Report on "Theory of Mind May Have Spontaneously Emerged in Large Language Models"

Paper Report as Part of the "Current Topics of Computational Social Science" Seminar

Taline Kehlenbach

RWTH Aachen

Abstract. This report summarises, questions and critisises the paper "Theory of Mind May Have Spontaneously Emerged in Large Language Models" by Kosinski

Keywords: Artificial Intelligence · Theory of Mind · Language Model

1 Introduction

Theory of Mind is known in psychology as the ability to understand and reason about one's own and other individual's mental states such as differing beliefs or knowledge states [4]. Most importantly, ToM is regarded to be a uniquely human ability [2]. The research presented in the paper titled "Theory of Mind May Have Spontaneously Emerged in Large Language Models" attempts to contradict this understanding and suggests that large language models may already possess Theory of Mind while not being specifically designed to solve problems for which Theory of Mind is needed. By conducting three studies on the reasoning capabilities of large language models (LLM) such as GPT-3 and GPT-4 Kosinski attempts to prove this hypothesis and compares the results to research on Theory of Mind in developmental psychology, giving an interesting insight in the status of Artificial Intelligence compared to human intelligence.

2 Related Work

2.1 Theory of Mind

While Theory of Mind and the research related to it in a purely psychological understanding is more of a foundation for the paper in question it is still important to know what the term means and the research it is related to, to grasp the context of the paper at hand.
Theory of Mind encompasses all abilities that infer various mental states based on contextual information. The term Theory of Mind (ToM) gained traction in the late 1970s and has since found firm footing in developmental psychology where the term is used to determine the existence and emergence of ToM in children. ToM is also a key part to enable predictions of behaviour based on an

agent's mental state.[4]
The research around ToM in child development plays an important role in Kosinski's work since he uses both well-cited methods and results from this field to first construct ToM tests for LLMs and consecutively compare the LLMs' performances to that of children in different age groups. Therefore, the relevant methods and research results will be shortly summised in the following. In order to assess a child's ability to differentiate reality from beliefs, false-belief tests were introduced by several researchers [8, 1, 6]. In theses tests children were exposed to situations in which they have to make distinctions between reality and the knowledge or belief of another person about reality and predict someone's behavior according to it. Kosinskis work is based on two kinds of these tests.
First, the Smarties Task developed by Perner et al. in 1987. In this experiment children around age 4 are shown a tube of "Smarties" (chocolate candy), asked its contents and then shown the real contents which were actually random objects and not candy. They would then be asked the questions what another child who did not see the real contents of the tube would expect to be inside, thus testing the reasoning skills about false beliefs of others. [6] Second, the Sally-Anne Task first used in a study from 1983 [8] and then again in 1985 [1] (this time coining the name Sally-Anne Task) tests the same reasoning ability of attributing false beliefs to others. In this task two agents are introduced (typically using puppets). One agent places an item and leaves the scene while the other agent moves the item before the first agent comes back to look for the item. The children are then asked about the expected behaviour of the first agent who did not witness the displacement of the item. The task is solved correctly when the child predicts that the first agent will search the item in the place where they left it before the second agent moved it, thus proving the child's ability to infer false beliefs of other people. [8, 1]
As for results in developmental psychology regarding ToM, Kosinski refers to the work of Wellman et al. which agglomorates several studies on ToM (and other attributes) of children mainly aged from 2.5 to 6 years old and interpolates their performance on both Smarties and Sally-Anne Tasks [7]. Known researchers in the field of developmental psychology are Frye and Moore or Wellman et al. who suggest that ToM emerges in children around ages 3 to 4 [5, 7]. Kosinski bases comparisons of LLM performance on ToM tests to human perfor

Strengths: - interesting premise - rewriting of existing ToM tests - well structured reasoning - in-depth showcasing and visualisation of findings - relation of ToM abilities to human age

Weaknesses: - assumption of model not knowing the ToM problem by rewriting it - missing data for study 3 - location and contents tasks are "remarkably equivalent"[7, ?] - age comparison is a bit wonky regarding the figures in [7, ?] i.e. reading the values has high error also studies with suspects aging well over 5 years are few in the used reference

7 Future Work

- unique ToM tasks with relation to human abilites (e.g. by age) to give context for progression of LLMs

8 First Section

8.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

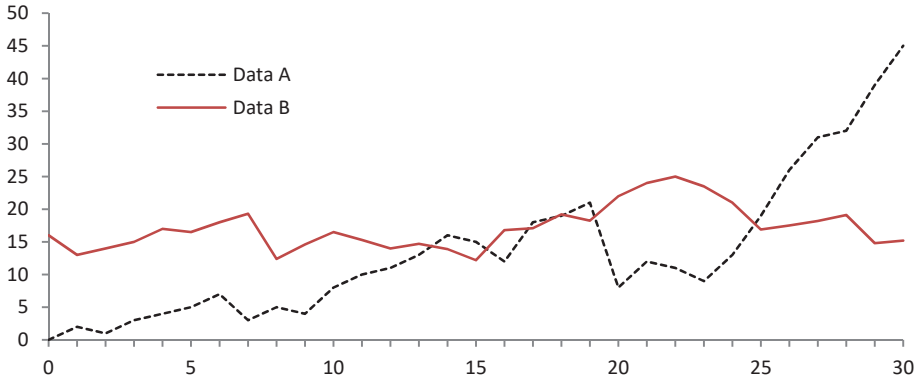

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

Acknowledgements Thanks to Professor Kosinski for answering some of my questions regarding his work and to Professor Wagner for assisting me during the seminar and writing of this paper.[3]

References

1. Baron-Cohen, S., Leslie, A.M., Frith, U.: Does the autistic child have a "theory of mind" ? Cognition **21**(1), 37–46 (1985). https://doi.org/https://doi.org/10.1016/0010-0277(85)90022-8, https://www.sciencedirect.com/science/article/pii/0010027785900228
2. Call, J., Tomasello, M.: Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences **12**(5), 187–192 (2008)
3. Doe, J.: The Book without Title. Dummy Publisher (2100)
4. Franchin, L.: Theory of Mind, pp. 1639–1644. Springer International Publishing, Cham (2022)
5. Frye, D., Moore, C.: Children's theories of mind: Mental states and social understanding. Psychology Press (2014)
6. Perner, J., Leekam, S.R., Wimmer, H.: Three-year-olds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology **5**(2), 125–137 (1987). https://doi.org/https://doi.org/10.1111/j.2044-835X.1987.tb01048.x, https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-835X.1987.tb01048.x
7. Wellman, H.M., Cross, D., Watson, J.: Meta-analysis of theory-of-mind development: The truth about false belief. Child Development **72**(3), 655–684 (2001). https://doi.org/https://doi.org/10.1111/1467-8624.00304, https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/1467-8624.00304
8. 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). https://doi.org/https://doi.org/10.1016/0010-0277(83)90004-5, https://www.sciencedirect.com/science/article/pii/0010027783900045

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890

3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017