Theory of mind in language, minds, and machines: a multidisciplinary approach

Dijk, B.M.A. van

Chapter 7

Theory of Mind in Large Language Models

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 standardised tests to gauge LLMs' robustness; (iii) prompting and scoring for open besides closed questions; and (iv) benchmarking LLM performance against that of children aged 7-10y 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 specialised 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.

This work was originally published as: Van Duijn, M.J.,* Van Dijk, B.M.A.,* Kouwenhoven, T.,* De Valk, W.M., Spruit, M.R., and Van Der Putten, P.W.H. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 389-402. Association for Computational Linguistics. (* denotes equal contribution.)

7.1 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, 2019) 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 (Achiam et al., 2024), 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, 2023; Kosinski, 2024; Webb et al., 2023), psychiatry (Kjell et al., 2023), epistemology (Sileo and Lernould, 2023), logic (Creswell et al., 2023), 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 (Binz and Schulz, 2024; Dillion et al., 2023).

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, 2012). We first address the question **how LLMs perform on standardised, language-based tasks used to assess ToM capabilities in humans.** We extend existing work in this area, surveyed in Section 7.2, in four ways:

1. By testing 11 LLMs (see Table 7.1) 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*...);

2. By using newly written versions of standardised tests with varying degrees of deviation from the originals;

3. By including open questions besides closed ones;

4. By benchmarking LLM performance against that of children aged 7-8 (n=37) and 9-10 (n=36) on the same tasks.

Section 7.3 contains details of our test procedures for both children and LLMs. After reporting the results in Section 7.4, we turn to the question **how variation in performance of the LLMs we tested can be explained** in Section 7.5. 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 a LLM to pass (or fail) ToM tests.

We are aware of issues regarding LLM training and deployment, for example regarding the biases they inherit (Bender et al., 2021; Lucy and Bamman, 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 chapter.

7.2 Background

Large language models

The field of Natural Language Processing (NLP) has been revolutionised by the advent of the Transformer architecture (Devlin et al., 2019; Vaswani et al., 2017) in 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 (Grand et al., 2022; Manning et al., 2020; Rogers et al., 2020). Models exclusively trained through such self-supervision constitute what we refer to as 'base-LLMs' in this chapter.

Base-LLMs can generate natural language when they are 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 chapter, 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.

A key realisation 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, datasets 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 (Grand et al., 2022; Reif et al., 2019), it is a legitimate empirical question which other capacities LLMs may have acquired as 'by-catch'.

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.

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 old children 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; Hutto, 2008; Milligan et al., 2007). Two-systems approaches (Apperly, 2012) essentially reconcile both sides by positing that our ToM 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 has been proposed (e.g. Arslan et al., 2017; Baker et al., 2011). More recently neural agents (Rabinowitz et al., 2018b) 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 datasets 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 language models (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. Ma et al., 2023a; Sap et al., 2022; Shapira et al., 2024; Sileo and Lernould, 2023; Srivastava et al., 2023), and studies that modified standardised ToM tests as used in developmental and clinical research for prompting LLMs (e.g. Brunet-Gouet et al., 2023; Bubeck et al., 2023; Chowdhery et al., 2022; Kosinski, 2024; Marchetti et al., 2023; Moghaddam and Honey, 2023; Ullman, 2023). This chapter adds to the latter tradition in four respects, as explained in the introduction.

7.3 Methods

Here we describe our tasks and procedures for testing LLMs and children.[1]

Theory of Mind tests

Sally-Anne test, first-order (SA1) – The Sally-Anne test (Baron-Cohen et al., 1985; Wimmer and Perner, 1983) 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.

[1] All code, materials, and data are available on OSF: `https://osf.io/426p9/`.

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 cause no problems to typically developed adults. To understand the situations, subjects must infer the characters' intentions, applying ToM. For example, in one of the test 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 7-10y. Our version features two different stories, fol-

lowed by true/false questions, 10 of which are 'intentionality' and 12 of which 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 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 (to match third-level intentionality questions, three elements from the story are combined, e.g. 'Sam was looking for a store where they sell post stamps. He told Pete that he had asked Helen about this').

Testing procedures

Scoring – For both children and LLMs test scores 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 mental 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, and IM scores were averaged over its items as well. All scores and ratings can be found on OSF.

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 memorising relevant patterns from the training data. Thus, deviations probe the degree to which performance on ToM tests in LLMs generalises. Deviation 0 was always the original test scenario (likely present in the training data); deviation 1 was a superficial variation on the original, e.g. with only objects and names changed (similar to Kosinski (2024)), whereas deviation 2 was a completely new scenario where only the ToM-phenomenon at issue

	LLMs	Source	Size
Base	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	Falcon-instruct	Penedo et al. (2023)	7B
	Flan-T5	Chung et al. (2024)	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	Achiam et al. (2024)	>340B

Table 7.1: 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.

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.

Testing LLMs

We leveraged 11 state-of-the-art LLMs: 4 base-LLMs and 7 instruct-LLMs (see Table 7.1). Inference parameters were set such that their output was as deterministic as possible (i.e. a temperature $\approxeq$ 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 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`] + [`question`] + [`LLM answer`] + [`'Why?'`]. This was not necessary for SS, given that experimental and motivation questions could be answered independently.

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 were prompted. It is important to note that our prompts were in general not optimised 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.

Testing 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 Figure 1.2 in

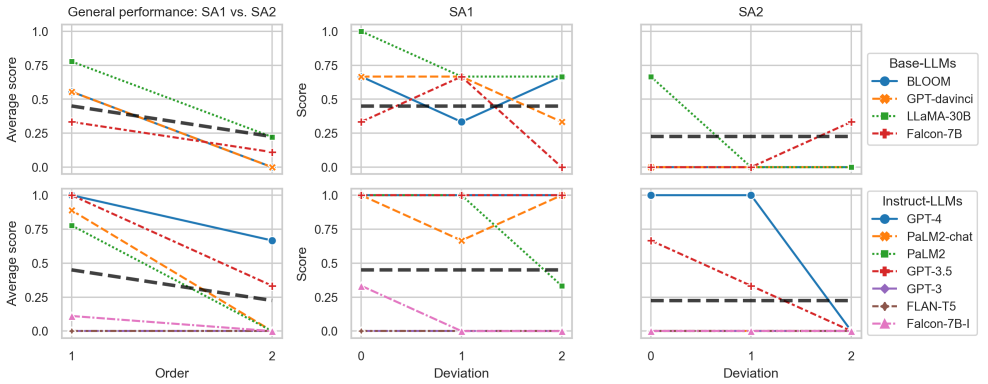

Figure 7.1: Performance on Sally-Anne tests for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts performance on first- and second-order ToM (i.e. SA1 vs. SA2), averaged over the original and rewritten test versions (deviations). Middle and right columns depict performance for SA1 and SA2 over levels of deviation from the original test (0, 1, and 2 as explained in Section 7.3). Dashed black lines indicate average child performance (n=37, age 7-8 years).

the Introduction) and navigating back and forth throughout the tests was enabled. Informed consent for each child was obtained from caregivers, 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.

7.4 Results

Sally-Anne Test

Overall performance on SA1 versus SA2 is given in Figure 7.1, left column. Most base-LLMs perform above child level on first-order ToM (BLOOM, Davinci, LLaMA-30B) but fall at 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 (2024) for GPT-3, which is probably due to the fact that Kosinski applied a text-completion approach whereas we prompted GPT-3 with open questions.

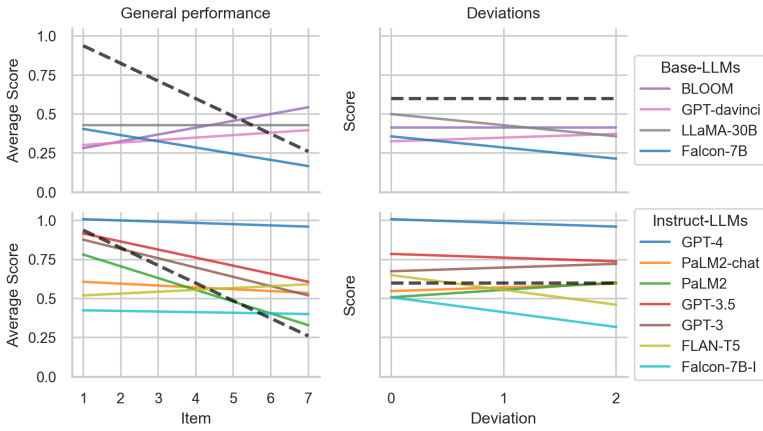

Figure 7.2: Performance on Strange Stories for base-LLMs (top row) and instruct-LLMs (bottom row). Left column shows overall performance, averaged over deviations from the original test. Right column shows performance over deviations, averaged over items. Dashed black lines indicate average child performance (n=37, 7-8y).

When we consider the performance on SA1 and SA2 over deviations (middle and right columns in Figure 7.1), 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 instruction-tuned version of this model (Falcon-7B-I) fails on all second-order questions.

Strange Stories Test

General performance on SS is given in Figure 7.2, left column. Whereas child performance declines as items become more complex (from 1 to 7; see Section 7.3), this is overall less the case for LLM performance. As a result, all models surpass child level at some point, except for the smallest model, Falcon-7B. All base-LLMs score below child level on most items but perform above child level on the most difficult ones, except Falcon-7B. 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 declines somewhat, while staying well above child level. Other instruct-LLMs show a mixed picture: PaLM2-chat and FLAN-T5 surpass child

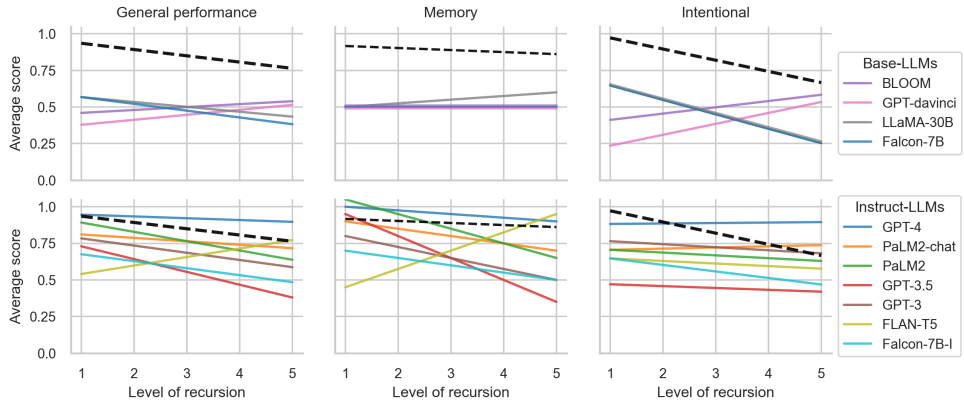

Figure 7.3: Performance on Imposing Memory test for base-LLMs (top row) and instruct-LLMs (bottom row). Left column depicts overall performance over five levels of recursion, averaged over deviations. Middle and right columns depict performance for Memory and Intentional questions. Dashed lines indicate average child performance (n=36, 9-10y).

level earlier than PaLM2. Interestingly, smaller FLAN-T5 outperforms larger PaLM2 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 7.2) 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 the SA test, deviation levels seem to have little negative impact for SS.

Imposing Memory Test

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).[2]

[2]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

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 IM 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 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 7.3 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 first-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.

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 $\bar{x} = 0.23$ for SA2. We have two complementary explanations for this. Firstly, as discussed in Section 7.3, 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 in-

questions; see e.g. Lewis et al. (2017). For a critical discussion of measuring recursive intentionality in general, see Wilson et al. (2023).

cluding 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 test, completed by the same group of children, we see the expected pattern that scores show a downward tendency as test items become increasingly difficult. The older group, aged 9-10, completed the IM. As discussed above, IM scores resonate with earlier work. Given that we see child performance not as the central phenomenon under observation in this chapter, but rather as a reference for LLM performance, further discussion is outside our scope.

7.5 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 formulations pushed down performance of even the most competitive models (e.g. GPT-4, GPT-3.5). This initially suggests that performance on ToM tasks does not generalise well beyond a few standard contexts in LLMs, in line with earlier work (Sap et al., 2022; Shapira et al., 2024; Ullman, 2023).

For the Strange Stories test 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 test, suggesting that the capacity to deal with non-literal language targeted by the Strange Stories*does* generalise 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 specialised prompting. Going one step further, we hypothesise that Sally-Anne is generally harder for LLMs given that this test relies less on a very specific sort of

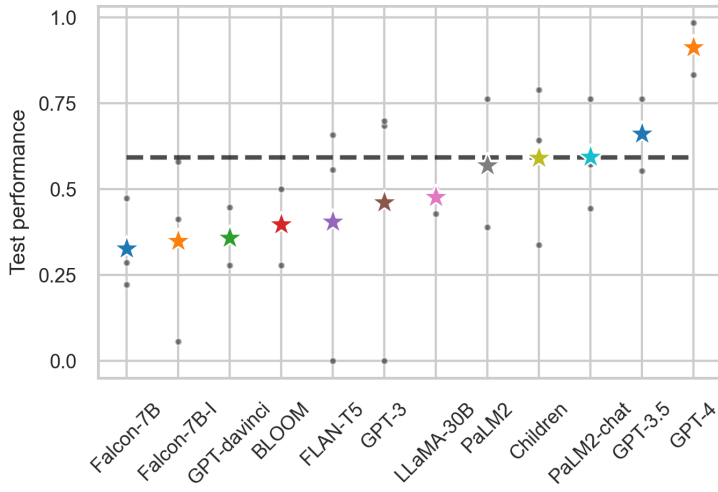

Figure 7.4: Grand mean performance (stars) of all mean test scores (dots) for children and LLMs.

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., 2024). The Imposing Memory test was the most challenging for both base- and instruct-LLMs. Since our version of this test 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 summarised in Figure 7.4. 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 for the SA test 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 see OSF).

Evidence is emerging that most LLM capacities are learned in the self-supervised pre-training phase (Gudibande et al., 2023; Ye et al., 2023), which suggests that base-

LLMs are essentially 'complete' models. Instruction-tuning, however, even in small amounts adds adherence to the desired interaction format and teaches LLMs, as it were, to apply their knowledge appropriately (Zhou et al., 2023a). We see a parallel between instruction-tuning and the role for *rewarding cooperative communication* 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 Grice, 1975; Verhagen, 2015). 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.

7.6 Conclusion

We have shown that the majority of recent LLMs operate below performance of children aged 7-10y on three standardised tests relevant to ToM. 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. We have interpreted these findings by drawing a parallel between instruction-tuning and rewarding cooperative interaction in human evolution. We conclude 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.