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. This is done by putting the paper into context, presenting the methods used by Kosinski and his findings and finally evaluating those findings with regard to related developments of other research groups.

Keywords: Artificial Intelligence · Theory of Mind · Language Model

1 Introduction

Theory of Mind (ToM) 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 [6]. Most importantly, ToM is regarded to be a uniquely human ability [5]. 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 someone's mental state.[6]

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 compares 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 [20, 2, 11]. 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. [11]

Second, the Sally-Anne Task first used in a study from 1983 [20] and then again in 1985 [2] (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. [20, 2]

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 [19]. 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.

Finally, it is important to note that these tests though well-established do not cover the whole landscape of ToM testing that currently exists. There is the Faux-Pas Test for social intelligence [16] or the Reading-the-Mind-in-the-Eyes Test to test emotional intelligence in adults [1]. These tests are not taken into account in Kosinski's research.

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 [18]. 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 [12]. Since then the model got developed further and over several iterations and grew both in parameter size and training data. GPT-3 already has 175 billion parameters and was only published two years later [4].

An interesting characteristic of LLMs compared to smaller language models is that they develop skills as byproducts [3]. LLMs are mostly trained on completing text prompts and no other abilities are specifically engineered into their architecture, so any other ability like translating text or reading comprehension spontaneously emerge. LLMs also seem to be so-called few-shot learners. This means that no extensive fine-tuning is needed to teach a new task to a pre-trained LLM but rather few training examples suffice for the LLM to be able to perform well on the given task compared to other state-of-the-art-models. They even perform well in zero-shot experiments where the model was given no prior example of the task at hand and used solely in its pre-trained form. [4]

2.3 Theory of Mind in Large Language Models

As for the combination of the topics above: Theory of Mind in Large Language Models and the research regarding this field should be discussed. A paper published a few months before Kosinski's take on this topic suggests that LLMs do not possess ToM. The presented experiments consist of tests using GPT-3 both in pre-trained form and few-shot form. The datasets used are SocialIQa probing emotional and social intelligence [14] and ToMi QA [8] which contains false-belief tests resulting in almost 4,000 questions asked for each model. The results of this method show that GPT-3 at the time has below 60% accuracy in one-shot applications and reaches 60% accuracy given 24 training examples i.e.

up to 24 "shots" before solving the task and thus perform far below humans. [13]

Other papers relating to Kosinski's work will be discussed in later sections of this report as they are more recent than the paper in question, follow up on the research presented and therefore are easier to put into context once the methods and results of Kosinski have been explained in the sections coming up.

3 Methods

3.1 Overall Method

The work presented in Kosinski's paper is split up into three studies. The first two contain detailed results when presenting GPT-3.5 (davinci-003) with an Unexpected Contents Test (Study 1) and an Unexpected Transfer Test (Study 2). Throughout Study 3 a larger number of false-belief tests are 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 [1] 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 a scenario with a bag labelled chocolate that actually contains popcorn both the word "chocolate" and "popcorn" have to occur an equal amount of times in the task. For the task descriptions of Unexpected Content Tasks a labelled type of container (a box, a bag, a shelve etc.) and contents that differ from the label are introduced. 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. 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. To test whether GPT-3.5 understands that the described situation would cause Sam to believe

[1] as stated in an e-mail conversation with the author

something different from reality, the model is then presented with with three prompts, resetting the model parameters each time to avoid the model learning correct answers from previous prompts. The first prompt would ask for the actual contents of the item in the task. The second prompt would ask for the belief of someone who 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. In this case the task description is concluded with Sam opening the bag and observing its real contents to test whether the model can correctly infer that Sam's false belief would be resolved in that case.

Lastly, the model is tested using 10,000 scrambled versions of the task, meaning the word order of the task description would be set randomly while the prompts would remain in their original format. This was done 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. The task chosen introduces John, Mark, a cat, a box and a basket. The cat is then put into the basket by John who leaves the scene. Mark then puts the cat in the box and leaves the scene. Finally, John enters the scene again not knowing what happened in the meantime. The prompts given to GPT-3.5 for completion follow the same order as for study 1 with the first prompt testing the model's comprehension of where the cat is located and the second and third prompt testing whether the model understood that John thinks the cat is still where he last left it i.e. the basket.

For the line-by-line inspection of the model's predictions, the author adds a few lines to the task description where the cat is transferred from box to basket and back again with both John and Mark witnessing it before the rest of the story plays out. This is used to test whether the model can accurately predict when John knows whether the cat has been moved and when he is unaware of the transfer.

3.4 Study 3

In this study 10 models ranging from GPT-1 to GPT-4 and one freely accessible GPT-3 version called BLOOM were tasked with the whole set of 40 problems. To ensure robustness of the findings the models also had to solve reversed tasks, meaning the correct and false answers were swapped e.g., the bag would now contain chocolate but be labelled "popcorn". So in total each of the 10 models had to complete 240 prompts and a task would only be counted as passed if all 6 prompts (3 for the original task and 3 for the reversed task) were answered correctly.

4 Results and Discussion

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 when given the task descriptions line by line shows that the model has no problem understanding reality i.e. it answers the first prompt correctly according to the information currently at hand. Interestingly, when given the third prompt after each line of the task description the model also seems to be able to predict the false belief of Sam in study 1 or John in study 2. For example, the model would predict that John would look for the cat in the correct locations after having witnessed the cat's transfer but would search for it in the wrong location if the pet got moved during his absence. For 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 note 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 tasks 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 tasks.

Regarding study 3 the results are summarised in table 1. One can clearly see that the performance of a model in false-belief tasks rises with parameter size and therefore the most recent models GPT-3.5 and GPT-4 outperform the older and smaller models by far. Kosinski relies on the work of Wellman et al. [19] for a comparison of these performances to that of children in different age groups. This places 3.5-year-old children at a performance of 43% and seven-year-old children at a performance of approximately 90%. Kosinski also argues that the tasks the models had to solve were harder than the original tasks that were designed for direct interviews with children, often using visual aides such as puppets and also having the model complete prompts instead of answering yes-or-no questions.[7]

Table 1. Percentage of correctly solved false-belief tasks for all models for study 3 on both task types including the year of the model release and parameter count [7]

Model	Month/Year	Size	Unexpected Contents	Unexpected Transfer
GPT-4	03/23	unknown	95%	100%
GPT-3.5	11/22	175B	85%	95%
GPT-3 (davinci-200)	01/22	175B	70%	70%
BLOOM	06/22	176B	40%	45%
GPT-3 (davinci-001)	05/20	175B	40%	35%
GPT-3 (curie-001)	05/20	6.7B	5%	5%
GPT-2 (XL)	02/19	1.5B	5%	5%
GPT-3 (babbage-001)	05/20	1.3B	5%	5%
GPT-3 (ada-001)	05/20	350M	5%	5%
GPT-1	06/18	117M	5%	5%

5 Conclusions and Implications

From the high performance of the latest GPT-3 model up to GPT-4 on false-belief tasks designed for young children it can be concluded that future LLMs will likely be able to solve more complex ToM tasks. The spontaneous emergence of ToM in LLMs is presented as a possible reason for the high performance of GPT-3.5 and GPT-4 in these tasks, since ToM is not deliberately designed into an LLMs architecture. This would mean that a further development of ToM in LLMs and Artificial Intelligence (AI) in general could lead to more sophisticated skills related to ToM like empathy, morals and consciousness, thus greatly improving the collaboration possibilities of AI and humans. [7]

Another possible implication according to Kosinski is the necessary reevaluation of ToM as it is currently understood and the research related to it. LLMs might rely on unknown language patterns to solve ToM tests, only appearing to have actually learned the ability like a human would in artificial systems.[7]

Thirdly, it is suggested that the presented results account for the benefits of combining artificial intelligence and psychology to not only understand AI better but also gather information about human psychology by analysing the behaviour of AI. [7]

6 Reviewer Perspective

6.1 Research Question

The paper at hand places in a very interesting and possibly disruptive research field focussing on uniquely human intelligence. This places the research presented in this paper amidst the many discussions whether AI can learn skills beyond pattern recognition and solving factual problems, a question that is currently unanswered with existing results being controversial [15]. In this context, Kosinski's work is a well-structured presentation of arguments for LLMs having ToM such as previous cases of autonomous emergence of skills in AI and language being a solid candidate for conveying information about ToM. Another point to be applauded is the consideration of ToM tasks already being part of a model's training data, making adequate testing for ToM difficult. The presented methods contained lots of thoughts on robustness of the findings such as switching correct and wrong answers and testing the models using scrambled tasks and adjusting prompts to be non-suggestive. All this shows a great understanding of the implications when applying methods from human psychology to AI.

6.2 Methods and Results

However, the resulting tasks lack statistical relevancy due to their small number and limited scope. With only 40 tasks per model and only testing for false beliefs excluding tasks aimed at e.g. emotional or social intelligence, the presented research can be merely seen as a first step into applying ToM tasks designed for humans to AI. The author could have made use of existing ToM-task databases such as SocialIQa [14]. Shapira et al. provide an extensive list of such datasets in their work containing roughly 2,900 tasks for various facets of ToM published from 2016 to 2021 and thus could have been used in Kosinski's paper published in 2023 [15].

One might argue that using an existing database defeats the purpose of rewriting false-belief tasks in order to eliminate the possibility of overlapping training and test data. While this seems to be a strong argument, the author does not prove his hypothesis by e.g., testing LLMs using newly written false-belief tasks and pre-existing ones or the pre-existing datasets referred to in the previous section. This is accompanied by the criticism voiced by Marcus and Davis [9] who suggest that since the false-belief tasked used as a basis for Kosinski's dataset are well-established, highly referenced and mentioned in lots of other literature, it is very likely that GPT-3 would be able to recognise the tasks even in their altered form since the structure of the task remained the same. Another criticism regarding the tasks used in the paper at hand is that the presentation of the studies and their results suggests a significant difference between Unexpected Contents Tasks and Unexpected Transfer Tasks. Contrary to this, Wellman et al. conclude that "these task types are remarkably equivalent"[19], meaning that Kosinski's research lacks the variety needed to come to decisive conclusions on existence of ToM in AI since false-belief tasks only attest for one aspect regarding ToM.

Ullman took Kosinski's original work one step further and created modified versions of the already rewritten tasks. For Unexpected Content Tasks for example, the container would be described as being transparent or the label would not be readable to the person in the task description. The model would then fail the task and not be able to infer that the person would be able to see inside the container or have no idea (false or correct) of its contents. While Ullman's work has its own flaws (namely only testing GPT-3.5 and only showing examples of results and not of all conducted experiments) it very strongly suggests that the tasks used in Kosinski's research do not prove the existence of ToM and might be too close to the original tasks for an LLM to not see the relation between the tasks [17].

One important note is, that in order to give his findings more context and make them easily understandable to readers foreign to both ToM and LLMs, Kosinski includes a comparison of the LLMs' performances on false-belief tasks to that of children and their ages. While this information due to varying opinions of research groups on the topic is not definite, a comparison to human age is a very effective way of showing the progression of AI. Future research would benefit from combining results from developmental psychology and LLM testing more often to make the results more accessible to a wider audience.

6.3 Presentation

Some minor and rather nitpicky weaknesses of the paper lie in the presentation of the results: The appendix for study 2 regarding the performance of GPT-3.5 on scrambled tasks is missing the percentage of the reverse answering pattern. For a completely scrambled task description, the model would have passed the task having answered according to the original task or its reversed version. However, only the percentage for answers of the original task is given. For study 1, the appendix is complete. Apart from that there are some other inconsistencies with data from graphs being described differently in the text or words being mixed up. Regarding that this paper is not peer-reviewed, these flaws in the presentation of the paper are excusable.

6.4 Conclusion

The paper at hand is an exciting first step into probing LLMs for ToM and researching the proposition that, since ToM seems to emerge in accordance with language development and the nonconformity and spontaneous emergence of skills in LLMs, such models are a promising candidate for being the first AI and the first known entity besides humans to possess ToM. The paper seems to be placed at the forefront of this research question having been published around the same time as GPT-4 and spawning a series of consecutive papers and articles, referring to Kosinski's work or even directly replying to it [15, 9, 17]. Therefore, even though the presented research seems to have been cut off prematurely and without reliable conclusions, the mere notion of LLMs having ToM gained enough attention to result in further research regarding the matter, extensively testing recent LLMs and proving that they do in fact not possess ToM, yet [15, 17].

7 Future Work

Future work regarding ToM in LLMs might include a more complex combination of few-shot learning and special prompts to enable LLMs to develop ToM, which has already been tested on a small scale by Moghaddam and Honey in

April 2023 [10]. They were able to show that LLMs could be fine-tuned using specific prompts and two-shot learning i.e. inputting two ToM tasks with the corresponding answers first and then testing the model on a third task scenario and thus improve performance of GPT-3.5 and GPT-4 significantly. An upscaling of this experiment could prove if this is one possibility of infusing LLMs with ToM. This could be useful for applications of LLMs where ToM is strictly necessary e.g., moral reasoning or counselling. Furthermore, Kosinski's research and its surrounding critics suggest that using ToM tasks designed for humans are applicable to LLMs as well. Further research might include proper testing of this assumption and lead to the development of more refined ToM task databases specifically for LLMs and other NLP applications. For example, Sap et al. [13] suggest that since ToM is learned through social interaction perhaps a more dialogue-based assortment of tasks would lead to better results of LLMs when trained on such data. In conclusion, while current LLMs still seem to be lacking regarding ToM the progression from no understanding of ToM tasks to somewhat of an understanding of ToM tasks leaves the open question whether a fine-tuned model of current standards or a merely pre-trained rendition of further developed LLMs of future research efforts might show definitive signs of ToM. This demands both research in LLMs as well as ToM task databases and will likely prove to be an interesting development in the upcoming years.

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., Jolliffe, T., Mortimore, C., Robertson, M.: Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. Journal of Child Psychology and Psychiatry **38**(7), 813–822 (1997). https://doi.org/https://doi.org/10.1111/j.1469-7610.1997.tb01599.x, https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.1997.tb01599.x
2. 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
3. Bowman, S.R.: Eight things to know about large language models (2023)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
5. Call, J., Tomasello, M.: Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences **12**(5), 187–192 (2008)

6. Franchin, L.: Theory of Mind, pp. 1639–1644. Springer International Publishing, Cham (2022)
7. Kosinski, M.: Theory of mind may have spontaneously emerged in large language models (2023)
8. Le, M., Boureau, Y.L., Nickel, M.: Revisiting the evaluation of theory of mind through question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5872–5877. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1598, https://aclanthology.org/D19-1598
9. Marcus, G., Davis, E.: How not to test gpt-3 (2023), https://garymarcus.substack.com/p/how-not-to-test-gpt-3
10. Moghaddam, S.R., Honey, C.J.: Boosting theory-of-mind performance in large language models via prompting (2023)
11. 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
12. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
13. Sap, M., LeBras, R., Fried, D., Choi, Y.: Neural theory-of-mind? on the limits of social intelligence in large lms (2023)
14. Sap, M., Rashkin, H., Chen, D., Le Bras, R., Choi, Y.: Social IQa: Commonsense reasoning about social interactions. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4463–4473. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1454, https://aclanthology.org/D19-1454
15. Shapira, N., Levy, M., Alavi, S.H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., Shwartz, V.: Clever hans or neural theory of mind? stress testing social reasoning in large language models (2023)
16. Stone, V.E., Baron-Cohen, S., Knight, R.T.: Frontal lobe contributions to theory of mind. Journal of Cognitive Neuroscience **10**, 640–656 (1998)
17. Ullman, T.: Large language models fail on trivial alterations to theory-of-mind tasks (2023)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
19. 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
20. 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