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[1]*, Tom Kouwenhoven[1]*,
Werner de Valk[1], **Marco Spruit[1,2]**, and **Peter van der Putten[1]**
[1]Leiden Institute of Advanced Computer Science
[2]Leiden University Medical Centre
Corresponding author: m.j.van.duijn@liacs.leidenuniv.nl

Abstract

To what degree should we ascribe cognitive capacities to Large Language Models (LLMs), such as the ability to reason about intentions and beliefs known as Theory of Mind (ToM)? Here we add to this emerging debate by (i) testing 11 base- and instruction-tuned LLMs on capabilities relevant to ToM beyond the dominant false-belief paradigm, including non-literal language usage and recursive intentionality; (ii) using newly rewritten versions of standardized tests to gauge LLMs' robustness; (iii) prompting and scoring for open besides closed questions; and (iv) benchmarking LLM performance against that of children aged 7-10 on the same tasks. We find that instruction-tuned LLMs from the GPT family outperform other models, and often also children. Base-LLMs are mostly unable to solve ToM tasks, even with specialized prompting. We suggest that the interlinked evolution and development of language and ToM may help explain what instruction-tuning adds: rewarding cooperative communication that takes into account interlocutor and context. We conclude by arguing for a nuanced perspective on ToM in LLMs.

1 Introduction

Machines that can think like us have always triggered our imagination. Contemplation of such machines can be traced as far back as antiquity (Liveley and Thomas, 2020), and peaked with the advent of all kinds of 'automata' in the early days of the Industrial Revolution (Voskuhl, 2013) before settling in computer science from the 1950s (Turing, 1950). Currently people around the world can interact with powerful chatbots driven by Large Language Models (LLMs), such as OpenAI's ChatGPT (OpenAI, 2023), and wonder to what degree such systems are capable of thought.

LLMs are large-scale deep neural networks, trained on massive amounts of text from the web. They are vastly complex systems: even if all details about their architecture, training data, and optional fine-tuning procedures are known (which is currently not the case for the most competitive models), it is very difficult to oversee their capabilities and predict how they will perform on a variety of tasks. Researchers from linguistics (Manning et al., 2020), psychology (Binz and Schulz, 2023b; Kosinski, 2023; Webb et al., 2023), psychiatry (Kjell et al., 2023), epistemology (Sileo and Lernould, 2023), logic (Creswell et al., 2022), and other fields, have therefore started to study LLMs as new, 'alien' entities, with their own sort of intelligence, that needs to be probed with experiments, an endeavour recently described as 'machine psychology' (Hagendorff, 2023). This not only yields knowledge about what LLMs are capable of, but also provides a unique opportunity to shed new light on questions surrounding our own intelligence (Dillion et al., 2023; Binz and Schulz, 2023a).

Here we focus on attempts to determine 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 behaviour in social settings (Apperly, 2010). We first address the question **how LLMs perform** on standardized, language-based tasks used to assess ToM capabilities in humans. We extend existing work in this area, surveyed in Section 2, in four ways: by (i) testing 11 models (see Table 1) 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) using newly written versions of standardized tests with varying degrees of deviation from the originals; (iii) including open questions besides closed ones; and (iv) benchmarking LLM performance against that of children aged 7-8 (n=37) and 9-10 (n=36) on the same tasks. Section 3 contains details of our

*Equal contribution.

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 389–402
December 6–7, 2023. ©2023 Association for Computational Linguistics

test procedures for both children and LLMs. After reporting the results in Section 4, we turn to the question **how variation in performance of the LLMs we tested can be explained** in Section 5. We conclude by placing our findings in the broader context of strong links between language and ToM in human development and evolution, and tentatively interpret what it means for an LLM to pass (or fail) ToM tests.

We are aware of issues regarding LLM training and deployment, for example regarding the biases they inherit (Lucy and Bamman, 2021; Bender et al., 2021), problems for educators (Sparrow, 2022), and ethical concerns in obtaining human feedback (Perrigo, 2023). Ongoing reflection on the use of LLMs is necessary, but outside the scope of this paper.

2 Background

2.1 Large Language Models

The field of Natural Language Processing (NLP) has been revolutionized by the advent of Transformer models (Vaswani et al., 2017; Devlin et al., 2019), deep neural networks that can induce language structures through self-supervised learning. During training, such models iteratively predict masked words from context in large sets of natural language data. They improve at this task by building representations of the many morphological, lexical, and syntactic rules governing human language production and understanding (Manning et al., 2020; Rogers et al., 2021; Grand et al., 2022). Models exclusively trained through such self-supervision constitute what we refer to as 'base-LLMs' in this paper.

Base-LLMs can generate natural language when prompted with completion queries ('A mouse is an ...'). They can also be leveraged successfully for an array of other challenges, such as question-answering and translation, which often requires task-specific fine-tuning or prompting with specific examples, known as few-shot-learning (Brown et al., 2020). This makes them different from a new generation of LLMs that we refer to as 'instruct-LLMs' in this paper, and to which the currently most competitive models belong. In instruction-tuning, various forms of human feedback are collected, such as ranking most suitable responses, which then forms the reward-signal for further aligning these models to human preferences through reinforcement learning (Ouyang et al., 2022). The resulting LLMs can be prompted with natural language in the form of instructions to perform a wide variety of tasks directly, amounting to zero-shot learning (Wei et al., 2022).

A key realization is thus that LLMs are given either no explicitly labelled data at all, or, in the case of instruct-LLMs, data with human labels pertaining to relatively general aspects of communicative interaction. As such they are part of a completely different paradigm than earlier language models that were trained on, for example, data sets of human-annotated language structures (e.g. Nivre et al., 2016). This means that when LLMs are capable of such tasks as solving co-reference relationships or identifying word classes (Manning et al., 2020), this arises as an *emergent* property of the model's architecture and training on different objectives. Given that such emergent linguistic capabilities have been observed (Reif et al., 2019; Grand et al., 2022), it is a legitimate empirical question which other capacities LLMs may have acquired as 'by-catch'.

2.2 Theory of Mind in Humans and LLMs

ToM, also known as 'mindreading', is classically defined as the capacity to attribute mental states to others (and oneself), in order to explain and anticipate behaviour. The concept goes back to research in ethology in which Premack and Woodruff (1978) famously studied chimpanzees' abilities to anticipate behaviour of caretakers. When focus shifted to ToM in humans, tests were developed that present a scenario in which a character behaves according to its *false beliefs* about a situation, and not according to the reality of the situation itself—which a successful participant, having the benefit of spectator-sight, can work out (see Section 3.1).

Initial consensus that children could pass versions of this test from the age of 4 was followed by scepticism about additional abilities it presumed, including language skills and executive functioning, which led to the development of simplified false-belief tests based on eye-gaze that even 15 month-olds were found to 'pass' (Onishi and Baillargeon, 2005). While this line of research also met important criticism (for a review see Barone et al., 2019), it highlights two key distinctions in debate from the past decades: implicit-behavioural versus explicit-representational and innate versus learned components of ToM. Some researchers see results from eye-gaze paradigms as evidence for a

native or very early developing capacity for belief-attribution in humans (Carruthers, 2013) and hold that performance on more complex tests is initially 'masked' by a lack of expressive skills (cf. also Fodor, 1992). Others have attempted to explain eye-gaze results in terms of lower-level cognitive mechanisms (Heyes, 2014) and argued that the capacity for belief-attribution itself develops gradually in interaction with more general social, linguistic, and narrative competencies (Heyes and Frith, 2014; Milligan et al., 2007; Hutto, 2008). Two-systems approaches (Apperly, 2010) essentially reconcile both sides by positing that our mindreading capacity encompasses both a basic, fast, and early developing component and a more advanced and flexible component that develops later.

In computational cognitive research, a variety of approaches to modelling ToM have been proposed (e.g. Baker and Saxe, 2011; Arslan et al., 2017). More recently neural agents (Rabinowitz et al., 2018) have been implemented, along with an increasing number of deep-learning paradigms aimed at testing first- and second-order ToM via question-answering. Initially this was done with recurrent memory networks (Grant et al., 2017; Nematzadeh et al., 2018) using data sets of classic false-belief tests from psychology, but after issues surfaced with simple heuristics for solving such tasks, scenarios were made more varied and challenging (Le et al., 2019). From the inception of BERT as one of the first LLMs (Devlin et al., 2019), we have seen roughly two approaches for testing ToM in LLMs: many different ToM scenarios integrated in large benchmark suites (e.g. Sap et al., 2022; Srivastava et al., 2023; Sileo and Lernould, 2023; Ma et al., 2023; Shapira et al., 2023), and studies that modified standardized ToM tests as used in developmental and clinical research for prompting LLMs (e.g. Kosinski, 2023; Ullman, 2023; Bubeck et al., 2023; Brunet-Gouet et al., 2023; Chowdhery et al., 2022; Moghaddam and Honey, 2023; Marchetti et al., 2023). This paper adds to the latter tradition in four respects, as listed in the introduction.

3 Methodology

Here we describe our tasks and procedures for testing LLMs and children; all code, materials, and data are on OSF: `https://shorturl.at/FQR34`.

3.1 ToM Tests

Sally-Anne test, first-order (SA1) — The Sally-Anne test (Wimmer and Perner, 1983; Baron-Cohen et al., 1985) is a classic first-order false belief test. It relies on a narrative in which Sally and Anne stand behind a table with a box and a basket on it. When Anne is still present, Sally puts a ball in her box. When Sally leaves, Anne retrieves the ball from the box and puts it in her own basket. The story ends when Sally returns and the participant is asked the experimental question 'Where will Sally look for the ball?' The correct answer is that she will look in her box. We followed up by asking a motivation question, 'Why?', to prompt an explanation to the effect of 'she (falsely) believes the object is where she left it'.

Sally-Anne test, second-order (SA2) — While SA1 targets the participant's judgement of what a character *believes* about the location of an unexpectedly displaced object, in SA2 the participant needs to judge what a character *believes* that *another character believes* about the location of an ice-cream truck (Perner and Wimmer, 1985). Sally and Anne are in a park this time, where an ice-cream man is positioned next to the fountain. Anne runs home to get her wallet just while the ice-cream man decides to move his truck to the swings. He tells Sally about this, but unknown to her, he meets Anne on the way and tells her too. Sally then runs after Anne, and finds her mother at home, who says that Anne picked up the wallet and went to buy ice cream. The experimental question now is 'Where does Sally think Anne went to buy ice cream?', with as correct answer 'to the fountain', also followed up with 'Why?', to prompt an explanation to the effect of 'Sally doesn't know that the ice-cream man told Anne that he was moving to the swings'.

Strange Stories test (SS) — The Strange Stories test (Happé, 1994; Kaland et al., 2005) depicts seven social situations with non-literal language use that can easily be misinterpreted, but causes no problems to typically developed adults. To understand the situations, subjects must infer the characters' intentions, applying ToM. For example, in one of the items a girl wants a rabbit for Christmas. When she opens her present, wrapped in a big enough box, it turns out that she received a pile of books. She says that she is really happy with her gift, after which subjects are asked the experimental question 'Is what the girl says true?', with correct answer 'No'. They can motivate their

answer after the question 'Why does she say this?', with as correct answer 'to avoid her parents' feelings being hurt'. Items increase in difficulty and cover a lie, pretend-play scenario, practical joke, white lie (example above), misunderstanding, sarcasm, and double bluff.

Imposing Memory test (IM) — The Imposing Memory test was originally developed by Kinderman et al. (1998), but the test has been revised several times; we rely on an unpublished version created by Anneke Haddad and Robin Dunbar (van Duijn, 2016), originally for adolescents, which we adapted thoroughly to make it suitable for children aged 7-10. Our version features two different stories, followed by true/false questions, 10 of which are 'intentionality' and 12 are 'memory' questions. For instance, in one story Sam has just moved to a new town. He asks one of his new classmates, Helen, where he can buy post stamps for a birthday card for his granny. When Helen initially sends him to the wrong location, Sam wonders whether she was playing a prank on him or just got confused about the whereabouts of the shop herself. He goes and asks another classmate, Pete, for help. As in the original IM, the intentionality questions involve reasoning about different levels of recursively embedded mental states (e.g., at third-level: 'Helen *thought* Sam *did not believe* that she *knew* the location of the store that sells post stamps'), whereas the memory questions require just remembering facts presented in the story (e.g., to match third-level intentionality questions, three elements from the story are combined: 'Sam was looking for a store where they sell post stamps. He told Pete that he had asked Helen about this').

3.2 Scoring Test Answers

Test scores for both children and LLMs were determined in the following way. For each of the SA1 and SA2 items, as well as for the seven SS items, a correct answer to the experimental question yielded 1 point. These answers were discrete and thus easy to assess ('box', 'fountain', 'no', etc.). For the motivation question a consensus score was obtained from two expert raters, on a range from 0-2, with 0 meaning a missing, irrelevant, or wrong motivation, 1 meaning a partly appropriate motivation, and 2 meaning a completely appropriate motivation that fully explained why the character in each scenario did or said something, or had a mental or emotional mind state. Thus, the maximum score for the SA1, SA2, and SS was 3 points per item, which were averaged to obtain a score between 0 and 1. For each correct answer to a true/false question in the IM, 1 point was given. All scores and ratings can be found on OSF.

3.3 Deviations

We tested the LLMs on the original SA and SS scenarios, but also on manually created *deviations* that increasingly stray from their original formulations, to prevent LLMs from leveraging heuristics and memorizing relevant patterns from the training data. Thus, deviations probe the degree to which performance on ToM tests in LLMs generalizes. Deviation 0 was always the original test scenario (likely present in the training data); deviation 1 was a superficial variation on the original with only e.g., objects and names changed (similar to Kosinski (2023)), whereas deviation 2 was a completely new scenario where only the ToM-phenomenon at issue was kept constant (e.g., 'second-order false belief' or 'irony'). Since our adaptation of the IM test has hitherto not been used or published, we did not include deviations for this test.

3.4 Test Procedures for LLMs

We leveraged 11 state-of-the-art LLMs: 4 base-LLMs and 7 instruct-LLMs (see Table 1). Inference parameters were set such that their output was as deterministic as possible (i.e. a temperature $\approx$ zero or zero where possible) improving reproducibility. Each inference was done independently to avoid in-context learning or memory leakage between questions. This means that for each question, the prompt repeated the following general structure: [*instruction*] + [*test scenario*] + [*question*].

Instruct-LLMs were prompted in a question-answering format that stayed as close as possible to the questionnaires given to children, without any further custom prompting or provision of examples. Instructions were also similar to those given to children (e.g. 'You will be asked a question. Please respond to it as accurately as possible without using many words.'). The 'Why'-questions in SA1 and SA2 were created by inserting the experimental question and answer the LLM gave into the prompt: [*instruction*] + [*test scenario*] + [*experimental question*] + [*LLM answer*] +[*'Why?'*]. This was not necessary for SS, given that experimental and motivation questions could be answered independently.

Base-LLMs	Source	Size
Falcon	Penedo et al. (2023)	7B
LLaMA	Touvron et al. (2023)	30B
GPT-davinci	Brown et al. (2020)	175B
BLOOM	Scao et al. (2022)	176B
Instruct-LLMs	”	”
Falcon-instruct	Penedo et al. (2023)	7B
Flan-T5	Chung et al. (2022)	11B
GPT-3 (text-davinci-003)	Ouyang et al. (2022)	175B
GPT-3.5-turbo	Ouyang et al. (2022)	175B
PaLM2	Anil et al. (2023)	175-340B
PaLM2-chat	Anil et al. (2023)	175-340B
GPT-4	OpenAI (2023)	>340B

Table 1: LLMs used in this study. Model sizes are undisclosed for GPT-4 and for PaLM2 and PaLM2-chat, thus we base ourselves on secondary sources for estimations; Knight (2023) and Elias (2023), respectively.

For base-LLMs, known to continue prompts rather than follow instructions, staying this close to the children's questionnaires was not feasible. For the SA and SS we therefore fed base-LLMs the scenario as described before, but formulated the questions as text-completion exercises (e.g. 'Sally will look for the ball in the '). Additionally, when creating the motivation questions for SA1 and SA2, we inserted the *correct* answer to the experimental question, instead of the LLM's answer. This was because base-LLMs so often derailed in their output that the method described for instruct-LLMs did not yield sensible prompts. Base-LLMs thus had an advantage here over children and instruct-LLMs, who were potentially providing a motivation following up on an incorrect answer they gave to the experimental question.

For the closed questions in the IM we attempted to streamline the output of base-LLMs by including two example continuations in the desired answer format. These examples were based on trivial information we added to the scenarios, unrelated to the actual experimental questions. For example: 'Helen: I wear a blue jumper today. This is [incorrect]', where it was added in the story that Helen wears a green jumper. This pushed nearly all base-LLM responses towards starting with '[correct]' or '[incorrect]', which we then assessed as answers to the true/false questions. We considered a similar prompt structure for SA and SS, amounting to adopting few-shot learning for base-LLMs throughout (Brown et al., 2020), but given that reformulating questions as text-completion exercises was by itself effective to get the desired output format, we refrained from inserting further differences from how instruct-LLMs are prompted. It is important to note that our prompts were in general not optimized for maximal test performance, but rather designed to stay as uniform and close to the way children were tested as possible, enabling a fair comparison among LLMs and with child performance.

3.5 Test Procedures for Children

Children were recruited from one Dutch and one international school in the South-West of the Netherlands: 37 children in the younger group (7-8y) and 36 children in the older group (9-10y). Children were administered digital versions of the SA and SS for the younger group, and of the IM for the older group, which they completed individually on tablets or PCs equipped with a touch screen. Test scenarios and questions were presented in a self-paced text format and all SA and SS questions were followed by an open text field in which they had to type their answer. As the IM features long scenarios, voice-overs of the text were included to alleviate reading fatigue. Here children had to answer by pressing yes/no after each question. To reduce memory bottlenecks, accompanying drawings were inserted (see OSF) and navigating back and forth throughout the tests was enabled. Informed consent for each child was obtained from caretakers, and the study was approved by the Leiden University Science Ethics Committee (ref. no. 2021-18). Test answers were evaluated and scored parallel to the approach for LLMs (Section 3.2).

4 Results

4.1 Sally-Anne

Overall performance on SA1 versus SA2 is given in Figure 1, left column. Most base-LLMs perform above child level on first-order ToM (BLOOM, Davinci, LLaMA-30B) but fall at or or below child level on second-order ToM. A similar pattern is visible for instruct-LLMs: most models perform well above child level on first-order (GPT-4, GPT-3.5, PaLM2-chat, PaLM2), but not on second-order ToM. Exceptions are GPT-4 and GPT-3.5: while degrading on second-order, they remain above child level. For both base- and instruct-LLMs, smaller models tend to perform worse (Falcon-7B, Falcon-7B-I, FLAN-T5) with GPT-3's structurally low scores as striking exception. This is inconsistent with results reported by (Kosinski, 2023) for GPT-3, which is probably due to the fact that Kosinski applied a text-completion approach whereas we

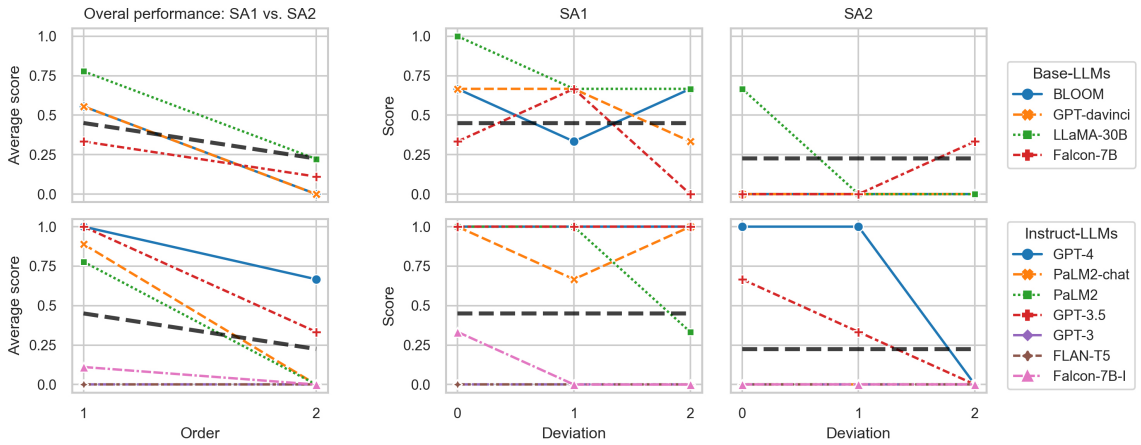

Figure 1: Performance on Sally-Anne tests for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts performance on first- and second-order ToM (i.e. SA1 vs. SA2), averaged over the original and rewritten test versions. Middle and left columns depict performance for SA1 and SA2 over levels of deviation from the original test (0, 1, and 2; see Section 3.3). Dashed lines indicate child performance (n=37, age 7-8 years).

prompted GPT-3 with open questions.

When we consider the performance on SA1 and SA2 over deviations (middle and right columns in Figure 1), we see once more that almost all LLMs struggle with second-order ToM, since performance decreases already on deviation 0 (i.e. the original test scenario), except for GPT-3.5 and GPT-4. Yet, it is the *combination* of second-order ToM and deviation 2 that pushes also GPT-3.5 and GPT-4 substantially below child levels, except for Falcon-7B, although the chat-optimized version of this model (Falcon-7B-I) fails on all second-order questions.

4.2 Strange Stories

General performance on SS is given in Figure 2, left column. Whereas child performance declines as items become more complex (from 1 to 7; see Section 3.1), this is overall less the case for LLM performance. For instruct-LLMs, we see that GPT-4 approaches perfect scores throughout. GPT-3 and GPT-3.5 perform at or close to child level on item 1, after which their performance somewhat declines, while staying well above child level. Other instruct-LLMs show a mixed picture: PaLM2-chat and FLAN-T5 surpass child level earlier than PaLM2. Interestingly, smaller FLAN-T5 outperforms large PaLM and PaLM2-chat on more difficult items. Falcon-7B-I, as smallest instruct-LLM, performs overall worst.

If performance is plotted over deviations (right column in Figure 2) we see little impact on most base-LLMs. For instruct-LLMs, it is striking that deviation levels have almost no effect on the larger models (GPT-4, PaLM2, PaLM2-chat, GPT-3, GPT-3.5), but do more dramatically lower performance of smaller models (FLAN-T5, Falcon-7B-I). In sum, base-LLMs perform below child level, except for the most complex items. Several large instruct-LLMs match or surpass child level throughout, others only for more complex items. Unlike for SA, deviation levels seem to have little negative impact.

4.3 Imposing Memory

The classical finding for the IM test is that error rates go up significantly for questions involving higher levels of recursive intentionality, but not for memory questions on matched levels of complexity, suggesting a limit to the capacity for recursive ToM specifically (Stiller and Dunbar, 2007).[1] We verified this for our child data (n=36) with two mixed linear models for memory and intentional questions with random intercepts. We included five predictors that were contrast-coded such that each predictor indicated the difference in average performance with the previous level. For intentional questions, only the difference between level two and one was significant ($\beta = -0.222, p < .05$), marking a cut-off point after which performance remained consistently low. For memory questions, performance

[1]While there is consensus in the literature that higher levels of intentionality are significantly harder for participants than lower levels, by various measures, there is debate about the difference with memory questions; see e.g. Lewis et al. (2017). For a critical discussion of measuring recursive intentionality in general, see Wilson et al. (2023).

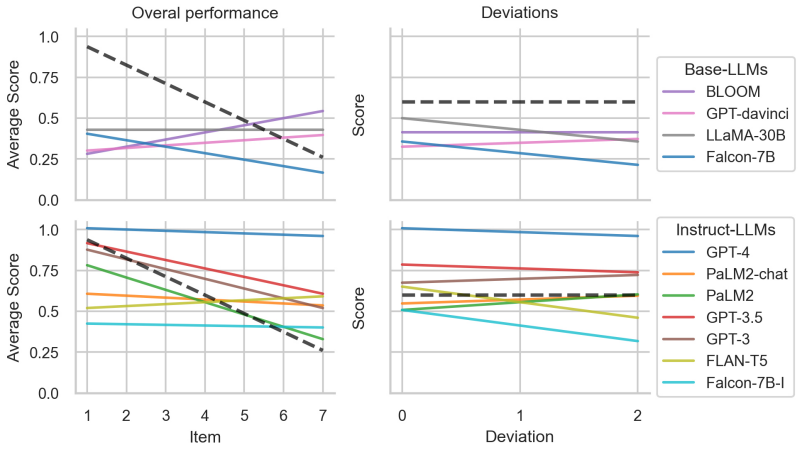

Figure 2: Performance on Strange Stories for base-LLMs (top row) and instruct-LLMs (bottom row). Left column shows overall performance, averaged over levels of deviation from the original test. Right column shows performance over deviation levels, averaged over items. Dashed lines indicate child performance (n=37, 7-8y).

remained high across all levels ($> .85$), except for level four, where scores were significantly lower than at level three ($\beta = -0.292, p < .00$), but went up again at level five ($\beta = 0.208, p < .00$). Thus, in line with earlier work, we find a cut-off point after which scores on intentionality questions remained consistently low, compared to scores on matched memory questions. We have no clear explanation for the dip in performance on memory questions at level four, but observe that it is driven by low scores on only one specific question out of a total of four for this level, which children may have found confusing.

In Figure 3 we see that all base-LLMs perform below child level, in general and on both intentionality and memory questions, and there is little variation in performance, except that larger base-LLMs (BLOOM, GPT-davinci) improve on higher levels of recursion. Regarding instruct-LLMs, we see largely the same picture, as they almost all perform below child level, in general and on both types of questions. The exception is GPT-4, which performs consistently well on all levels and stays above child level after second-order intentionality. For the difference between memory and intentional questions, instruct-LLMs perform better on easier memory questions, and drop towards the end, while on intentional questions, they already start lower and stay relatively constant. Lastly, it is remarkable that FLAN-T5, as one of the smallest instruct-LLMs, overall increases performance as recursion levels go up, and ends at child level. For GPT-3.5, which performs worst of all instruct-LLMs on this task, we see the exact opposite.

4.4 Notes on Child Performance

It can be observed that performance for SA was overall low compared to what could be expected from children aged 7-8 years: $\bar{x} = 0.45$ for SA1 and $\bar{x} = 0.225$ for SA2. We have two complementary explanations for this. Firstly, as discussed in Section 3.5, children had to read the tests on a screen, after which they had to type answers in open text fields. This is a challenging task by itself that relies on additional skills including language proficiency, conscientiousness, digital literacy, and more. Secondly, whereas 'passing' originally only means that a child can work out where Sally will look (for the ball, or for Anne on her way to buy ice cream), we also asked for a motivation, which makes the test more demanding. For the SS, completed by the same group of children, we see the expected pattern that scores show a downward tendency as test items increase in difficulty. The older group, aged 9-10, completed the IM. As discussed in Section 4.3, scores resonate with earlier work. Given that we see child performance not as the central phenomenon under observation in this paper, but rather as a reference for LLM performance, further discussion is outside our scope.

5 Discussion

Summing up the results for the Sally-Anne tests, while it is less surprising that base-LLMs and smaller instruct-LLMs struggle with increasing test complexity and deviations, it is striking that second-order ToM immediately perturbs some large instruct-LLMs (e.g. PaLM2-chat), and that adding deviations from the original test formula-

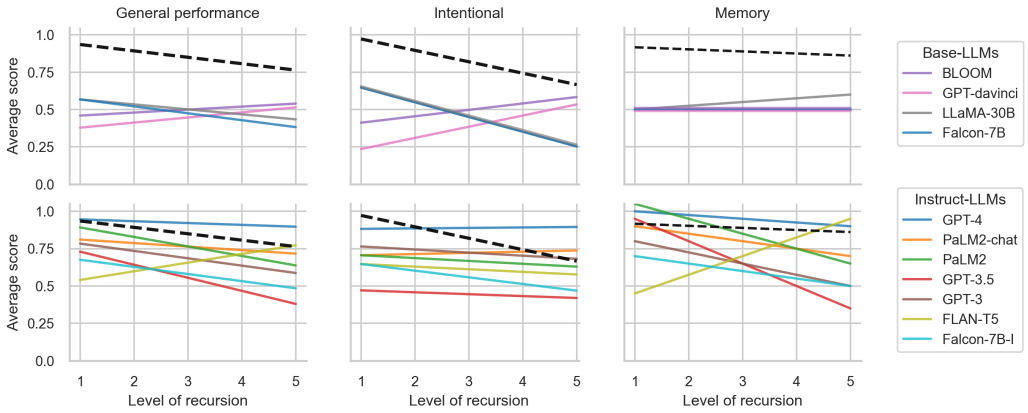

Figure 3: Performance on Imposing Memory test for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts overall performance over five levels of recursion, averaged over deviations. Middle and left columns depict performance for Memory and Intentional questions. Dashed lines indicate child performance (n=36, 9-10y).

tions pushed performance of even the most competitive models down (e.g. GPT-4, GPT-3.5). This initially suggests that performance on ToM tasks does not generalize well beyond a few standard contexts in LLMs, in line with earlier work (Sap et al., 2022; Shapira et al., 2023; Ullman, 2023).

For the Strange Stories we saw that base-LLMs perform generally below child level. Most instruct-LLMs perform close to or above child level, particularly as items become more complex and child performance drops much more dramatically than LLM performance. Levels of deviation from the original test formulation seem to have made almost no impact for the SS, suggesting that the capacity to deal with non-literal language targeted by the Strange Stories test *does* generalize to novel contexts. We conclude that instruct-LLMs are quite capable at interpreting non-literal language, a skill that in humans involves ToM. Since the training data of LLMs includes numerous books and fora, which are typically rich in irony, misunderstanding, jokes, sarcasm, and similar figures of speech, we tentatively suggest that LLMs are in general well-equipped to handle the sort of scenarios covered in the Strange Stories. This should in theory include base-LLMs, but it could be that their knowledge does not surface due to the test format, even after specialized prompting. Going one step further, we hypothesize that Sally-Ann is generally harder for LLMs given that this test relies less on a very specific sort of advanced language ability, but more on a type of behaviourally-situated reasoning that LLMs have limited access to during training (see also Mahowald et al., 2023).

The Imposing Memory test was the most challenging for both base- and instruct-LLMs. Since our version of it was never published before, it constitutes another robustness test, which only GPT-4 as largest instruct-LLM seems to pass well.

The gap between base- and instruct-LLMs is best summarized in Figure 4. Here we see that no base-LLM achieves child level: all LLMs approaching or exceeding child performance are larger instruct-LLMs. Our adapted prompts and insertion of correct answers for motivation questions did not make a difference. We suggest that another issue for base-LLMs, besides the prompt format, was prompt length. This was highest for IM, which can explain why they struggled most with this test. Prompt length, in relation to the models' varying context window sizes and ability to engage in what Hagendorff et al. (2023) call chain-of-thought reasoning, merits further research (see also Liu et al., 2023). We tested whether there was a difference between model performance on closed versus open questions across all three tasks, but found no signal: the models that struggled with closed questions were also those that performed low on open questions (for more details and additional information on prompting, see Appendix A on OSF).

Evidence is emerging that most LLM capacities are learned during self-supervised pre-training (Gudibande et al., 2023; Ye et al., 2023), which suggests that base-LLMs are essentially 'complete' models. Yet instruction-tuning, even in small amounts (Zhou et al., 2023), adds adherence to the desired interaction format and teaches LLMs, as it were, to apply their knowledge appropriately. We see a parallel between instruction-tuning and the role for *rewarding cooperative communication*

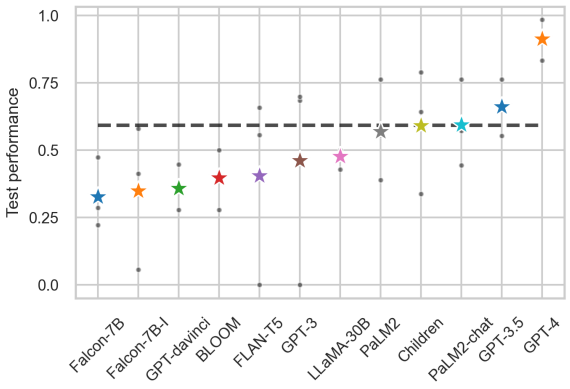

Figure 4: Grand mean performance (stars) of all mean test scores (dots) for children and LLMs.

in human evolution and development. It has been argued extensively that human communication is fundamentally cooperative in that it relies on a basic ability and willingness to engage in mental coordination (e.g Verhagen, 2015; Grice, 1975). It is a key characteristic of the socio-cultural niche in which we evolved that, when growing up, we are constantly being rewarded for showing such willingness and cooperating with others to achieve successful communicative interactions (Tomasello, 2008). Reversely, if we do not, we are being punished, explicitly or implicitly via increasing social exclusion (David-Barrett and Dunbar, 2016). This brings us back to our context: instruction-tuning essentially rewards similar cooperative principles, but punishes the opposite, which may amount to an enhanced capacity for *coordinating with an interaction partner's perspective*, in humans and LLMs alike. This is reflected in performance on ToM tasks, which are banking on this capacity too.

Finally, we do not claim that LLMs that performed well also have ToM in the way that humans have it. Validity of cognitive tests such as those used in ToM research is a general issue (e.g. van Duijn, 2016). Yet for humans ToM tests are validated 'quick probes': decades of research have shown that proficiency on such tests *correlates* with an array of real-world social and cognitive abilities (Beaudoin et al., 2020). For LLMs we are in a very early stage of figuring out what is entailed by proficon ToM tests: on the one hand it is impressive that some models show a degree of robust performance, without explicit training on ToM. On the other hand it remains an open question whether this amounts to any actual capacities in the social-cognitive domain, in which they are clearly very differently grounded (if at all) compared to humans.

For future research we believe in the format of testing models that differ in other respects than just size, on a varied array of tasks, with multiple tests per test item, to gain further insight into the aspects that explain variability in performance. For this, more openness about architecture and training procedures of current and future LLMs is imperative. In addition, we believe to have contributed to the debate by benchmarking LLM results on child data, but more of this is needed. We had limited samples and age distributions, and tests were not presented in optimal ways (see Section 3.5).

We emphasize that our results need to be seen within the time frame of late Spring 2023. The fast pace with which LLMs are currently released and, in some cases, updated, makes them a moving target. There are indications that specific capacities of models from the GPT-family have declined over time, perhaps as a result of such updates (e.g., handling math problems and producing code; Chen et al., 2023). Future studies need to address how such developments impact the capacities assessed in this paper.

6 Conclusion

We have shown that a majority of recent Large Language Models operate below performance of children aged 7-10 on three standardized tests relevant to Theory of Mind. Yet those that are largest in terms of parameters, and most heavily instruction-tuned, surpass children, with GPT-4 well above all other models, including more recent competitors like PaLM2-chat and PaLM2 (see Figure 4). We have interpreted these findings by drawing a parallel between instruction-tuning and rewarding cooperative interaction in human evolution. We concede that researching the degree to which LLMs are capable of anything like thought in the human sense has only just begun, which leaves the field with exciting challenges ahead.

Acknowledgements

This research took place in the context of the project *A Telling Story*, financed by the Dutch Research Council NWO (VI.Veni.191C.051). We are grateful to the children and their caregivers and teachers for participating in our research, and we thank Li Kloostra, Lola Vandame, and three anonymous reviewers for their help and constructive feedback.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403v3*.

Ian Apperly. 2010. *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press.

Burcu Arslan, Niels A Taatgen, and Rineke Verbrugge. 2017. Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in psychology*, 8:275.

Chris Baker and Rebecca Saxe. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Pamela Barone, Guido Corradi, and Antoni Gomila. 2019. Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57:101350.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Marcel Binz and Eric Schulz. 2023a. Turning large language models into cognitive models. *arXiv preprint arXiv.2306.03917*.

Marcel Binz and Eric Schulz. 2023b. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Eric Brunet-Gouet, Nathan Vidal, and Paul Roux. 2023. Do conversational agents have a theory of mind? a single case study of chatgpt with the hinting, false beliefs and false photographs, and strange stories paradigms.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4.

Peter Carruthers. 2013. Mindreading in infancy. *Mind & Language*, 28(2):141–172.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Tamas David-Barrett and Robin I. M. Dunbar. 2016. Language as a coordination tool evolves slowly. *R. Soc. open sci.*, 3:160259.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Jennifer Elias. 2023. Google's newest A.I. model uses nearly five times more text data for training than its predecessor. Accessed on: 2023-05-30.

J.A. Fodor. 1992. A theory of the child's theory of mind. *Cognition*, 44(3):283–296.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.

Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and semantics. Vol. 3: Speech acts*, pages 41–58. Academic Press, New York.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms.

T. Hagendorff, S. Fabi, and M. Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Compututer Science*.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Francesca G.E. Happé. 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 Developmental disorders*, 24(2):129–154.

Cecilia Heyes. 2014. False belief in infancy: a fresh look. *Developmental Science*, 17(5):647–659.

Cecilia M. Heyes and Chris D. Frith. 2014. The cultural evolution of mind reading. *Science*, 344(6190):1243091.

Daniel D. Hutto. 2008. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. The MIT Press.

Nils Kaland, Annette Møller-Nielsen, Lars Smith, Erik Lykke Mortensen, Kirsten Callesen, and Dorte Gottlieb. 2005. The Strange Stories test - a replication study of children and adolescents with Asperger syndrome. *European child & adolescent psychiatry*, 14(2):73–82.

P. Kinderman, R. Dunbar, and R. P. Bentall. 1998. Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, (2):191–204.

Oscar Kjell, Katarina Kjell, and H Andrew Schwartz. 2023. Ai-based large language models are ready to transform psychological health assessment.

Will Knight. 2023. A new chip cluster will make massive ai models possible. Accessed on: 2023-05-30.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. 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)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Penelope A. Lewis, Amy Birch, Alexander Hall, and Robin I. M. Dunbar. 2017. Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, 12(7):1063–1071.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Genevieve Liveley and Sam Thomas. 2020. Homer's intelligent machines: AI in antiquity.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Antonella Marchetti, Cinzia Di Dio, Angelo Cangelosi, Federico Manzi, and Davide Massaro. 2023. Developing chatgpt's theory of mind. *Frontiers in Robotics and AI*, 10.

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. 2007. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Kristine H. Onishi and Renée Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Josef Perner and Heinz Wimmer. 1985. "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

Billy Perrigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Accessed on: 2023-01-25.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or neural theory of mind? stress testing social reasoning in large language models.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic. *arXiv preprint arXiv.2305.03353*.

Jeff Sparrow. 2022. 'Full-on robot writing': the artificial intelligence challenge facing universities. Accessed on: 2023-01-25.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex

Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Mish-

erghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

James Stiller and Robin IM Dunbar. 2007. Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104.

Michael Tomasello. 2008. *Origins of Human Communication*. MIT Press, Cambridge, MA.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX:433–460.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Max J van Duijn. 2016. *The lazy mindreader: a humanities perspective on mindreading and multiple-order intentionality*. Ph.D. thesis, Leiden University.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Arie Verhagen. 2015. Grammar and cooperative communication. In Ewa Dabrowska and Dagmar Divjak, editors, *Handbook of Cognitive Linguistics*, pages 232–252. De Gruyter Mouton, Berlin, München, Boston.

Adelheid Voskuhl. 2013. One introduction: Androids, enlightenment, and the human-machine boundary.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Robert Wilson, Alexander Hruby, Daniel Perez-Zapata, Sanne W. van der Kleij, and Ian A. Apperly. 2023. Is recursive "mindreading" really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 152(5):1454–1468.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.