Neural Theory-of-Mind?
On the Limits of Social Intelligence in Large LMs

Maarten Sap[♠♢] **Ronan Le Bras**[♠] **Daniel Fried**[♢] **Yejin Choi**[♠♡]

[♠]Allen Institute for AI, Seattle, WA, USA

[♢]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

[♡]Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA

maartensap@cmu.edu

Abstract

Social intelligence and Theory of Mind (ToM), i.e., the ability to reason about the different mental states, intents, and reactions of all people involved, allow humans to effectively navigate and understand everyday social interactions. As NLP systems are used in increasingly complex social situations, their ability to grasp social dynamics becomes crucial.

In this work, we examine the open question of social intelligence and Theory of Mind in modern NLP systems from an empirical and theory-based perspective. We show that one of today's largest language models (GPT-3; Brown et al., 2020) lacks this kind of social intelligence out-of-the box, using two tasks: SocialIQa (Sap et al., 2019b), which measures models' ability to understand intents and reactions of participants of social interactions, and ToMi (Le et al., 2019), which measures whether models can infer mental states and realities of participants of situations.

Our results show that models struggle substantially at these Theory of Mind tasks, with well-below-human accuracies of 55% and 60% on SocialIQa and ToMi, respectively. To conclude, we draw on theories from pragmatics to contextualize this shortcoming of large language models, by examining the limitations stemming from their data, neural architecture, and training paradigms. Challenging the prevalent narrative that only scale is needed, we posit that person-centric NLP approaches might be more effective towards neural Theory of Mind.

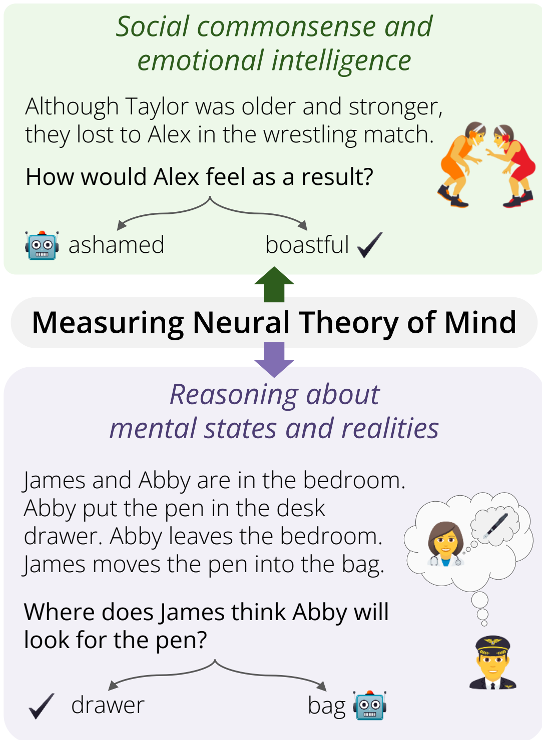


Figure 1: Theory of Mind is the ability for humans to reason about the intents, reactions, and mental states of others. We asses these abilities in LLMs through two question-answering tasks that measure social commonsense and emotional intelligence (SocialIQa; top) and reasoning about people's mental states and realities (ToMi; bottom); finding that GPT-3 (🤖) struggles on both tasks. We discuss why that may be, drawing from theories of the pragmatics of language.

1 Introduction

With the growing prevalence of AI and NLP systems in everyday social interactions, the need for AI systems with *social intelligence* and *Theory of Mind* (ToM), i.e., the ability to infer and reason about the intents, feelings, and mental states of others, becomes increasingly evident (Pereira et al., 2016; Langley et al., 2022). For humans, Theory of Mind is a crucial component that enables us to interact and communicate effectively with each other (Premack and Woodruff, 1978; Apperly, 2010). It allows us, for example, to infer that someone likely feels boastful instead of ashamed after winning a wrestling match (Fig. 1; top). In addition, ToM also enables us to reason about people's mental realities, e.g., if someone was out of the room while a pen was moved, she will likely search for the pen where she last saw it instead of where it was moved to (Fig. 1; bottom).

While humans develop it naturally, TOM and social intelligence remain elusive goals for modern AI systems (Choi, 2022), including large neural language models (LLMs). With advances in scaling the sizes of models and datasets, these LLMs have proven very impressive at generating human-like language for conversational, summarization, or sentence continuation settings, often with zero to few examples to learn from (Brown et al., 2020; Clark et al., 2021; Chowdhery et al., 2022). However, increasing scrutiny has shed light on the shortcomings of these LLMs, showing that they often fall prey to spurious correlational patterns instead of displaying higher-order reasoning (Elkins and Chun, 2020; Dale, 2021; Marcus, 2022).

In line with EMNLP 2022's theme, we examine the open research question of whether and how much LLMs—which are the backbone of most modern NLP systems—exhibit social intelligence and TOM abilities. Using some of the largest English models in existence (GPT-3; Brown et al., 2020), we demonstrate that out-of-the-box LLMs struggle at two types of reasoning abilities that requisites for Theory of Mind (shown in Fig. 1). We argue that these reasoning abilities are necessary but not sufficient for Theory of Mind, and that larger models will likely provide upper bounds on what equivalent-but-smaller models are capable of.

We first assess whether LLMs can reason about *social commonsense and emotional intelligence* with respect to social interactions (§3), using the SOCIALIQA benchmark (Sap et al., 2019b) illustrated in Fig. 1 (top). Results show our best performing few-shot GPT-3 setup achieving only 55% accuracy, lagging >30% behind human performance. Furthermore, social reasoning about the protagonists of situations is easier for GPT-3 (5-15% absolute difference) compared to reasoning about other secondary participants.

Second, we measure LLMs' ability to *understand other people's mental states and realities* in short stories (§4). We use the TOMI QA benchmark (illustrated in Fig. 1; bottom; Le et al., 2019), which was inspired by the classic Sally-Ann False Belief Theory of Mind test (Baron-Cohen et al., 1985). Here, our results show that GPT-3 models peak at 60% accuracy on questions about participants' mental states, compared to 90–100% on factual questions.

Our novel insights show that reasoning about social situations and false beliefs still presents a significant challenge for large language models, despite their seemingly impressive performance on tasks that could require social intelligence (e.g., story generation, dialogues). In §5, we first examine these shortcomings; drawing on theories of the pragmatics of language, we speculate that the type of texts in LLMs' training datasets could substantially limit learning social intelligence. Then, we outline some possible future directions towards socially aware LLMs, reflecting on the feasibility of interactional data selection, person-centric inductive biases, and interaction-based language learning. Our findings suggest that only increasing the scale of LLMs is likely not the most effective way to create socially aware AI systems, challenging a prevalent narrative in AI research (Narang and Chowdhery, 2022).

2 Theory of Mind & Large LMs

Why do LLMs need Theory of Mind? Social intelligence, Theory of Mind, and commonsense reasoning have been a longstanding but elusive goal of artificial intelligence for decades (Gunning, 2018; Choi, 2022). These reasoning abilities are becoming increasingly necessary as AI assistants are used in situations that require social intelligence and Theory of Mind in order to operate effectively (Wang et al., 2007; Dhelim et al., 2021; Langley et al., 2022). For example, new technologies are emerging where AI is used to *interact* and *adapt* to users (Bickmore and Picard, 2005; Jaques, 2019), e.g., voice assistants, and tutoring systems; or where AI helps *enhance communication* between multiple users, e.g., email autocomplete (Chen et al., 2019), AI-assisted counseling (Kearns et al., 2020; Allen, 2020; Sharma et al., 2021), or facilitated discussion (Rosé et al., 2014).

As we move beyond just asking single-turn questions to social and interactive AI assistants, higher-order reasoning becomes necessary (McDonald and Pearson, 2019). For example, AI systems should be capable of more nuanced understanding, such as ensuring an alarm is on if someone has a job interview the next morning (Dhelim et al., 2021), knowing to call for help when an elderly person falls (Pollack, 2005), inferring personality and intentions in dialogues (Mairesse et al., 2007; Wang et al., 2019), reasoning about public commitments (Asher and Lascarides, 2013), predicting

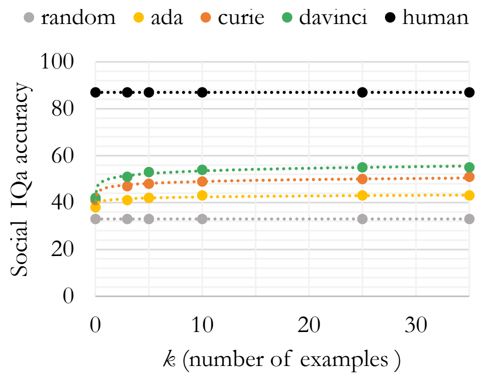


Figure 2: Accuracy on the SOCIALIQA dev. set, broken down by LLM model type and size, as well as number of few-shot examples (k).

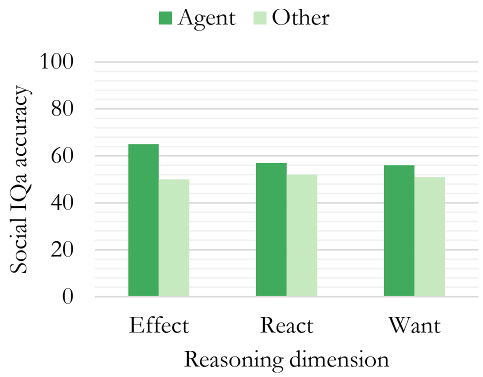


Figure 3: Comparing the accuracy of GPT-3-DAVINCI (35-shot) on SOCIALIQA when the reasoning is about the main agent of the situation versus others.

emotional and affective states (Litman and Forbes-Riley, 2004; Jaques et al., 2020), and incorporating empathy, interlocutor perspective, and social intelligence (Kearns et al., 2020; Sharma et al., 2021).

What is Theory of Mind? Theory of Mind (TOM) describes the ability that we, as humans, have to ascribe and infer the mental states of others, and to predict which likely actions they are going to take (Apperly, 2010).[1] This ability is closely related to (interpersonal) social intelligence (Ganaie and Mudasir, 2015), which allows us to navigate and understand social situations ranging from simple everyday interactions to complex negotiations (Gardner et al., 1995).

Interestingly, the development of Theory of Mind and language seem to happen around similar ages in children (Sperber and Wilson, 1986; Wellman, 1992; Miller, 2006; Tauzin and Gergely, 2018).[2] Theories of the pragmatics of language and communication can frame our understanding of this link (Rubio-Fernandez, 2021), positing that one needs to reason about an interlocutor's mental state (TOM) to effectively communicate and understand language (Grice, 1975; Fernández, 2013; Goodman and Frank, 2016; Enrici et al., 2019).[3]

[1]While Theory of Mind is well developed in most adults (Ganaie and Mudasir, 2015), reasoning and inference capabilities can be influenced by age, culture, neurodiversity, or developmental disorders (Korkmaz, 2011).

[2]The direction of the TOM-language association is still debated (de Villiers, 2007). Some researchers believe language development enables TOM-like abilities (Pyers and Senghas, 2009; Rubio-Fernandez, 2021). On the other hand, some argue that language develops after TOM since preverbal infants already could possess some level of TOM-like abilities (Onishi and Baillargeon, 2005; Southgate and Vernetti, 2014; Poulin-Dubois and Yott, 2018).

[3]Most cognitive studies on this subject focus on the English language, which is not representative of the wide variation of language structures, and thus limits the cognitive conclusions one can draw about the link between language and Theory of Mind (Blasi et al., 2022).

3 SOCIALIQA: Do LLMs have Social Intelligence and Social Commonsense?

A crucial component of Theory-of-Mind is the ability to reason about the intents and reactions of participants of social interactions. To measure this, we use the dev. set of the SOCIALIQA QA benchmark (Sap et al., 2019b), which was designed to probe social and emotional intelligence in various everyday situations. This benchmark covers questions about *nine* social reasoning dimensions, drawn from the ATOMIC knowledge graph (Sap et al., 2019a).

SOCIALIQA instances consist of a context, question, and three answer choices, written in English. Each question relates to a specific reasoning dimension from ATOMIC: six dimensions focus on the pre- and post-conditions of the *agent* or protagonist of the situation (e.g., needs, intents, reactions, next actions), and three dimensions focus on the post-conditions of *other* participants involved in the situation (reaction, next action, effect). In total, there are 1954 three-way QA tuples; see Tab. 1 for examples, and Tab. 3 in Appendix A for per-dimension counts.

3.1 Probing LLMs with SOCIALIQA

To probe our language models, we use a k-shot language probing setup, following Brown et al. (2020). We select the answer that has the highest likelihood under the language model conditioned on the context and question, as described in Appendix C.

To test the limits of what the models can do, we select k examples that have the same ATOMIC reasoning dimension as the question at hand, varying k

	Situation	Answers	Focus
(a)	Remy was working late in his office trying to catch up. He had a big stack of papers. What does Remy need to do before this?	✓ Needed to be behind Be more efficient 🤖 Finish his work	Agent
(b)	Casey wrapped Sasha's hands around him because they are in a romantic relationship. How would you describe Casey?	✓ Very loving towards Sasha 🤖 Wanted Being kept warm by Sasha	Agent
(c)	Tracy held a baby for 9 months and then gave birth to addison. What will happen to Tracy?	Throw her baby at the wall 🤖 Cry ✓ Take care of her baby	Agent
(d)	Kai gave Ash some bread so they could make a sandwich. How would Kai feel afterwards?	✓ Glad they helped Good they get something to eat 🤖 Appreciative	Agent
(e)	Aubrey was making extra money by babysitting Tracey's kids for the summer. What will Tracy want to do next?	🤖 Save up for a vacation ✓ Let Aubrey know that they are appreciated Pay off her college tuition	Others
(f)	The people bullied Sasha all her life. But Sasha got revenge on the people. What will the people want to do next?	Do whatever Sasha says 🤖 Get even ✓ Flee from Sasha	Others
(g)	After everyone finished their food they were going to go to a party so Kai decided to finish his food first. What will others want to do next?	✓ Eat their food quickly 🤖 Throw their food away Go back for a second serving	Others
(h)	Aubrey fed Tracy's kids lunch today when Tracy had to go to work. What will happen to Aubrey?	Be grateful ✓🤖 Get paid by Tracy Get yelled at by Tracy	Agent
(i)	Sasha was the most popular girl in school when she accepted Jordan's invitation to go on a date. What will Jordan want to do next?	Plan a best friends outing with Sasha ✓🤖 Plan a romantic evening with Sasha Go on a date with Valerie	Others

Table 1: Examples of SOCIALIQA questions, which person the questions focus on (*Agent*, *Others*), and the human gold answers (✓) and GPT-3-DAVINCI predictions (🤖).

from 0 to 35 in increments of 5. We use three GPT-3 model sizes: GPT-3-ADA (smallest), and GPT-3-CURIE and GPT-3-DAVINCI (two largest).

3.2 SOCIALIQA Results

Shown in Fig. 2, GPT-3 models perform substantially worse than humans (>30% less) on SOCIALIQA,[4] and also worse than models finetuned on the SOCIALIQA training set (>20%; Lourie et al., 2021).[5] Although it is not surprising that GPT-3-DAVINCI reaches higher accuracies than GPT-3-ADA and GPT-3-CURIE, the gains are small, which suggests that increasing model size might not be enough to reach human-level accuracy. These findings are in line with recent BIG-Bench results on SOCIALIQA with the BIG-G (128B parameters; Srivastava et al., 2022) and PaLM (353B parameters; Chowdhery et al., 2022) LLMs, which lag behind humans with 45% and 73% accuracy, respectively (see Fig. 7 in Appendix A.2).

Focusing on GPT-3-DAVINCI, while increasing the number of examples k improves performance, the differences are marginal after k=10 examples (only 1% increase from 10 to 35 examples). This suggest that performance either plateaus or follows a logarithmic relationship with increasing number of conditioning examples.

Finally, we examine the differences in GPT-3-DAVINCI with respect to which participant is the focus. Shown in Fig. 3, we find that GPT-3-DAVINCI performs consistently better on agent-centric questions, compared to other-oriented questions. Shown in the example predictions in Tab. 1, GPT-3-DAVINCI often confuses which participant is being asked about. In example (e), after Aubrey babysat for Tracy, GPT-3-DAVINCI fails to predict that Tracy will likely want to "*let Aubrey know they are appreciated*," and instead mistakenly predicts that Tracy will want to "*save up for vacation*," which is what Aubrey would likely do. GPT-3-

[4]We find similar results when using INSTRUCTGPT (Ouyang et al., 2022) instead of GPT-3-DAVINCI.

[5]Lourie et al. (2021) achieves 83% on the test set, as shown on the AI2 SOCIALIQA leaderboard.

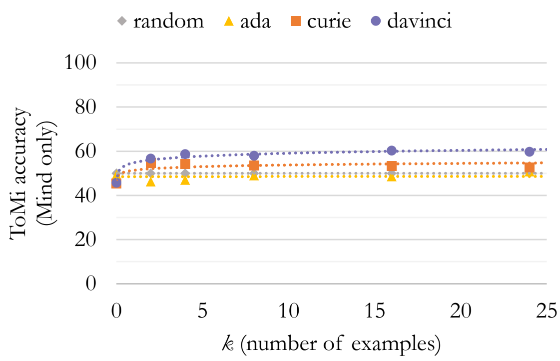


Figure 4: Accuracy on the ToMi dev. set Mind questions of varying sizes of GPT-3, and with varying number of examples (k).

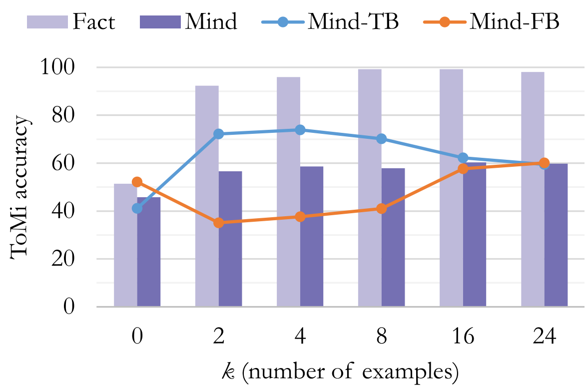


Figure 5: Accuracy of GPT-3-DaVinci by number of examples (k), by reasoning type (Fact vs. Mind; Mind-Tb vs. Mind-Fb).

DaVinci displays a similar participant confusion in example (f) in Tab. 1.

4 ToMi: Can LLMs Reason about Mental States and Realities?

Another key component of Theory of Mind is the ability to reason about mental states and realities of others, recognizing that they may be different than our own mental states. As a measure of this ability in humans, psychologists developed the *Sally Ann false-belief test* (Wimmer and Perner, 1983), in which two people (Sally and Ann) are together in a room with a ball, a basket, and a box, and while Sally is away, Ann moves the ball from the basket to the box. When asked where Sally will look for her ball, Theory of Mind allows us to infer that Sally will look in the basket (where she left the ball), instead of in the box (where the ball is, unbeknownst to Sally).

To measure the false-belief abilities of LLMs, we use the ToMi QA dataset of English Sally-Ann-like stories and questions (Le et al., 2019).[6] ToMi stories were created using a stochastic rule-based algorithm that samples two participants, an object of interest, and a set of locations or containers, and weaves together a story that involves an object being moved (see Tab. 2). All questions have two possible answers: the original object location, and the final object location.

We investigate how LLMs answer the ToMi story-question pairs, distinguishing between questions about factual object locations (Fact) and questions about where participants think objects are located (i.e., their mental states; Mind). The Fact questions either ask about the object's original (Fact-Mem) or final (Fact-Real) location. The Mind questions cover first-order (e.g., "*where will Abby look for the object?*"; Mind-1st) and second-order beliefs (e.g., "*where does James think that Abby will look for the object?*"; Mind-2nd). We further distinguish the Mind questions between true belief (Tb) and false belief (Fb), i.e., stories where a participant was present or absent when an object was moved, respectively.

Importantly, answering the Mind questions requires Theory of Mind and reasoning about realities and mental states of participants—regardless of the true- or false-belief setting—whereas Fact questions do not require such ToM. There are a total of 1861 two-way QA pairs in our ToMi probe set, with 519 Fact and 1342 Mind questions (see Tab. 4 in Appendix B for more detailed counts).

4.1 Probing LLMs with ToMi

We use the k-shot probing setup to test this ToM component in LLMs, with $k \in \{2, 4, 8, 16, 24\}$. We select k examples of the same reasoning type (i.e., Fact-Mem, Mind-1st, etc.), ensuring a 50-50 split between true- and false-belief examples for the Mind questions. As before, we test GPT-3-Ada, GPT-3-Curie, and GPT-3-DaVinci.

4.2 ToMi Results

Shown in Fig. 4, our results indicate that GPT-3 models struggle substantially with the ToMi questions related to mental states (Mind), reaching 60% accuracy in the best setup. As expected, the best performance is reached with GPT-3-DaVinci compared to smaller models which do not surpass

[6] ToMi is a more challenging version of the rule-based datasets by Nematzadeh et al. (2018) and Grant et al. (2017), as it contains randomly inserted distractor actions that prevent trivial reverse engineering.

	Type	Story	Question		Answers
(a)	FACT	Sophia entered the study. Noah entered the study. The dress is in the treasure chest. Noah exited the study. Hannah entered the garden. Sophia moved the dress to the box.	Where is the dress really?	✓🤖 	box treasure chest
(b)	M-1-FB	Noah entered the garden. Nathan entered the garden. Evelyn likes the pumpkin. The banana is in the basket. Nathan exited the garden. Noah moved the banana to the suitcase.	Where will Nathan look for the banana?	✓🤖 	basket suitcase
(c)	M-2-TB	Lily entered the patio. Aiden is in the patio. Mila entered the patio. Mila hates the radish. The coat is in the box. Aiden moved the coat to the crate. Mila exited the patio.	Where does Aiden think that Mila searches for the coat?	✓🤖 	crate box
(d)	M-1-TB	Elizabeth entered the cellar. Carter entered the cellar. The slippers is in the crate. Elizabeth moved the slippers to the container. Carter exited the cellar.	Where will Carter look for the slippers?	✓ 🤖	container crate
(e)	M-1-FB	Evelyn entered the living room. Jackson entered the playroom. James entered the playroom. The beans are in the treasure chest. James exited the playroom. Jackson moved the beans to the pantry. Jackson exited the playroom. James entered the living room.	Where will James look for the beans?	✓ 🤖	treasure chest pantry
(f)	M-2-FB	Isla likes the potato. Ella entered the laundry. Oliver entered the laundry. The slippers are in the box. Ella exited the laundry. Oliver moved the slippers to the basket. Isla entered the office.	Where does Ella think that Oliver searches for the slippers?	🤖 ✓	basket box

Table 2: Example stories in the TOMI dev. dataset, with GPT-3-DAVINCI predictions (with k=16 examples) and gold answers. "Type" denotes reasoning type, M-1 and M-2 denote MIND-1st and MIND-2nd, resp.

55% accuracy; however, as before, the gains from scaling up GPT-3 are very small. Similarly, increasing the number of few-shot examples beyond $k = 4$ does not substantially improve performance, corroborating findings on SOCIALIQA.

Further examining GPT-3-DAVINCI with respect to question types, we show that the model struggles substantially more with questions about mental states (55–60% for $k > 0$) compared to factual questions (90–100% for $k > 0$; Fig. 5; columns). Furthermore, the difference between performance on MIND-TB and MIND-FB questions shows an interesting pattern when conditioning on an increasing number of examples k (Fig. 5; lines): GPT-3-DAVINCI's MIND-TB accuracy first increases, peaks at $k = 4$, then decreases. This peak seems to be due to the model defaulting to the most recent object location (i.e., the correct MIND-TB answer), as illustrated in example (e) in Tab. 2. Apparent in Fig. 10 in Appendix B, this recency bias is a phenomenon that has been previously documented in LLMs (O'Connor and Andreas, 2021). In general, GPT-3-DAVINCI's comparably poor performance for MIND-TB and MIND-FB questions at $k > 8$ suggests that it cannot properly answer questions about participants' mental states and realities.

5 Discussion: Towards NLP with Neural Theory of Mind

Most humans develop social intelligence and Theory of Mind naturally. However, in this work, we showed that these abilities do not emerge automatically in large-pretrained language models. These shortcomings contrast with the wealth of successes of LLMs at a variety of tasks, including tasks that potentially require social intelligence. For example, GPT-3 has been shown to generate stories with emotional arcs that are virtually indistinguishable from human-written stories (Clark et al., 2021). Additionally, recent work has used GPT-3 to generate social commonsense knowledge related to protagonists of situations (West et al., 2022). While those findings suggest some level of social and emotional intelligence in LLMs, our explorations highlight the limits of these abilities, and raise the open question: *how can we create NLP systems with true social intelligence and Theory of Mind?*

To begin answering this question, we first discuss the current LLMs training paradigm (§5.1), drawing from theories of pragmatics to examine why these models are not learning social intelligence efficiently. Then, we outline some possible future directions to bias models towards Theory of Mind (§5.2), through person-centric neural architectures, data selection, and training objectives.

5.1 The Pragmatics of "Static" Text

To understand why LLMs are still struggling with social intelligence, we examine LLMs' training paradigm through the lens of *pragmatics*. As discussed in §2, pragmatics provides a connection between language development and Theory of Mind (Sperber and Wilson, 1986; Miller, 2006; Tauzin and Gergely, 2018): learning to communicate effectively with language requires reasoning about what our interlocutor knows or does not know (Grice, 1975; Fernández, 2013; Goodman and Frank, 2016; Enrici et al., 2019).[7]

One major use of language by people is to communicate about relationships and personal experiences (Clark and Schaefer, 1989; Dunbar, 1993). This is fundamentally different from the training data of LLMs, which consists of language found in what we call *static* texts: documents that are written for a general audience and are relatively self-contained and topically focused (e.g., news articles, books, Wikipedia articles; Gao et al., 2020; Dodge et al., 2021). Such static text is typically written such that readers only require the language itself as input, which they then combine with their world knowledge and commonsense to understand its meaning (Graesser et al., 1994).

If AI systems are to learn social intelligence and Theory of Mind, we posit that static text has certain limitations, from a pragmatics lens, outlined below.

Reporting bias. Following Grice's maxim of quantity (Grice, 1975), static text often avoids redundancy by omitting content that is known by both the author and the reader (Clark and Brennan, 1991). Also known as reporting bias (Gordon and Van Durme, 2013; Lucy and Gauthier, 2017), this phenomenon likely limits LLMs' ability to learn social commonsense knowledge from static text.

Lack of communicative intent and alternatives. A corollary to reporting bias, static text does not provide any direct access to *communicative intent* (why words were used) or to *alternatives* (which words were not used, and why). This reasoning about intents, alternatives, and their implications is highly predictive of the pragmatic inferences people draw about their interlocutors (Goodman and Frank, 2016) — for example, when someone answers *Where does Taylor live?* with *Somewhere in the U.S.*, it implies that they likely do not know or do not want to share the exact location, since, if they did, they would have been more specific. This poses a likely limitation that LLMs only learn what words are used, but not which words were not used, and why.

Lack of communicative effects. Language is primarily learned (Wells and Bridges, 1981; Tomasello et al., 2005) and used (Clark, 1996) in collaborative and interactive settings (Clark and Schaefer, 1989), which allow interlocutors to give immediate feedback to each other on whether their language was understood (Clark and Krych, 2004) or should be adjusted (Krauss and Weinheimer, 1966), and observe the perlocutionary effects that their language has on their partners (Austin, 1975). Since static text has no such feedback, LLMs learn from all texts, as if they were all equally understandable by readers.

Centering theory. At any given time, most text focuses on describing one protagonist and their relation to their surroundings, according to Centering Theory (Grosz et al., 1995). As such, main characters and their mental states are more likely to be *described*, whereas other participants might only be *mentioned*. Additionally, main characters or protagonists are more likely to be referred to with pronouns, whereas secondary characters with their names.

Thus, a model trained purely on static text might not learn to reason about social intelligence or mental states and realities of different characters of situations; they might not even inherently learn to resolve coreference for multiple characters (Sakaguchi et al., 2020). In fact, challenges of coreference resolution could explain why GPT-3 models struggle on SocialIQa which contains questions with pronouns, and centering theory and main character biases in static text could explain why models find non-protagonist questions more challenging. On the other hand, ToMi does not contain any pronouns, and thus requires social intelligence beyond coreference resolution.

5.2 Future directions towards LLMs with Theory of Mind

While there is no one best path towards LLMs with social intelligence and Theory of Mind, it seems

[7] Note here that, in contrast to other work (Bender and Koller, 2020; Bisk et al., 2020), we do not focus on whether LLMs "understand" language, instead we examine whether LLMs can answer questions about the emotions and mental states of participants of situations.

likely that progress will require challenging the standard paradigm of training on static text with the language modeling objective. Based on our findings and the limitations we discussed, we reflect on some possible directions forward.

Beyond static text as training data? Perhaps the key is in the data: the knowledge contained in static text might be too limited for models to learn social intelligence, for reasons described in §5.1 Socially grounded text (containing elaborations of communicative intents, character mental states, speaker identities, etc.) could enable more efficient learning of Theory of Mind abilities (Bender and Koller, 2020; Bisk et al., 2020; Hovy and Yang, 2021), similar to how visual groundings can help with learning physical knowledge (Zhang et al., 2022a). Examples of such datasets include "Social Stories," which are devised to help individuals with autism improve their interpersonal skills (Gray, 1995), or the Story Commonsense (Rashkin et al., 2018) and GLUCOSE (Mostafazadeh et al., 2020) commonsense-annotated story datasets. Alternatively, perhaps interactional texts, such as dialogues and other datasets that were explicitly created to require reasoning about mental states, could help with neural Theory of Mind (Bara et al., 2021).

Nevertheless, the scale of training datasets seems to be crucial for LLMs (Kaplan et al., 2020; Chowdhery et al., 2022), which poses a challenge: text datasets rich in social intelligence and interactions are not easily found naturally due to reporting biases, and they are costly to create (Rashkin et al., 2018; Mostafazadeh et al., 2020). Promising results on commonsense reasoning suggest a possible hybrid approach: LLMs could be jointly or sequentially trained on static text and commonsense knowledge bases or socially grounded or interactional text (Bosselut et al., 2019; Hwang et al., 2021), first trained on static text and then enhanced for commonsense knowledge via reinforcement learning (Zhou et al., 2021).

Person-centric neural inductive biases? While more socially grounded training data could help, LLMs might also learn social intelligence better if they are designed with person-centric inductive biases and training objectives. Hinting at this, prior work has shown that training entity-centric neural architectures on text with entity coreference information yields more entity-aware LLMs, both in recurrent (Henaff et al., 2017; Ji et al., 2017; Yang et al., 2017; Liu et al., 2019) and Transformer-based models (Févry et al., 2020; De Cao et al., 2020; Rosset et al., 2020; Zhang et al., 2022c).

However, Theory of Mind and social intelligence require much richer social grounding than coreference chains, which is challenging to obtain for supervised settings, especially at the scale that LLMs require. Thus, unsupervised approaches to adding inductive biases to models could be a promising solution. Future work could look to cognitive science and neuroscience research for possible directions (Langley et al., 2022), such as exploring LLMs' equivalents of human concept cells (i.e., sets of neurons that activate for important people or concepts; Bowers, 2017; Calvo Tapia et al., 2020).

Alternatively, examining the internal or latent representations of LLMs could point to future directions towards inductive biases for neural Theory of Mind. As an example, recent work has found evidence of latent representations of grounded semantics in models trained only on static text (Li et al., 2021), which can be tied to real-world grounding with a small amount of additional supervised training (Patel and Pavlick, 2022). Future work might similarly analyze deep learning models for representations of Theory of Mind, toward augmenting the models with structure or objectives that surface and strengthen these representations.

Interactive and experiential grounding? It is possible, nevertheless, that socially grounded data and person-centric inductive biases will not suffice. Some researchers have argued that language understanding could only emerge from interactions and experiences (Bender and Koller, 2020; Bisk et al., 2020). Likely, this applies to Theory of Mind and social intelligence as well, due to lack of communicative intents and alternatives in static text. Future work could explore approaches grounded more explicitly in interaction, intents, and alternatives, e.g., by explicitly predicting possible next steps and learning why predictions were wrong. In fact, promising research has shown that using an interactive learning or multi-agent communication paradigm can enable some Theory of Mind capabilities of models (Hawkins et al., 2019; Lazaridou et al., 2020; Zhu et al., 2021; Wang et al., 2022).

However, there are limits to the types of Theory of Mind that can be learned from interactive simulations, which are often task-specific (e.g., describing objects in an image; Lazaridou et al., 2020; Steinert-Threlkeld et al., 2022). Furthermore, models that were trained in interactive simulation settings often struggle to generalize beyond the simulation environment (Ludwin-Peery et al., 2021; Mu and Goodman, 2021). Based on promising results by Lazaridou et al. (2020); Zhu et al. (2021), future work might create generalizable LLMs with neural Theory of Mind through hybrid approaches that combine pretraining with interactive learning: updating models trained on static text using supervision either from humans (Stiennon et al., 2020; Ouyang et al., 2022; Scheurer et al., 2022) or from proxies for human behavior or social environments (Ammanabrolu et al., 2022a,b) based on broad coverage LLMs (Perez et al., 2022).

Probing and evaluating ToM While neural Theory of Mind and social intelligence may remain an elusive goal for some time, developing measures of those abilities in systems can be done in tandem. We encourage further research in developing benchmarks that measure specific social abilities in LLMs (e.g., Sap et al., 2019b; Zadeh et al., 2019), especially those that minimize annotation artifacts and spurious correlations (Schwartz et al., 2017; Gururangan et al., 2018; Le et al., 2019). Additionally, we encourage further investigations into probing the latent knowledge within LLMs (Tenney et al., 2019; Li et al., 2021) or examining how LLMs handle entities and people (Onoe et al., 2022; Schuster and Linzen, 2022), which could shed light onto better data choices and inductive biases towards neural Theory of Mind and social intelligence.

6 Conclusion

We explore the open question of whether and how much modern large-scale language models (LLMs) can reason about social intelligence and Theory of Mind. Our results show that out-of-the-box LLMs struggle substantially with these abilities, which we argue are necessary but not sufficient aspects of Theory of Mind. Specifically, GPT-3's social intelligence as measured by SocialIQa lags behind humans (>30%), and the model struggles to answer ToMi questions about mental states (55-60%) compared to factual questions (90–100%). In light of these shortcomings, we critically examine the large language model pretraining paradigm from a pragmatics-based perspective, and discuss possible directions towards enabling true social intelligence in NLP systems.

We make our preprocessed datasets available at `http://maartensap.com/neuralToM`.

7 Limitations

Our work focuses on investigating the Theory of Mind abilities in large pretrained language models, but we focus on accessing GPT-3 (Brown et al., 2020) through an API, since we do not have access to some of the larger models out there (PaLM; Chowdhery et al., 2022) nor do we have the computational resources to run an open-source version of GPT-3 (OPT; Zhang et al., 2022b). We hypothesize that results would not be drastically different with such models, based on the low accuracy displayed on SocialIQa in the recently released BIG-Bench experiments (Srivastava et al., 2022). Nevertheless, we hope developers of larger LLMs will investigate these ToM abilities to confirm or refute our findings.

We measure the ability to answer questions about people's mental states using ToMi, which is an automatically constructed corpus of stories involving people, objects, and locations. The automatic nature of the creation process could induce biases and artifacts, such as objects being in locations that are plausible but not typical (e.g., bananas in a closet), which could influence model's ability to answer questions properly. Based on the near-perfect accuracy on the factual questions, however, this may not be a significant issue. Future work should investigate more naturalistic settings to probe this ability in LLMs.

A potential limitation of our work is that models could latch onto surface patterns and spurious correlations in our two datasets. For example, theoretically, a model prompted with many ToMi examples may be able to reverse-engineer the data creation algorithm to find the solution to each question. However, this would be a bigger limitation if our claims were that LLMs *do* have social intelligence and Theory of Mind; instead, given that our results show low performance on these tasks even though they are potentially easier due to correlational patterns, this would indicate that LLMs have potentially even less reasoning abilities.

Additionally, while we operationalize our measure of social intelligence and Theory of Mind through two specific tasks, SocialIQa and ToMi, these abilities are much broader. As noted earlier, we view these benchmarks as necessary but not sufficient conditions for LLMs to have ToM; solving the benchmarks does not imply that LLMs have ToM, but LLMs with ToM should be able to solve them. We hope that future research will further investigate other aspects of Theory of Mind abilities in large pretrained LMs, drawing on social science research. For example, future work could make use of the "unexpected content" task (Gopnik and Astington, 1988) or the "George Washington University Social Intelligence Test" (Hunt, 1928) to measure the social intelligence of LLMs.

Finally, the focus on English language LLMs and benchmarks for Theory of Mind is another limitation of our work. Echoing recent cognitive science work that argues the need for non-English cognitive science investigations (Blasi et al., 2022). Specifically, false-belief abilities are greatly influenced by language structure and grammar (Boeg Thomsen et al., 2021; Zhang and Zhou, 2022).

Broader Sociotechnical Implications

AI systems are part of a broader sociotechnical system that also involves individual motivations and societal norms (Johnson and Verdicchio, 2017). As such, per a contextualist view of AI (instead of utopian or dystopian; Barbour, 1992), we envision AI systems with social intelligence and Theory of Mind being used in ways that enhance human's lives, autonomy, and agency (Chan, 2022). In parallel, we strongly support the development and research of policy and regulation, to prevent misuses of AI with social intelligence (Wischmeyer and Rademacher, 2020; Crawford, 2021; Reich et al., 2021).

Acknowledgements

We would like to thank Jack Hessel, Rowan Zellers, Jena D. Hwang, Prithviraj Ammanabrolu for their feedback on preliminary versions of this work, and Anna Jafarpour and Noah Goodman for fruitful cognitive science discussions about the research. We also thank the anonymous reviewers for their thoughtful comments. This research was supported by the Allen Institute for AI and the DARPA MCS program through NIWC Pacific (N66001-19-2-4031).

References

Summer Allen. 2020. Artificial intelligence and the future of psychiatry. *IEEE pulse*, 11(3):2–6.

Prithviraj Ammanabrolu, Renee Jia, and Mark Riedl. 2022a. Situated dialogue learning through procedural environment generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8099–8116, Dublin, Ireland. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022b. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6.

John Langshaw Austin. 1975. *How to Do Things with Words*. Clarendon Press.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *EMNLP*.

Ian G Barbour. 1992. Ethics in an age of technology. *The Gifford lectures*.

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

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of ACL*.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *EMNLP*, pages 8718–8735, Online. Association for Computational Linguistics.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Overreliance on english hinders cognitive science. *Trends Cogn. Sci.*

Ditte Boeg Thomsen, Anna Theakston, Birsu Kandemirci, and Silke Brandt. 2021. Do complement clauses really support false-belief reasoning? a longitudinal study with english-speaking 2- to 3-year-olds. *Dev. Psychol.*, 57(8):1210–1227.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Jeffrey S Bowers. 2017. Grandmother cells and localist representations: a review of current thinking. *Language, Cognition and Neuroscience*, 32(3):257–273.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are Few-Shot learners. In *NeurIPS*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Carlos Calvo Tapia, Ivan Tyukin, and Valeri A Makarov. 2020. Universal principles justify the existence of concept cells. *Scientific reports*, 10(1):7889.

Anastasia Chan. 2022. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. *AI and Ethics*.

Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-Time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2287–2295, New York, NY, USA. Association for Computing Machinery.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won, Chung Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard Guy, Gur-Ari Pengcheng, Yin Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. Technical report, Google.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *ACL*.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *J. Mem. Lang.*, 50(1):62–81.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Kate Crawford. 2021. *Atlas of AI*. Yale University Press.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *ICLR*.

Jill de Villiers. 2007. The interface of language and theory of mind. *Lingua. International review of general linguistics. Revue internationale de linguistique generale*, 117(11):1858–1878.

Sahraoui Dhelim, Huansheng Ning, Fadi Farha, Liming Chen, Luigi Atzori, and Mahmoud Daneshmand. 2021. IoT-Enabled social relationships meet artificial social intelligence. *IEEE Internet of Things Journal*, 8(24):17817–17828.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *EMNLP*.

R I M Dunbar. 1993. Coevolution of neocortical size, group size and language in humans. *The Behavioral and brain sciences*, 16(4):681–694.

Katherine Elkins and Jon Chun. 2020. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 1(1):17212.

Ivan Enrici, Bruno G Bara, and Mauro Adenzato. 2019. Theory of mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature ofhuman communication. *Pragmatics & Cognition*, 26(1):5–38.

Camila Fernández. 2013. Mindful storytellers: Emerging pragmatics and theory of mind development. *First language*, 33(1):20–46.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *EMNLP*.

MY Ganaie and Hafiz Mudasir. 2015. A Study of Social Intelligence & Academic Achievement of College Students of District Srinagar, J&K, India. *Journal of American Science*, 11(3):23–27.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Howard Gardner, Robert M Hanson, and Steve Hamilton. 1995. *How Are Kids SMART?: Multiple Intelligences in the Classroom*. National Professional Resources, Incorporated.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, 20(11):818–829.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

A C Graesser, M Singer, and T Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371–395.

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

Carol A Gray. 1995. Teaching children with autism to "read" social situations. *Teaching children with autism: Strategies to enhance communication and socialization*, pages 219–242.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

David Gunning. 2018. Machine common sense concept paper.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Robert D Hawkins, Minae Kwon, Dorsa Sadigh, and Noah D Goodman. 2019. Continual adaptation for efficient machine communication. In *CoNLL*.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *ICLR*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *NAACL*. Association for Computational Linguistics.

Thelma Hunt. 1928. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (COMET-)ATOMIC2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003, Online. Association for Computational Linguistics.

Natasha M Jaques. 2019. *Social and affective machine learning*. Ph.D. thesis, Massachusetts Institute of Technology.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A Smith. 2017. Dynamic entity representations in neural language models. In *EMNLP*, pages 1831–1840, Copenhagen, Denmark. Association for Computational Linguistics.

Deborah G Johnson and Mario Verdicchio. 2017. Reframing AI discourse. *Minds and Machines*, 27(4):575–590.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

William R Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. A Wizard-of-Oz interface and persona-based methodology for collecting health counseling dialog. In *CHI Extended Abstracts*, CHI EA '20, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Baris Korkmaz. 2011. Theory of mind and neurodevelopmental disorders of childhood. *Pediatr Res*, 69(5 Pt 2):101R–8R.

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

R M Krauss and S Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *J. Pers. Soc. Psychol.*, 4(3):343–346.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence and Applications*, 5.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

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.

Alessandro Lenci. 2023. Understanding natural language understanding systems. a critical analysis.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *ACL*, pages 1813–1827.

Diane J. Litman and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 351–358, Barcelona, Spain.

Fei Liu, Luke Zettlemoyer, and Jacob Eisenstein. 2019. The referential reader: A recurrent entity network for anaphora resolution. In *ACL*, pages 5918–5925, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *AAAI*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *RoboNLP@ACL*.

Ethan Ludwin-Peery, Neil R Bramley, Ernest Davis, and Todd M Gureckis. 2021. Limits on simulation approaches in intuitive physics. *Cognitive psychology*, 127:101396.

F Mairesse, M A Walker, M R Mehl, and R K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Gary Marcus. 2022. Deep learning is hitting a wall. *Nautilus*.

Gary Marcus and Ernest Davis. 2023. How not to test GPT-3.

Kelsey R McDonald and John M Pearson. 2019. Cognitive bots and algorithmic humans: toward a shared understanding of social intelligence. *Current Opinion in Behavioral Sciences*, 29:55–62.

Carol A Miller. 2006. Developmental relationships between language and theory of mind. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 15(2):142–154.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. Glucose: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586.

Jesse Mu and Noah Goodman. 2021. Emergent communication of generalizations. In *NeurIPS*.

Sharan Narang and Aakanksha Chowdhery. 2022. Pathways language model (PaLM): Scaling to 540 billion parameters for breakthrough performance. `https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html`. Accessed: 2022-5-5.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. 2018. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

851–864, Online. Association for Computational Linguistics.

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

Yasumasa Onoe, Michael J Q Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. In *Findings of NAACL*.

OpenAI. 2023. GPT-4 technical report. Technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models.

Martha E Pollack. 2005. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI magazine*, 26(2):9–9.

Diane Poulin-Dubois and Jessica Yott. 2018. Probing the depth of infants' theory of mind: disunity in performance across paradigms. *Developmental science*, 21(4):e12600.

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

Jennie E Pyers and Ann Senghas. 2009. Language promotes false-belief understanding: evidence from learners of a new sign language. *Psychol. Sci.*, 20(7):805–812.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *ACL*.

Rob Reich, Mehran Sahami, and Jeremy M Weinstein. 2021. *System error: Where big tech went wrong and how we can reboot*. Hodder & Stoughton.

Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. 2014. Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning @ Scale*, L@S '14, pages 197–198, New York, NY, USA. Association for Computing Machinery.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-Aware language model pretraining. In *NeurIPS*.

Paula Rubio-Fernandez. 2021. Pragmatic markers: the missing link between language and theory of mind. *Synthese*, 199(1):1125–1158.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*. Data downloaded from `http://maartensap.com/social-iqa/data/socialIQa_v1.4_withDims.tgz`.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback. In *ACL Workshop on Learning with Natural Language Supervision. 2022.*

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *NAACL*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Victoria Southgate and Angelina Vernetti. 2014. Belief-based action prediction in preverbal infants. *Cognition*, 130(1):1–10.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Aarohi Srivastava, Abhinav Rastogi, Abhishek B 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, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew D. La, Andrew Kyle 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 Karakacs, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C'esar Ferri Ram'irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Gonz'alez, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. 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, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng' 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, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, M Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, M'aty'as Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pi-Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ram'on Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Lebras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Sameh Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Kumar Reddy, Sneha Priscilla Makini, Soo hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko,

Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O. Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, W Vossen, Xiang Ren, Xiaoyu F Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yuntao Bai, Zachary Seid, Zhao Xinran, Zhuoye Zhao, Zi Fu Wang, Zijie Jay Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Shane Steinert-Threlkeld, Xuhui Zhou, Zeyu Liu, and CM Downey. 2022. Emergent communication finetuning (ec-ft) for pretrained language models. In *Emergent Communication Workshop at ICLR 2022*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Tibor Tauzin and György Gergely. 2018. Communicative mind-reading in preverbal infants. *Scientific reports*, 8(1):9534.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *ACL*.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–91; discussion 691–735.

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

Annalisa Valle, Davide Massaro, Ilaria Castelli, and Antonella Marchetti. 2015. Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *European journal of psychological assessment: official organ of the European Association of Psychological Assessment*, 11(1):112–124.

Fei-Yue Wang, Kathleen M Carley, Daniel Zeng, and Wenji Mao. 2007. Social computing: From social informatics to social intelligence. *IEEE intelligent systems*, 22(2):79–83.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. 2022. ToM2C: Target-oriented multi-agent communication and cooperation with theory of mind. In *ICLR*. arxiv.org.

Henry M Wellman. 1992. The child's theory of mind. *The MIT Press series in learning, development, and conceptual change.*, 358.

Gordon Wells and Allayne Bridges. 1981. *Learning through interaction: volume 1: the study of language development*, volume 1. Cambridge University Press.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *NAACL*.

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.

Thomas Wischmeyer and Timo Rademacher, editors. 2020. *Regulating Artificial Intelligence*. Springer, Cham.

Zichao Yang, Phil Blunsom, Chris Dyer, and Wang Ling. 2017. Reference-Aware language models. In *EMNLP*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. Visual commonsense in pretrained unimodal and multimodal models. In *NAACL 2022*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xiaowen Zhang and Peng Zhou. 2022. Linguistic cues facilitate children's understanding of belief-reporting sentences. *First Lang.*, 42(1):51–80.

Zhihan Zhang, Wenhao Yu, Chenguang Zhu, and Meng Jiang. 2022c. A unified Encoder-Decoder framework with entity memory. In *EMNLP*.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Bill Yuchen Lin, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. *ICLR*, abs/2011.07956.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12901–12911. PMLR.

dimension	definition	count
	Before the event	
xIntent	Why does X cause the event	238
xNeed	What does X need to do before the event	228
xAttr	How would X be described?	287
	After the event	
Effect		*218*
xEffect	What effects does the event have on X?	99
oEffect	What effects does the event have on others?	119
React		*415*
xReact	How does X feel after the event?	223
oReact	How do others feel after the event?	192
Want		*568*
xWant	What would X likely want to do after the event?	338
oWant	What would others likely want to do after the event?	230
total		**1954**

Table 3: SOCIALIQA dev. set statistics, broken down by question reasoning type and their definitions from ATOMIC.

A SOCIALIQA Details

A.1 Data Preprocessing

We downloaded the SOCIALIQA training and dev. datasets from the publicly available SOCIALIQA website.[8] This version of the SOCIALIQA dataset contains the original ATOMIC dimensions that workers were prompted with to create a question, as well as the correspondence between questions and which character they focus on (agent or other). To ensure consistency, for each context, question, and answer, we normalize the casing to start with a capital letter if the text does not already.

A.2 Further SOCIALIQA results

In addition to results discussed in §3.2, we report further SOCIALIQA results here.

SOCIALIQA broken down by reasoning dimension. We break down the best performing GPT-3-DAVINCI (35-shot) setup by reasoning dimension. Shown in Fig. 6, we find that GPT-3-DAVINCI struggles most with questions related to what people needed to do before a situation could take place (Need). Conversely, questions related to a situation's agent's intent (Intent) and the effect of the situation on the agent (Effect) are seemingly easier for GPT-3-DAVINCI. Future work should explore

[8] http://maartensap.com/social-iqa/data/socialIQa_v1.4_withDims.tgz

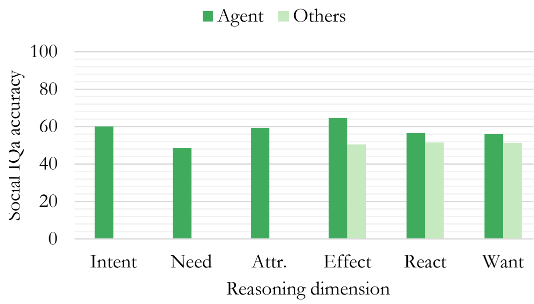


Figure 6: Comparing the accuracy of GPT-3-DAVINCI (35-shot) when over all nine reasoning dimensions.

LLMs's reasoning abilities along each of these dimensions in further detail.

BIG-Bench and PaLM results on SOCIALIQA. To further corroborate that LLMs struggle with SOCIALIQA, we show the performance of the non-publicly available BIG-G (Srivastava et al., 2022) and PaLM (Chowdhery et al., 2022) LLMs, along with the GPT-3 models, in Fig. 7. Both models are proprietary LLMs developed and tested on the 200+ datasets in BIG-Bench by Google / DeepMind.

While they are not discussed in the main BIG-Bench paper, the SOCIALIQA results for few-shot settings up to k=3 for BIG-G and k=5 for PaLM can be found on the BIG-Bench github website (accessed on 2022-11-10). Plotted in Fig. 7, both the BIG-G and PaLM LLMs lag behind humans with 45% and 73% peak accuracy, respectively.

B TOMI Details

B.1 Data Preprocessing

We generated TOMI stories using the github repository provided by Le et al. (2019). The code generated 5994 training and 5994 dev. stories. From those, we removed the story-question pairs which wrongly answered TOM-requiring questions from an omniscient perspective (i.e., answered MIND-FB questions from an omniscient perspective instead of the perspective of the character) which we noticed upon manual data inspection.[9] After this filtering, 5190 training and 5170 dev. stories remained.

For the final TOMI dev. set, we used stratified sampling to obtain similar numbers of story-question pairs for all types (FACT-REAL, FACT-MEM, MIND-1st-FB, MIND-1st-TB, MIND-2nd-FB and MIND-2nd-TB). The exact counts are

[9] We do not know why these datapoints were generated.

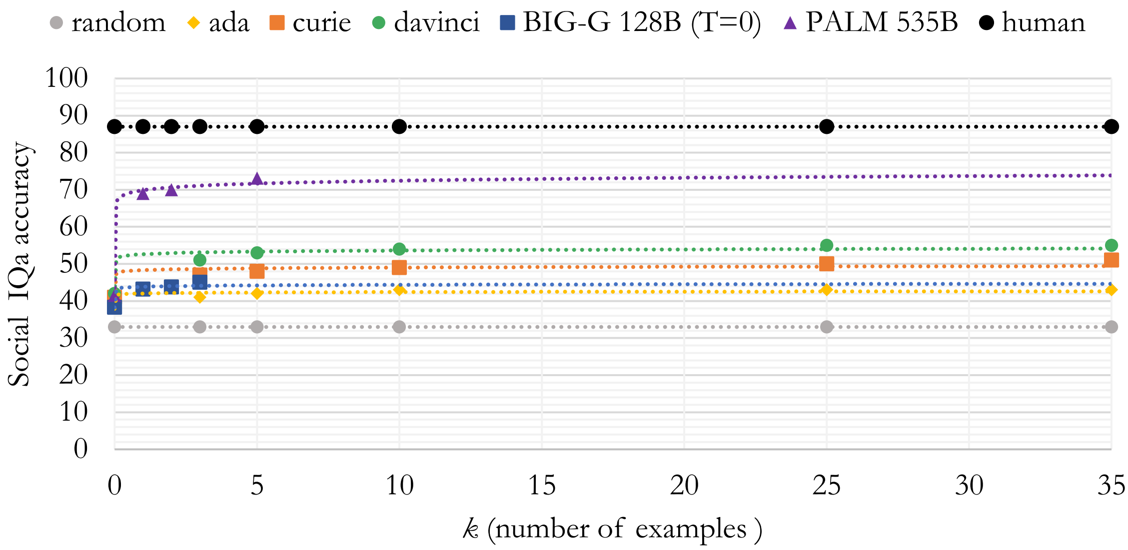


Figure 7: Expanded version of Fig. 2, depicting the accuracy on the SOCIALIQA dev. set, broken down by LLM model type and size, as well as number of few-shot examples (k). Here, we also include the accuracy results of the PaLM (Chowdhery et al., 2022) and BIG-G (Srivastava et al., 2022) LLMs, taken from the `BIG-Bench github` repository on 2022-11-10.

FACT	519
FACT-MEM	278
FACT-REAL	241
MIND	1342
MIND-TB	778
MIND-1st-TB	389
MIND-2nd-TB	389
MIND-FB	564
MIND-1st-FB	231
MIND-2nd-FB	333
total	1861

Table 4: TOMI dev. set statistics, broken down by question reasoning type.

shown in Tab. 4. We release our final preprocessed TOMI dev. dataset at `http://maartensap.com/neuralToM/ToMi-finalNeuralTOM.csv`

B.2 Further TOMI results

Shown in Fig. 8-10, we provide additional results to supplement those in §4.2.

Performance by model size, number of examples, and MIND versus FACT. In Fig. 8, we show the different accuracies that GPT-3 models of various sizes, prompted with various number of examples, for TOMI MIND and FACT questions. This plot shows the same accuracies as Fig. 4, with the addition of the FACT accuracies. These results show that in the few-shot prompting setup, GPT-3-CURIE and GPT-3-DAVINCI can achieve near perfect performance on factual questions about object locations (FACT), but struggle substantially more on questions related to mental states (MIND). Surprisingly, GPT-3-ADA struggles with both factual and mental state questions, possibly due to its smaller size.

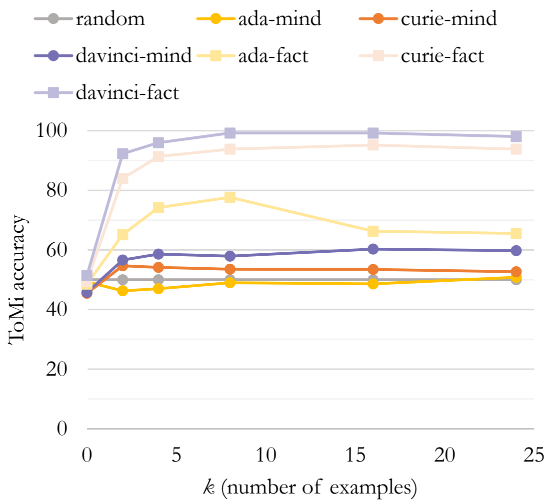


Figure 8: Examining the accuracy of GPT-3 of different sizes with different number of few-shot examples (k) on TOMI-MIND vs. TOMI-FACT questions.

Performance by question order. In Fig. 9, we break the GPT-3-DAVINCI performance down by TOM order (i.e., MIND-1st, MIND-2nd). Results show that with a number of examples between

2 and 16, GPT-3-DAVINCI performs better on MIND-1st questions (e.g., "Where will Sally look for the ball?") and struggles more with MIND-2nd questions (e.g., "Where does Ann think that Sally will look for the ball?"). This difference is somewhat diminished but still present for k=24 few-shot examples. These results somewhat mirror how humans struggle with increasingly higher-order ToM questions (Valle et al., 2015).

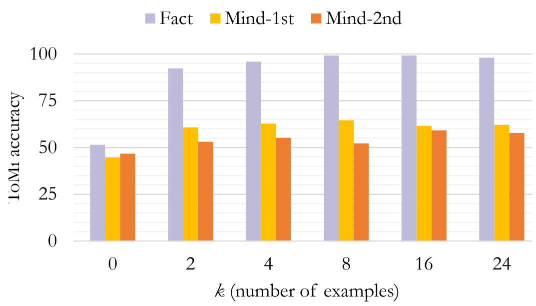


Figure 9: Comparing the accuracy of GPT-3-DAVINCI by the question reasoning type, specifically FACT vs. MIND-1st vs. MIND-2nd.

Recency bias in predictions. We further examine the results from §4.2, looking at GPT-3-DAVINCI's rate of predicting the location where the object was moved to (i.e., FACT-REAL). Shown in Fig. 10, GPT-3-DAVINCI accurately learns to almost always predict the last object location for FACT-FACT-REAL questions, and almost never for FACT-FACT-MEM locations.

Interestingly, the rates of selecting the last object location for MIND questions follows a concave pattern. This helps shed light onto the concave accuracy pattern seen in Fig. 5 for MIND-TB (and convex pattern for MIND-FB). Likely, in the few-shot setting with $2 < k < 8$, GPT-3-DAVINCI defaults to the most recently mentioned object location due to recency bias, which has been previously documented in LLMs (O'Connor and Andreas, 2021).

C GPT-3 Access and Probing Details

To probe our language models, we use a k-shot language probing setup, following Brown et al. (2020). Specifically, we concatenate the context (c) and question (q) together with proper punctuation, and assign the model prediction to the answer (a_i, $i \in 1, 2, 3$) with the highest conditional likelihood under the language model: $\arg\max_i p_{\text{LM}}(a_i \mid c, q, \mathcal{C}_k)$ where $\mathcal{C}_k$ denotes the k training examples, for which we provide the context, question, and correct answer concatenated. Note that we explored various probing setups and formats, such as QA-oriented formats and normalizing by marginal likelihood of each answer $p_{\text{LM}}(a)$ (as also explored in Brown et al., 2020), but found very little difference in performance.

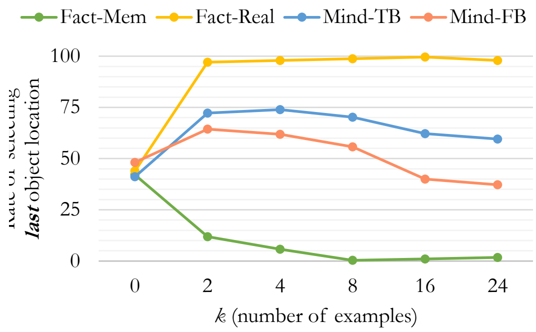


Figure 10: We plot the proportion of examples for which GPT-3-DAVINCI selects the last object location (i.e., in "reality").

We access GPT-3 through the OpenAI API.

model	OpenAI endpoint	Probing type	Description
GPT-3-DAVINCI	`davinci`	LM, MC	GPT-3 model, pretrained solely on the language modeling objective on large amounts of web text (Brown et al., 2020).
GPT-3.5-IFT	`text-davinci-002`	LM, MC	GPT-3 based model, finetuned to follow instructions and generate "high-quality" text using supervised finetuning (Ouyang et al., 2022).
GPT-3.5-RLHF	`text-davinci-003`	LM, MC	Similar to GPT-3.5-IFT, but finetuned with reinforcement learning from human feedback (RLHF) instead of supervised finetuning (Ouyang et al., 2022).
GPT-3.5-Turbo	`gpt-3.5-turbo-0301`	MC	Version of GPT-3.5-RLHF that is optimized for chat at 1/10th the computational cost (snapshot model from March 1st 2023) (Ouyang et al., 2022).
GPT-4	`gpt-4-0314`	MC	Newest model from OpenAI, claimed to be more capable than previous models (snapshot model from March 14th 2023) (OpenAI, 2023).

Table 5: List of new OpenAI models and APIs that we query with our SOCIALIQA and TOMI tasks. Probing type denotes whether the API allows for scoring candidate answers independently (LM) or if it requires showing candidate answers in multiple choice format (MC).

D What About ChatGPT or GPT-4? Effect of Instruction-tuning & RLFH

Our analyses in this paper have focused on large language models that are simply trained on the language modeling objective (e.g., GPT-2, GPT-3). However, in recent years, many improvements in LLMs have come from *instruction-finetuning* (IFT) and reinforcement learning from human feedback (RLHF), which are the key to the success of ChatGPT (or GPT-3.5; Ouyang et al., 2022) and GPT-4 (OpenAI, 2023). Despite the opacity of how these models were trained,[10] the question of whether they have better neural Theory of Mind than LM-objective-only LLMs is still of interest.

We quantify the performance of the newer set of OpenAI models (listed in Table 5) on a randomly selected 400 examples from the SOCIALIQA and TOMI datasets. Since these models have been increasingly capable, we focus on examining only their performance in zero-shot settings, as a stress-test of their TOM abilities. The newer APIs (GPT-3.5-Turbo, GPT-4) no longer provide the ability to score sequences, and thus, prevent the language modeling probing setup used in §3 and §4. As such, we also compare two types of probing: language modeling as described in §C (*LM probing*), and multiple-choice probing, where we provide the two or three answer candidates prepended with letter choices (A, B, C) and prompt the model to generate the letter for the correct answer (*MC probing*).[11]

D.1 SOCIALIQA: Social Intelligence and Social Commonsense

Shown in Figure 11, our results show that instruction-tuning and RLHF do indeed improve performance on zero-shot social and emotional intelligence question answering in SOCIALIQA. With LM probing, supervised instruction-finetuning (GPT-3.5-IFT: 53%) improves performance over vanilla language modelling (GPT-3-DAVINCI: 45%) slightly less than reinforcement learning (GPT-3.5-RLHF: 55%) . Notably, multiple-choice probing only improves over random chance (33%) with RLHF models (with 60, 67, and 79% for GPT-3.5-RLHF, GPT-3.5-Turbo, and GPT-4, respectively).

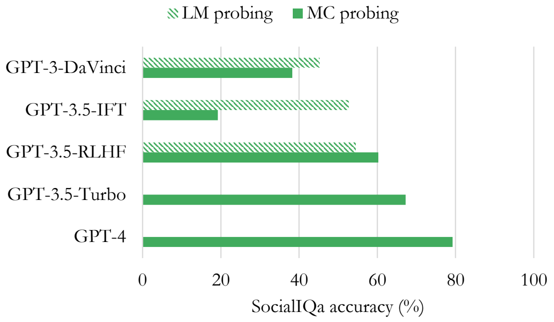
Focusing on the newest models, GPT-3.5-Turbo reaches 67.2% performance, still >20% below human performance reported in (Sap et al., 2019b). However, surprisingly, GPT-4 performance increases to 79.3%, within 10% of human performance on SOCIALIQA.

Interestingly, however, all models and probing setups perform worse on questions pertaining to non-main characters (Figure 11b), with GPT-4 showing a 7% decrease in accuracy. This suggests that reporting biases due to centering theory, as discussed in §5, may still play a role even for these

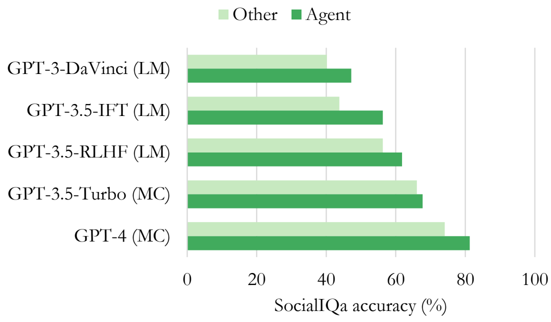
[10] OpenAI has stated that they will not be releasing any useful details. In their system report for GPT-4, they state: "*Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar*" (OpenAI, 2023).

[11] For example, for SOCIALIQA, we prepend the example with the instructions "`You are a multiple-choice answering system that responds with either A, B, or C.`"

New Content Added 2023-03-30

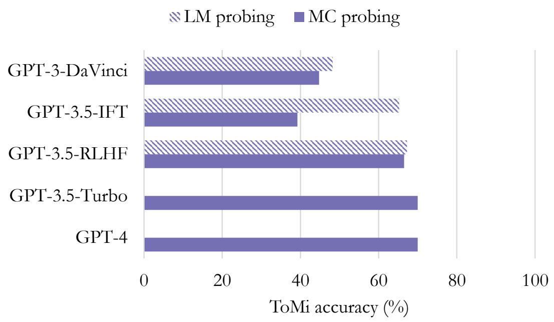


(a) SOCIALIQA accuracy with different probing methods.

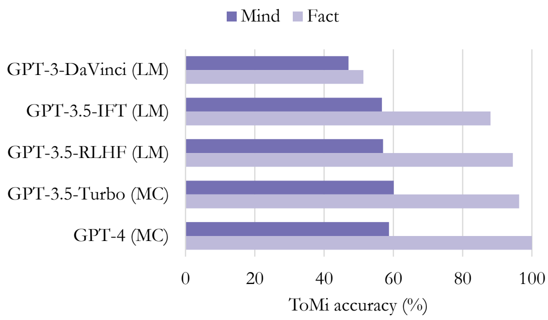


(b) SOCIALIQA accuracy broken down by character.

Figure 11: SOCIALIQA results for new instruction-tuned and reinforcement learning LLMs, broken down by probing type (11a) and by focus character of the question (11b), on a random subset of 400 examples.



(a) TOMI accuracy with different probing methods.



(b) TOMI accuracy broken down by question type.

Figure 12: TOMI results for new instruction-tuned and reinforcement learning LLMs, broken down by probing type (12a) and by question type (12b), on a random subset of 400 examples.

extreme scale models.

Caveat: it is increasingly likely that models were trained on benchmark data (either due to data collection via the online GUI or API, or by scraping web text without filtering). Indeed, the GPT-4 system report notes that some BIG-Bench test data (Srivastava et al., 2022), which contains the SOCIALIQA dev. dataset, was present in the training data of GPT-4 (OpenAI, 2023, footnote 5). Without more transparency on the training data and possible data contamination, conclusions about models achieving near-human performance cannot be drawn.

D.2 TOMI: Reasoning about Mental States and Realities

Results on TOMI are plotted in Figure 12, showing that instruction-finetuning and RLHF only somewhat improve the models' ability to reason about mental states and realities of others. Compared to GPT-3-DAVINCI's performance which is essentially random, both GPT-3.5-IFT and GPT-3.5-RLHF's performance increases by 17% and 19%, respectively. However, in contrast to SOCIALIQA, performance of newer models (GPT-3.5-Turbo, GPT-4) does not substantially improve over GPT-3.5 models.

When breaking down the accuracy by question type (Figure 12b), we find that the performance of these models improves only on factual questions (fact), but stays low for questions about mental states of participants (mind). Indeed, GPT-3.5-Turbo reaches the "best" performance on the MIND subset of TOMI with 60% accuracy, surpassing GPT-4's 59% accuracy.

D.3 Discussion

Based on our new results, it is not clear that the newer generation of models have achieved neural Theory of Mind, corroborating findings by Ullman (2023), (Lenci, 2023), and Marcus and Davis (2023) and debunking claims of "emergence of TOM in LLMs" by Kosinski (2023) and Bubeck et al. (2023). While models may achieve higher accuracy on social intelligence tasks, their ability to model mental states and realities of others is still very far from humans (only 10% over random chance).

New Content Added 2023-03-30

Examining why these instruction-tuned or RLHF models perform somewhat better remains an open question, hindered by the lack of transparency in pretraining and instruction data. Possibly, instruction-tuned models are better able to learn social intelligence due to the more interactional nature of instruction following or dialogue responding compared to static text. However, improvements could solely be due to development and test data leakage as acknowledged by OpenAI (OpenAI, 2023), calling for the development of better evaluation and probing methods for these TOM abilities. Additionally, approaches such as person-centric inductive biases as well as interactive, experiential, or multimodal grounding could improve their ability to model mental states, as discussed in §5.2.