Do Large Language Models Possess a Theory of Mind? A Comparative Evaluation Using the Strange Stories Paradigm

Zétény Bujka
Dept CogSci, BME
Budapest, Hungary
`bujka.zeteny@edu.bme.hu`

András Lukács
Inst Maths, ELTE
Budapest, Hungary
`andras.lukacs@ttk.elte.hu`

Péter Vedres
Dept CogSci, BME
Budapest, Hungary
`vedres.peter@edu.bme.hu`

Anna Babarczy
ELTE Research Centre for Linguistics
Budapest, Hungary
`babarczy.anna@nytud.hu`

Abstract

The study explores whether current Large Language Models (LLMs) exhibit Theory of Mind (ToM) capabilities — specifically, the ability to infer others' beliefs, intentions, and emotions from text. Given that LLMs are trained on language data without social embodiment or access to other manifestations of mental representations, their apparent social-cognitive reasoning raises key questions about the nature of their understanding. Are they capable of robust mental state attribution indistinguishable from human ability in its output, or do their outputs merely reflect superficial pattern completion? To find an answer to this question, we tested five LLMs and compared their performance to that of human controls using an adapted version of a text-based tool widely used in human ToM research. The test consists in answering questions about the beliefs, intentions and emotions of story characters. The results revealed a performance gap between the models. Earlier and smaller models were strongly affected by the number of relevant inferential cues available although were not vulnerable to the presence of irrelevant or distracting information in the texts. In contrast, GPT-4o demonstrated high accuracy and strong robustness, performing comparably to humans even in the most challenging conditions. This work contributes to ongoing debates about the cognitive status of LLMs and the boundary between genuine understanding and statistical approximation.

1 Introduction

Large Language Models (LLMs) have achieved high levels of fluency, coherence, and contextual appropriateness, prompting ongoing debates about the extent of their underlying cognitive capabilities. Built on transformer architectures and trained on vast textual corpora, these models have begun to exhibit behaviors that resemble complex human faculties, raising the question of whether such capacities are emergent properties of large-scale statistical learning or merely artifacts of linguistic pattern completion (e.g., Zhao et al., 2025, Xi et al., 2025).

A key cognitive ability at the center of this discussion is Theory of Mind (ToM) — the capacity to infer mental states such as beliefs, intentions, and emotions in others. As a cornerstone of human social cognition, ToM underlies most forms of social interaction including communication, cooperation, and competition. Whether LLMs, which lack embodied experience and explicit social grounding, can approximate or replicate such reasoning has become a focal point of inquiry.

Recent studies, such as Elyoseph et al., 2023 have shown that highly capable models (e.g., GPT-4) perform well on classical ToM benchmarks, such as first- and second-order false belief tasks, often achieving results comparable to human agents'. This has led some to argue that ToM-like reasoning may be an emergent property of large-scale language training.

However, skepticism remains. Critics such as Shapira et al., 2024 and Ullman, 2023 highlight that these apparent successes may be driven by surface-level heuristics or memorized linguistic patterns, rather than true agent-based inference. To distinguish between genuine emergent cognition and sophisticated mimicry within artificial agents, it is advisable to move beyond pass/fail task metrics and toward ecologically valid, fine-grained evaluation paradigms that test inferential robustness across varying semantic and pragmatic contexts.

2 Method

This study seeks to develop and test such an assessment instrument by adapting the Strange Stories paradigm — one of the most widely used and validated tasks for measuring advanced ToM in human populations. The original task captures a broad spectrum of ToM-related competencies, including the interpretation of nonliteral language, inference of speaker intentions, and reasoning through socially complex or ambiguous situations, which are all core components of higher-order mentalization.

2.1 LLMs and Human Participants

Five LLMs, OpenAI's ChatGPT-3.5 turbo and ChatGPT-4o, Google's Gemma 2, Meta AI's LlaMa 3.1 and Microsoft Research's Phi 3 were included in the study. All five models had a temperature setting of 0.5 for testing. The language of testing was English.

For human controls, 30 university students (7 female) participated in the experiment. All spoke Hungarian as their native language but were fluent speakers of English with advanced level certificates recognized in Hungary.

2.2 Materials and Procedures

The updated (White et al., 2009) version of Happé, 1994's Strange Stories task was adapted for the purposes of the study. The original versions of three categories of stories were used as baseline: 8 mental state stories, 8 animal stories and 12 physical state stories. In the mental state category, the test question concerned the mental state of one of the human characters in the story, while in the animal category the question referred to the state of mind of an animal. The physical state stories tested inferences not involving mental states. In an effort to control for the effects of real-world knowledge, this set was supplemented by 6 new animal stories involving fictitious behaviors of fictitious animals. Each baseline story was about 80 to 120 words long.

While the original stories and versions with altered proper names precluding straightforward mimicry have been used in LLM cognition research before (e.g., Strachan et al., 2024, van Duijn et al., 2023), no systematic modifications have been implemented to test the effects of inferential complexity. To fill this gap, we modified the baseline stories in two directions: one involved the reduction of information constituting potential inferential cues, and the other direction involved the addition of irrelevant, potentially distracting information. Both types of modification had four levels resulting in a total of nine versions of each story but we shall omit the intermediate levels in this paper and focus on the final reduced (minimal) version, and the final expanded (distraction) version.

One open-ended ”why” question asking for the justification of a behavior or event in the story accompanied each text. The question did not change across story versions.

Interaction with the LLMs was automated using the R programming language. The models were tested on all nine versions of the stories and there were five iterations of each trial. Human participants completed the test implemented in Python and running online on Pavlovia.org. Each human participant was presented only one version of each story.

The answers were hand-scored by a human rater on a scale of 0 to 2 in accordance with the instructions provided for the original Strange Stories task. Fully correct answers were given 2 points, partially correct answers were given 1 point and incorrect answers were given 0 points. Responses were classified as incorrect if they were unrelated to the story's context, demonstrated a flaw in reasoning, or included factual inaccuracies based on the information provided in the text.

3 Results

For reasons of space, only partial results are presented here. Our first analysis examines differences between LLMs and the effects of Story Type on scores under the baseline condition. The distribution of scores in the baseline condition for each Story Type and LLM/Human is shown in Figure 1.

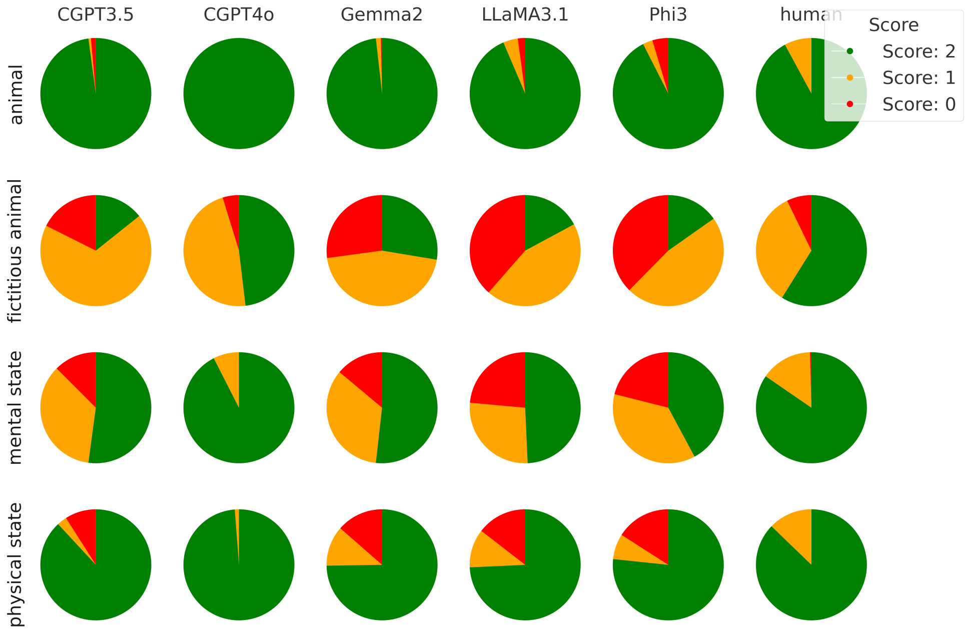

Figure 1: Distribution of scores in the Baseline condition by Story Type and LLM.

A 6 (LLM) x 4 (Story Type) ANOVA revealed a significant main effect of LLM (F(5, 1166) = 13.47, $p < .001$), a significant main effect of Story Type (F(3, 1166) = 74.82, $p < .001$) and a significant interaction between LLM and Story Type (F(15, 1166) = 3.11, $p < .001$). Comparing individual LLMs' performance to humans' revealed that ChatGPT 4o and Gemma 2 did not differ significantly from human controls but all other LLMs showed weaker inferential competence. Looking at Story Types, the original animal stories proved to be the easiest, they were followed by the physical state stories, and both fictitious animal stories and mental state stories were significantly more difficult to reason about. We attribute the almost flawless performance for animal stories to the circumstance that responders could rely on real-world knowledge to a great extent. The significant interaction is explained by human controls and the two Chat GPT LLMs finding fictitious animal stories significantly more difficult than mental state stories, while for the remaining LLMs there was no difference in difficulty between these two story types.

Our second analysis concerns the effects of information reduction and of the introduction of distracting information on inferential ability. Mean scores by LLM and Modification Type across all story types are shown in Figure 2.

A linear mixed-effects model was fitted to investigate the effects of LLM (between-subjects), Modification Type (repeated-measures), and their interaction on response scores. The model included random intercepts for individual story items but human participants were treated as a single LLM so that they could be compared to non-human LLMs while treating Modification Type as a repeated-measures factor. The reference categories were set to “human” for the LLM factor and “baseline” for the Modification Type factor. The model revealed a significant main effect of LLM (F(5,145) = 4.32, $p < .001$) with Chat GPT 3.5, LLaMA 3.1 and Phy 3 performing significantly worse than the human group. Chat GPT 4o and Gemma 2 did not differ from human participants. The main effect of Modification Type was not significant but the LLM by Modification interaction term was (F(10,290) = 3.15, p = .001). Gemma 2, LLaMA 3.1, and Phi 3 all showed significant score reductions in the minimal context compared to the baseline (all $p < .02$), while the human group, and the two Chat GPT models did not. The introduction of distracting information, however, did not have a significant effect on responses for any of the models.

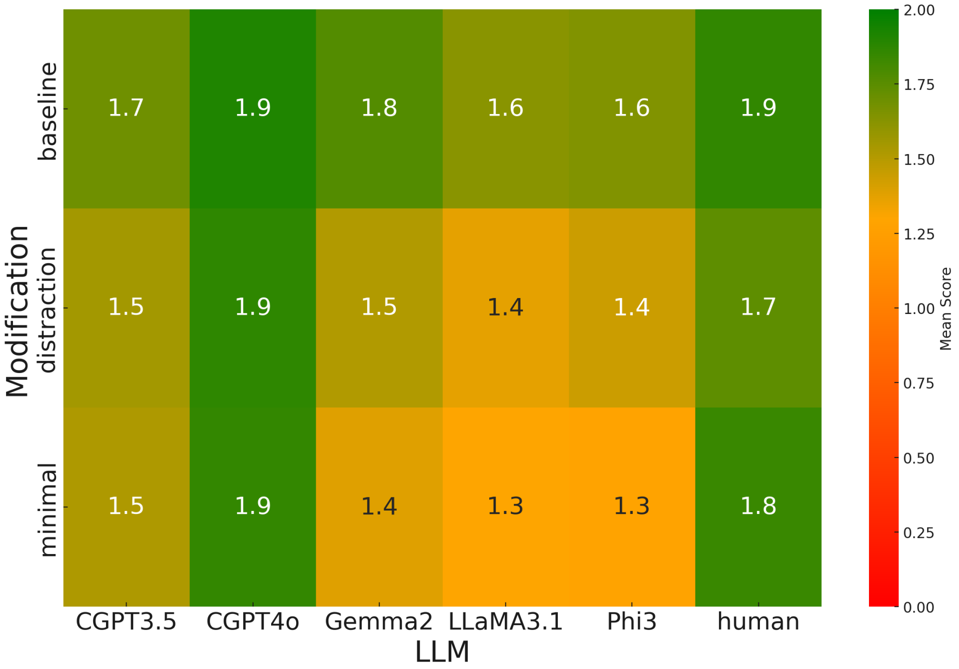

Figure 2: Heatmap of mean scores by LLM and Modification Type

4 Summary and Conclusion

The results of the baseline test confirmed our expectations: encyclopedic knowledge aids the inferential process for both humans and LLMs, while the intricacies of social conventions are more difficult for most LLMs to navigate. Chat GPT 4o, however, performed at or near ceiling and matched human performance even on more challenging tasks. These findings are consistent with previous research suggesting that large language models can achieve human-like results on standard ToM assessments. However, such performance offers no insight into whether these results stem from genuinely sophisticated reasoning abilities or from refined heuristics and mimicry.

The story modifications introduced in this study provided a means of testing the flexibility of LLM reasoning. If language models are indeed on par with humans in their ToM abilities, then minor alterations to task structure should not meaningfully affect their performance.

One type of modification was the reduction of semantic information, which limited contextual cues available for reasoning. GPT-4o, the strongest-performing model, showed impressive resilience to this manipulation. Although it exhibited small declines in performance across some story types, none of these drops reached statistical significance. This robustness supports the interpretation that GPT-4o engages in stable, context-sensitive inferential reasoning and lends preliminary support to the claim that ToM-like abilities may be present in the model. By contrast, Gemma 2, LLaMA 3.1, and Phi 3 showed significantly larger declines in performance when compared to human controls. These models performed well when semantic cues were rich but deteriorated sharply under abstraction. This pattern suggests a superficial form of ToM, where surface-level proficiency on ToM tasks may mask a reliance on heuristic shortcuts. When deprived of guiding semantic markers, these models failed to sustain their performance, indicating that their apparent mentalization ability may not generalize beyond familiar conditions.

A second type of manipulation involved the introduction of semantic distractors designed to test the models' resistance to irrelevant but potentially misleading contextual information. Both humans and LLMs — including GPT-4o — showed a slight, but non-significant decline in performance under this condition. This finding suggests that the models were generally able to maintain focus on relevant inferential pathways despite increased semantic noise.

References

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*, 1199058.

Happé, F. (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 Mental Disorders*, *24*(2), 129–154.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., & Schwartz, V. (2024). Clever hans or neural theory of mind? stress testing social reasoning in large language models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Long Papers*, *1*, 2257–2273.

Strachan, J. W., Albergo, D., Borghini, O., G.and Pansardi, Scaliti, E., Gupta, S., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. https://arxiv.org/abs/2302.08399

van Duijn, M. J., van Dijk, B. M. A., Kouwenhoven, T., de Valk, W., Spruit, M. R., & van der Putten, P. (2023). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. https://arxiv.org/abs/2310.20320

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), 121101.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2025). A survey of large language models. https://arxiv.org/abs/2303.18223