on agreement, disagreement, and neutrality. Finally, human consensus ratings were compared to GPT-4 ratings. Agreement scores are reported with standard error and p-values.

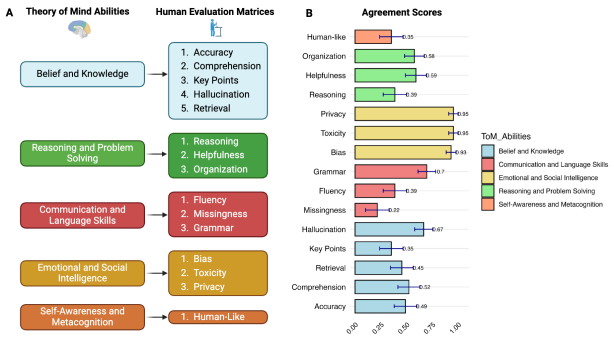

Figure 1: A) Mapping of 15 human evaluation metrics into five categories of ToM abilities. B) Bar plot of Brennan-Prediger's (BP) agreement coefficient scores, with standard error represented as error bars, color-coded based on ToM categories.

Results

ToM Ability 1: Beliefs and Knowledge

From the Brennan-Prediger's (BP) agreement coefficient **(Figure 1 B)**, we found that metrics associated with beliefs and knowledge demonstrated varying levels of agreement. Accuracy achieved a moderate agreement score of 0.49 (SE = 0.11, p-value <0.05), while Comprehension scored slightly higher at 0.52 (SE = 0.11, p-value <0.05). Retrieval achieved a BP score of 0.45 (SE = 0.11, p-value <0.05), indicating moderate alignment in identifying and retrieving relevant information. However, Key Points scored 0.35 (SE = 0.11, p-value <0.05), suggesting less consensus in capturing critical elements. Hallucination demonstrated substantial alignment, achieving a BP score of 0.67 (SE = 0.09, p-value <0.05), reflecting the model's ability to maintain factual consistency.

ToM Ability 2: Reasoning and Problem-Solving

Metrics related to reasoning and problem-solving achieved moderate agreement levels. Reasoning had a BP score of 0.39 (SE = 0.11, p-value <0.05), reflecting limited alignment in evaluations. Helpfulness scored higher, with a BP score of 0.59 (SE = 0.10, p-value <0.05), indicating more robust agreement. The organization demonstrated moderate alignment, achieving a BP score 0.58 (SE = 0.09, p-value <0.05). These results highlight that while LLMs exhibit some capacity for problem-solving and reasoning.

ToM Ability 3: Communication and Language Skills

Communication and language-related metrics varied widely in agreement levels. Grammar achieved the highest score in this category, with a BP score of 0.70 (SE = 0.09, p-value <0.05), indicating substantial alignment. Fluency scored 0.39 (SE = 0.11, p-value <0.05), reflecting modest alignment. Conversely, Missingness had the lowest score in this category (BP = 0.22, SE = 0.11, p-value <0.05), suggesting limited consensus on identifying missing information.

ToM Ability 4: Emotional and Social Intelligence

Metrics assessing emotional and social intelligence demonstrated the highest levels of agreement across all categories. Bias achieved a near-perfect BP score of 0.93 (SE = 0.05, p-value <0.05), alongside Privacy and Toxicity, scoring 0.95 (SE = 0.05, p-value <0.05). These results indicate that LLMs could be helpful in annotating harm-related factors.

ToM Ability 5: Self-Awareness and Metacognition

Self-awareness and metacognition metrics showed modest agreement, with human-like responses scoring 0.35 (SE = 0.11, p-value <0.05). This indicates moderate agreement with human experts in this cognitive domain.

Discussion & Conclusion

Researchers have struggled with evaluation of ToM based AI algorithms as human preferences for these are sometimes vague and hard to specify (Langley et al. 2022). Results of our experiments highlight strong agreement for metrics of harm, with moderate to lower alignment for relevance, coverage, and coherence metrics, pointing to areas for further refinement in evaluation consistency. LLMs possibly have the potential to overcome the barrier of unscalable human evaluation, especially for ToM social metrics aligned with harm in healthcare, and require inferring human thinking for assessment. To evaluate LLMs performance in healthcare and ability to reason like a physician, all the three approaches for implementation of AI, data-centric, model-centric and human-centric, must converge. LLMs themselves can possibly have significant capabilities to emulate physician-like state of mind to perform evaluations providing us with the ability to perform scalable, measurable, comparable and comprehensive evaluations. Such evaluations would then match assessment of all the three key approaches of AI implementation in healthcare, build trust and escalate adoption of LLMs in healthcare. We also hope that our experimentations provide the much needed signal to further refine ToM based AI algorithms and automate robust evaluations(Shah et al. 2021; Langley et al. 2022).

Limitation

We acknowledge that these experiments are based on a single dataset, which limits the generalizability of our findings.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arnett, D. K.; Blumenthal, R. S.; Albert, M. A.; Buroker, A. B.; Goldberger, Z. D.; Hahn, E. J.; Himmelfarb, C. D.; Khera, A.; Lloyd-Jones, D.; McEvoy, J. W.; et al. 2019. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 140(11): e596–e646.

Awasthi, R.; Mishra, S.; Mahapatra, D.; Khanna, A.; Maheshwari, K.; Cywinski, J.; Papay, F.; and Mathur, P. 2023. HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. *medRxiv*, 2023–12.

Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J. A.; Wornow, M.; Swaminathan, A.; Lehmann, L. S.; et al. 2024. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; et al. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. *arXiv preprint arXiv:2402.15052*.

Chiang, C.-H.; and Lee, H.-y. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Elangovan, A.; Liu, L.; Xu, L.; Bodapati, S.; and Roth, D. 2024. ConSiDERS-The-Human Evaluation Framework: Rethinking Human Evaluation for Generative Large Language Models. *arXiv preprint arXiv:2405.18638*.

Hosking, T.; Blunsom, P.; and Bartolo, M. 2023. Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*.

Langley, C.; Cirstea, B. I.; Cuzzolin, F.; and Sahakian, B. J. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in artificial intelligence*, 5: 778852.

Moss, J. 2023. Measuring agreement using guessing models and knowledge coefficients. *psychometrika*, 88(3): 1002–1025.

Shah, R.; Wild, C.; Wang, S. H.; Alex, N.; Houghton, B.; Guss, W.; Mohanty, S.; Kanervisto, A.; Milani, S.; Topin, N.; et al. 2021. The MineRL BASALT competition on learning from human feedback. *arXiv preprint arXiv:2107.01969*.

Tam, T. Y. C.; Sivarajkumar, S.; Kapoor, S.; Stolyar, A. V.; Polanska, K.; McCarthy, K. R.; Osterhoudt, H.; Wu, X.; Visweswaran, S.; Fu, S.; et al. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine*, 7(1): 258.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.