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: SOCIAL IQA (Sap et al., 2019b), which measures models’ ability to understand intents and reactions of participants of social interactions, and TO MI (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 SOCIAL IQA and TO MI, 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.

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 ($\S3$), using the SOCIALIQ 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 ($\S4$). 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 $\S5$, 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
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). 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). 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).

1While 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).

2The 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).

3Most cognitive studies on this subject focus on the English language, which is not representative of the wide variation of

3 SOCIALIQ: 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 SOCIALIQ 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).

SOCIALIQ 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 SOCIALIQ

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 $k$ language structures, and thus limits the cognitive conclusions one can draw about the link between language and Theory of Mind (Blasi et al., 2022).
These findings are in line with recent BIG-Bench results on the AI2 SOCIAL IQA datasets, which might not be enough to reach human-level accuracy. Shown in Fig. 2, GPT-3 models perform suboptimally compared to human performance, suggesting that increasing model size 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 SOCIAL IQA Results

Shown in Fig. 2, GPT-3 models perform substantially worse than humans (>30% less) on the SOCIAL IQA datasets, and also worse than models finetuned on the SOCIAL IQA training set (>20%; Lourie et al., 2021). 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 SOCIAL IQA 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 suggests 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 that they are appreciated," and instead mistakenly predicts that Tracy will want to "save up for vacation," which is what Aubrey would likely do.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Answers</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>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?</td>
<td>Needed to be behind</td>
<td>Agent</td>
</tr>
<tr>
<td>Casey wrapped Sasha’s hands around him because they are in a romantic relationship. How would you describe Casey?</td>
<td>Very loving towards Sasha</td>
<td>Agent</td>
</tr>
<tr>
<td>Tracy held a baby for 9 months and then gave birth to Addison. What will happen to Tracy?</td>
<td>Throw her baby at the wall</td>
<td>Agent</td>
</tr>
<tr>
<td>Kai gave Ash some bread so they could make a sandwich. How would Kai feel afterwards?</td>
<td>Glad they helped</td>
<td>Agent</td>
</tr>
<tr>
<td>Aubrey was making extra money by babysitting Tracey’s kids for the summer. What will Tracy want to do next?</td>
<td>Save up for a vacation</td>
<td>Others</td>
</tr>
<tr>
<td>The people bullied Sasha all her life. But Sasha got even. What will the people want to do next?</td>
<td>Eat their food quickly</td>
<td>Others</td>
</tr>
<tr>
<td>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?</td>
<td>Eat their food quickly</td>
<td>Others</td>
</tr>
<tr>
<td>Aubrey fed Tracy’s kids lunch today when Tracy had to go to work. What will happen to Aubrey?</td>
<td>Be grateful</td>
<td>Agent</td>
</tr>
<tr>
<td>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?</td>
<td>Plan a romantic evening with Sasha</td>
<td>Others</td>
</tr>
</tbody>
</table>

Table 1: Examples of SOCIAL IQA questions, which person the questions focus on (Agent, Others), and the human gold answers (√) and GPT-3-DAVINCI predictions (✓).
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).\textsuperscript{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 ToMi 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

---

\textsuperscript{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.
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 SCoALiQA.

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 archi-

![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.](image)
tectures, 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 SOCIALIQ 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...
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 re-

flect 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. Alterna-
tively, 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 (Bura et al., 2021).

Nevertheless, the scale of training datasets seems
to be crucial for LLMs (Kaplan et al., 2020; Chowd-
bery et al., 2022), which poses a challenge: text
datasets rich in social intelligence and interactions
are not easily found naturally due to reporting bi-
ases, and they are costly to create (Rashkin et al.,
2018; Mostafazadeh et al., 2020). Promising re-

sults on commonsense reasoning suggest a possi-
ble hybrid approach: LLMs could be jointly or
sequentially trained on static text and common-
sense 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 rein-
forcement 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 infor-
mation 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 corefer-
ence chains, which is challenging to obtain for su-
ervised settings, especially at the scale that LLMs
require. Thus, unsupervised approaches to adding
inductive biases to models could be a promising so-
lution. 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 con-
cepts; Bowers, 2017; Calvo Tapia et al., 2020).

Alternatively, examining the internal or latent
representations of LLMs could point to future di-
rections towards inductive biases for neural Theory
of Mind. As an example, recent work has found ev-
dence of latent representations of grounded seman-
tics 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 train-
ing (Patel and Pavlick, 2022). Future work might
similarly analyze deep learning models for repre-
sentations 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 un-
derstanding 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 com-
municative intents and alternatives in static text.
Future work could explore approaches grounded
more explicitly in interaction, intents, and alterna-
tives, 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 capa-
bilities 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 simula-
tions, 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 these 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 SOCIALIQ 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 SOCIALIQ 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, SOCIALIQ 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 (John 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**


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


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


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


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


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


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.


diagram


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.


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


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.


Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Emergent communication fine-tuning (ec-ft) for pretrained language models. In Emergent Communication Workshop at ICLR 2022.


BERT rediscovers the classical NLP pipeline. In ACL.


A unified Encoder-Decoder framework with entity memory. In EMNLP.

The child’s theory of mind. The MIT Press series in learning, development, and conceptual change., 358.

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. 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

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 ToM Details

B.1 Data Preprocessing

We generated ToM 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. After this filtering, 5190 training and 5170 dev. stories remained.

For the final ToM 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

---

**Note:**

8http://maartensap.com/social-qa/data/socialIQa_v1.4_withDims.tgz

9We 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.

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

<table>
<thead>
<tr>
<th>Reasoning Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT</td>
<td>519</td>
</tr>
<tr>
<td>FACT-MEM</td>
<td>278</td>
</tr>
<tr>
<td>FACT-REAL</td>
<td>241</td>
</tr>
<tr>
<td>MIND</td>
<td>1342</td>
</tr>
<tr>
<td>MIND-TB</td>
<td>778</td>
</tr>
<tr>
<td>MIND-1st-TB</td>
<td>389</td>
</tr>
<tr>
<td>MIND-2nd-TB</td>
<td>389</td>
</tr>
<tr>
<td>MIND-FB</td>
<td>564</td>
</tr>
<tr>
<td>MIND-1st-FB</td>
<td>231</td>
</tr>
<tr>
<td>MIND-2nd-FB</td>
<td>333</td>
</tr>
<tr>
<td>total</td>
<td>1861</td>
</tr>
</tbody>
</table>

shown in Tab. 4. We release our final preprocessed ToMi dev. dataset at [http://maartensap.com/neuralToM/ToMi-finalNeuralTOM.csv](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).

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: \( \text{arg max}_i p_{LM}(a_i | c, q, C_k) \) where \( C_k \) denotes the \( k \) training examples,