Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests

Max van Duijn[1]*, Bram van Dijk[2]*, Tom Kouwenhoven[1]*,
Werner de Valk[1], Marco Spruit[1,2], and Peter van der Putten[1]

[1] LIACS, Leiden University, The Netherlands
[2] Leiden University Medical Center, The Netherlands

Abstract. How do large language models perform on Theory of Mind tests typically used to evaluate human reasoning about intentions and beliefs of others? In this abstract we present findings from our work published in [8] to contribute to this evolving debate. We compare the performance of 11 base and instruction-tuned LLMs against 73 children aged 7-10 on standardized ToM tests, in more depth than earlier work. Our results show that in non-trivial set-ups the majority of LLMs operate below child level, with the exception of heavily instruction-tuned LLMs.

1 Introduction

Large language models (LLMs) are complex systems; even if all architecture, data, and fine-tuning details are known (which is currently not the case for most competitive models), it is difficult to assess their capabilities and performance on a variety of tasks. Researchers from fields such as linguistics [17], psychology [3, 15, 18], mental health [10, 14] and other fields [5, 16] have therefore started to study LLMs as new, 'alien' entities, with their own intelligence, that needs to be probed with experiments, an endeavor recently described as 'machine psychology' [11]. This not only advanced our understanding about LLM capabilities but also provides a unique opportunity to shed light on questions surrounding human intelligence [4, 6, 7].

In [8] we focus on determining to what degree LLMs demonstrate a capacity for Theory of Mind (ToM), defined as the ability to work with beliefs, intentions, desires, and other mental states, to anticipate and explain behavior in social settings [1]. We first address the question of how LLMs perform on standardized, language-based tasks used to assess ToM capabilities in humans. We extend existing work in this area (see [8]), in four ways: (i) by testing 11 models for a broader suite of capabilities relevant to ToM beyond just the dominant false-belief paradigm, including non-literal language understanding and recursive intentionality (A *wants* B to *believe* that C *intends*...); (ii) by using newly written

*Equal contribution. Corresponding author m.j.van.duijn@liacs.leidenuniv.nl

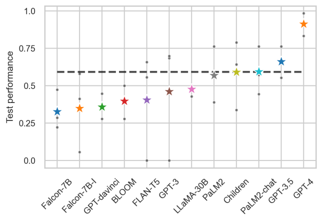

Fig. 1: Grand mean performance (stars) of all mean test scores (dots) for children (SA & SS younger children, IM older children) and LLMs.

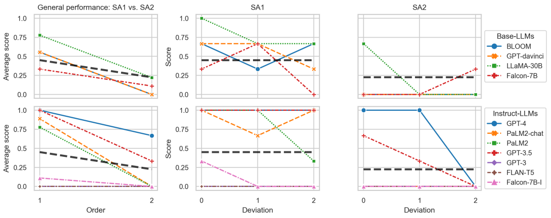

Fig. 2: Performance on Sally-Anne tests for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts performance on first- and second-order ToM (i.e. SA1 vs. SA2). Middle and right columns depict performance for SA1 and SA2 over levels of deviation from the original test. Dashed lines indicate child performance (n=37, age 7-8 years).

versions of standardized tests with varying degrees of deviation from the originals; (iii) by including open questions besides closed ones; and (iv) by benchmarking LLM performance against that of children aged 7-8 (n=37) and 9-10 (n=36) on the same tasks.

2 Methodology

Here we list our tasks for testing LLMs and children at a high level, detailed descriptions are in our paper [8]; all code, materials, and data are on OSF: `https://shorturl.at/FQR34`. In contrast to some related work, we use a broad set of ToM tests: the **Sally-Anne test, first-order (SA1)** [19], a classic false belief test, and the **Sally-Anne test, second-order (SA2)**, that targets the belief of another person about another person; the **Strange Stories test (SS)** [12], which covers scenarios of increasing complexity: a lie, pretend-play scenario, practical joke, white lie, misunderstanding, sarcasm, and double bluff; and the **Imposing Memory test (IM)** [13], which we adapted for children aged 7-10 from an unpublished version by Anneke Haddad and Robin Dunbar [9].

3 Results

Figure 1 shows that across all tests, averages for GPT-3.5 and GPT-4 exceed performance of children. However, when zooming in on specific test results, a more nuanced picture emerges (e.g., figure 2) showing that most LLMs struggle with more complex settings (more results in [8]). We are now extending this work towards multi-modal ToM scenarios, where stories are combined with images, with interesting results in terms of accuracy and confidence, and some surprisingly strong results on heavily obfuscated stories, underscoring the need for further critical evaluation of ToM tests themselves in the context of LLMs [2].

References

1. Apperly, I.: Mindreaders: the Cognitive Basis of "Theory of Mind". Psychology Press (2010)
2. van Berkel, R.: Large multimodal models and theory of mind. Bachelor's Thesis, LIACS, Leiden University, the Netherlands (2024)
3. Binz, M., Schulz, E.: Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences **120**(6), e2218523120 (2023)
4. Binz, M., Schulz, E.: Turning large language models into cognitive models. In: The Twelfth International Conference on Learning Representations (2024)
5. Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., Cao, Y., Hu, M., Lai, Y., Xiong, Z., Huang, M.: ToMBench: Benchmarking Theory of Mind in large language models (2024)
6. van Dijk, B., Kouwenhoven, T., Spruit, M., van Duijn, M.J.: Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12641–12654. Association for Computational Linguistics, Singapore (Dec 2023)
7. Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can AI language models replace human participants? Trends in Cognitive Sciences **27**(7), 597–600 (2023)
8. van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., van der Putten, P.: Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In: Jiang, J., Reitter, D., Deng, S. (eds.) CoNLL. pp. 389–402. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.conll-1.25
9. van Duijn, M.J.: The lazy mindreader: a humanities perspective on mindreading and multiple-order intentionality. Ph.D. thesis, Leiden University (2016)
10. Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., Hadar-Shoval, D.: Can large language models "read your mind in your eyes"? JMIR Mental Health **10** (2023)
11. Hagendorff, T.: Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988 (2023)
12. Happé, F.G.: 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 (1994)
13. Kinderman, P., Dunbar, R., Bentall, R.P.: Theory-of-mind deficits and causal attributions. British Journal of Psychology **89**, 191–204 (1998)
14. Kjell, O., Kjell, K., Schwartz, H.A.: AI-based large language models are ready to transform psychological health assessment (2023)
15. Kosinski, M.: Theory of Mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083 (2023)
16. Ma, X., Gao, L., Xu, Q.: ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In: Jiang, J., Reitter, D., Deng, S. (eds.) Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). pp. 15–26. Association for Computational Linguistics, Singapore (Dec 2023)
17. Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., Levy, O.: Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences **117**(48), 30046–30054 (2020)

18. Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M.S., Becchio, C.: Testing theory of mind in large language models and humans. Nature Human Behaviour (2024)
19. Wimmer, H., Perner, J.: Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1), 103–128 (1983)