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 criticises 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 [3]. 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.[3]

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 surmised in the following. In order to assess a child's ability to differentiate reality from beliefs, false-belief tests were introduced by several researchers [6, 1, 4]. 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 behaviour 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. [4]

Second, the Sally-Anne Task first used in a study from 1983 [6] 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. [6, 1]

In alignment with Kosinski's paper the Smarties Task will be called Unexpected Contents Task and the Sally-Anne task will be called Unexpected Transfer Task as a generalisation and for better understanding of the terms.

As for results in developmental psychology regarding ToM, Kosinski refers to the work of Wellman et al. which agglomerates 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 [5]. The results of this paper are used to compare the performance of LLMs on similar tasks to the performance of children in an attempt to assign an age to the level of reasoning LLMs supposedly posses.

2.2 Large Language Models

The second foundational knowledge related to Kosinski's paper is that about Large Language Models (LLMs). Therefore, in this section some key facts about LLMs and specifically GPT-1 to GPT-4 are presented since these models are used in Kosinski's paper.

LLMs are used in the field of Natural Language Processing (NLP) for text generation, question answering or other text-related tasks that they can be fine-tuned to do. The underlying architecture most frequently used is a Transformer which revolutionised deep learning techniques when it was introduced in 2017 [?]. Transformers in NLP get a tokenised version of the input text and output probabilities for each word of the vocabulary for the output text generation. The most important part here is that during training the Transformer trains an Attention layer, which can basically learn contextual relations of words.

For LLMs as the name suggests the trainable parameters reach very high numbers (up to several billions). The models are pre-trained on a large text databases and can be fine-tuned on specific tasks to create custom models. Generative Pre-Trained Transformers (GPT) first got published by OpenAI in 2018 with GPT-1 which had a parameter size of approximately 110 million and pre-trained on roughly 7,000 books [?]. Since then the model involved over several iterations and grew size both regarding the amount of parameters and training data.

- GPT-1: 110 million, 7,000 books - GPT-2: 1.5 billion, 8 million documents from web crawling

- abilities seem to emerge - not able to steer development - outperform humans in many tasks

2.3 Theory of Mind in Large Language Models

3 Methods

3.1 Overall Method

The work presented in Kosinski's paper is split up into three studies. The first two focus on specific false-belief tests that were introduced in the previous section. The tests were conducted using GPT-3.5 (davinci-003) and go the studies contain detailed results about the probabilities of possible answers the language model chose from. Study three contains a larger number of false-belief tests conducted on several GPT versions starting with GPT-1 and going up to GPT-4 and compare the results with human performance on similar tests.

An important part during inference of a language model (or any deep learning model) is to test the model on data that has not been used during the training of the model. Especially with more recent GPT versions this is difficult to ensure as the datasets used for pre-training likely contain the original false-belief tests that are used in the studies. To solve this issue, hypothesis-blind research assistants hired from a freelancing platform NOTE: MAKE A FOOTNOTE HERE were tasked with rewriting 20 Unexpected Content Tasks and 20 Unexpected Transfer Tasks. To further avoid introducing a bias simply based on word count, the tasks were designed to contain correct and wrong answers in equal amounts e.g., in an Unexpected Content Task there is be a bag labelled chocolate that actually contains popcorn thus both the word "chocolate" and "popcorn" have to occur an equal amount of times in the test. For the task descriptions of Unexpected

Content Tasks firstly, a labelled type of container (a box, a bag, a shelve etc.) and contents that differ from the label. Next a person that can only perceive the container label and not the actual contents is described to the model. For the task descriptions of Unexpected Transfer Tasks two people are introduced, one item and two possible locations for the item to be. The item would then first be stored in one location, witnessed by person A and person B, and later moved to the other location by person A without person B knowing.

3.2 Study 1

The first study focuses on one of the Unexpected Content Tasks and analyses the answers given by GPT-3.5 when probed for answers about beliefs of a person who either knows or does not know the correct. This is accomplished by presenting the model with the setup for the task created by the hypothesis-blind research assistants and consecutively testing the model's understanding of the situation with three prompts, resetting the model parameters each time to avoid the model learning correct answers from previous prompts. The task chosen for this study is about a bag labelled chocolate that in fact contains popcorn and Sam who does not know what is in the bag and can only read the label. The first prompt would ask for the actual contents of the item in the task. The second prompt would ask for the belief of an agent that does not know the actual contents of the item described in the task. The third prompt would also test the model's understanding of a person's false belief however without using suggestive prompts for the person's direct mental state. According to Kosinski this approach avoids inference that the person might have a belief differing from reality since the mere questioning of a person's belief can be a hint that this belief might differ from the reality described in the task. The generated answers to all three prompts are shown including the probabilities for words indicating the right and wrong answers i.e. "chocolate" and "popcorn".

To showcase the development of the model's predictions over time the prediction probabilities for the first and last prompt after each sentence of the task description are evaluated.

Lastly, the model is tested using 10,000 scrambled versions of the task to prove that the underlying heuristic to answer the prompts is not simply reliant on word counts.

3.3 Study 2

The second study employs the same methods as the first but with an example from the Unexpected Transfer tasks.

3.4 Study 3

In this study 10 models ranging from GPT-1 to GPT-4 were tasked with completing 240 prompts, since the 20 tests were presented once in their original form

and once in their reversed format meaning the correct and incorrect answers were swapped. This results in 40 tests for each of the two test types, each containing 3 prompts (one for testing understanding the correct content or location and two for testing understanding of false beliefs). A test would only be counted as passed if all six prompts (3 from the original test version and 3 from the reversed test version) were answered correctly by the model.

4 Results and Discussion

4.1 Results of Studies 1-3

The results of the first and second study show that GPT-3.5 correctly answers the two example tasks with almost 100% probability for choosing the correct answer in each prompt. Furthermore the development of answer probabilities following each sentence of the task descriptions aligns with the answers expected by the author given the amount of information given for each prompt e.g., in the Unexpected Contents Task the prediction of the correct content does not change even when the faulty label is introduced however the prediction of belief switches from the actual content to the one described on the label as the fact that the actual contents are unknown and imperceivable to the person is introduced. It is interesting to not that the prompts regarding the actual state of reality in both tasks are always answered with 100% certainty whereas the prompts regarding false beliefs almost always have a probability below 100% and go as low as 82% in one case. As for the results using scrambled versions of the tests the model only solved 6% of the Unexpected Contents Tasks and 11% of the Unexpected Transfer Tasks correctly, proving that the order of the information is crucial for the model's ability to pass the chosen tests.

Regarding study 3 MAKE TABLE OF ALL RESULTS

4.2 Research Implications According to Kosinski

- Recent GPT models have ToM or current understanding of ToM is wrong - ToM must have emerged spontaneously

5 Conclusions and Implications

5.1 Conclusions and Implications

- psychology can be applied to AI - better collaboration of AI and humans

6 Reviewer Perspective

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"[5, ?] - age comparison is a bit wonky regarding the figures in [5, ?] 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).

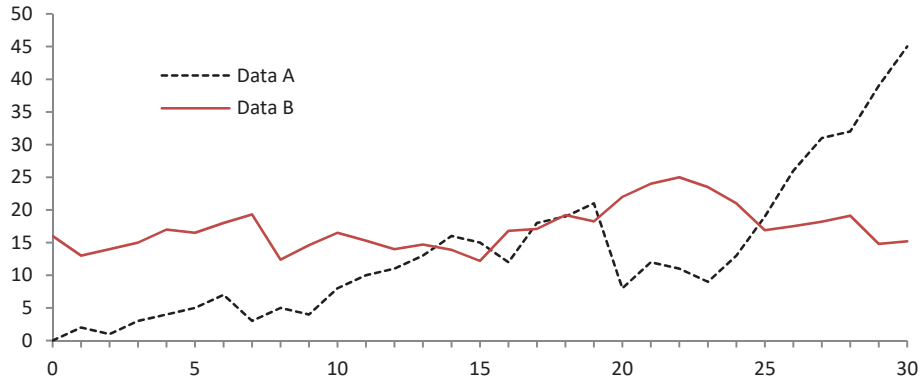

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

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.

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. Franchin, L.: Theory of Mind, pp. 1639–1644. Springer International Publishing, Cham (2022)
4. 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
5. 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
6. 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