Does reading words help you to read minds? A comparison of humans and LLMs at a recursive mindreading task

Cameron R. Jones
Department of Cognitive Science
UC San Diego,
cameron@ucsd.edu

Sean Trott
Department of Cognitive Science
UC San Diego,
strott@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego,
bkbergen@ucsd.edu

Abstract

There is considerable debate about the origin, mechanism, and extent of humans' capacity for recursive mindreading: the ability to represent beliefs about beliefs about beliefs (and so on). Here we quantify the extent to which language exposure could support this ability, using a Large Language Model (LLM) as an operationalization of distributional language knowledge. We replicate and extend O'Grady, Kliesch, Smith, and Scott-Phillips (2015)'s finding that humans can mindread up to 7 levels of embedding using both their original method and a stricter measure. In Experiment 2, we find that GPT-3, an LLM, performs comparably to humans up to 4 levels of embedding, but falters on higher levels, despite being near ceiling on 7th-order non-mental control questions. The results suggest that distributional information (and the transformer architecture in particular) can be used to track complex recursive concepts (including mental states), but that human mentalizing likely draws on resources beyond distributional likelihood.

Keywords: theory of mind; large language models; recursive mindreading; distributional statistics

Introduction

Humans keep track of others' mental states: a phenomenon known as mentalizing, mindreading, or theory of mind. *Recursive* mindreading occurs when the tracked mental states are themselves about others' mental states (e.g. knowing that Andy is embarrassed that Bella knows he likes Carl). While it is well established that humans can can track beliefs at 1-2 levels of recursive embedding (Apperly, 2012; Wellman, Cross, & Watson, 2001), [1] there is considerable disagreement about whether humans have the need or the capacity to process more deeply embedded mental representations.

Theoretical arguments for deeply recursive ToM in humans point to our evolutionary origins and the pragmatic demands of communication. Our social and cultural environment may have provided evolutionary incentives to develop recursive mindreading, including social learning (Sperber, 2000), reputation management, and story-telling (Dunbar, 2009). Gricean pragmatic theories of communication (Grice, 1975) require constant monitoring of recursively embedded mental representations (e.g. knowing$_0$ that Bob knows$_1$ that you know$_2$ that if he had meant$_3$ all of the papers have been graded he would have said "all" and not "some".) Rational Speech Act theory (Goodman & Frank, 2016) further formalizes these pragmatic principles, suggesting that listeners interpret utterances using a probabilistic model of the speaker's model of the listener's interpretative process. Sperber and Wilson (1986)'s relevance theory posits that everyday communicative interaction relies, in principle, on a 4th order recursive mental representation (e.g. Peter believes$_0$ that Mary wants$_1$ Peter to know$_2$ that Mary wants$_3$ Peter to know$_4$ that her glass is empty), extending to higher levels when listeners are considering utterances about mental representations.

Others, however, have suggested that deeply recursive mindreading may be too cognitively demanding for human comprehenders. Sperber and Wilson (1986) suggest that recursive mutual knowledge need not be deliberatively inferred in every interaction, as long as it is mutually inferable. A variety of simpler processes have been proposed that achieve the functional effects of recursive mindreading without requiring explicit mental models, such as inferring shared attention from eye contact (Gómez, 1994) and associating common ground with speakers generically rather than specific speaker-listener dyads (Shintel & Keysar, 2009). Heyes (2014) discusses a range of generic spatial and attentional mechanisms she calls 'submentalizing' that could explain apparent mind-reading without explicit recursive representations. However, debates around whether simpler mechanisms could account for complex mindreading have proven challenging to resolve without explicit models of these processes and their predictive power.

The debate over recursive mindreading relates to a wider debate about the origins and mechanisms behind humans' theory of mind. One dominant theory is that we have an evolved biological endowment for mentalizing (Bedny, Pascual-Leone, & Saxe, 2009). Our capacity to represent others' thoughts is likely to have been important for our evolution as a highly social species (Dunbar, 2009; Sperber, 2000). Alternative theories argue that mindreading is acquired across the lifetime based on our social experience (Hughes et al., 2005), and in particular our exposure to language (de Villiers & de Villiers, 2014; P. L. Harris, 2005). Language provides a rich framework to both share and represent our inner mental lives. Mental state verbs such as *know* and *believe* provide explicit cues to unobservable mental states (J. R. Brown, Donelan-McCall, & Dunn, 1996), while syntactic structures like sentential complements (*Sarah believes that X*) allow us to recursively embed others' thoughts (Hale & Tager-Flusberg, 2003). However, existing work has not quantified to what extent language exposure alone could account for complex mentalizing behavior such as recursive mindreading.

[1] Following (O'Grady et al., 2015) we count levels excluding the focal individual: i.e. I think$_0$ Mary thinks$_1$ John wants$_2$ cake.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Here, we address that question using a Large Language Model (LLM) as a *distributional baseline* to ask: to what extent can purely statistical information about the distribution of words in language be used to predict human mentalizing behavior. The distribitional hypothesis (Firth, 1957; Z. S. Harris, 1954) posits that human language comprehension is—in part—based on learning statistical relationships about which words are likely to follow other words. LLMs provide a computational operationalization of the distributional hypothesis; they learn to predict words on the sole basis of distributional language statistics. They lack innate biases and feedback from social interaction, allowing us to test the sufficiency of language exposure alone to explain behavior. Importantly, for this analysis we focus on base or "vanilla" LLMs that have not been additionally fine-tuned using reinforcement learning from human feedback (RLHF; Ouyang et al., 2022). Although RLHF may improve performance, it constitutes an additional training signal beyond distributional language statistics, complicating the inferences we can draw.

In spite of their conceptual simplicity and lack of access to non-linguistic resources, LLMs have been shown to predict diverse measures of human language comprehension (Chang & Bergen, 2023) and brain activity (Michaelov, Coulson, & Bergen, 2022). More specifically, LLMs have been found to perform at above-chance levels on the False Belief task, which measures 1st-order mindreading (Trott, Jones, Chang, Michaelov, & Bergen, 2023). However, an ongoing debate concerns whether these results reflect evidence of a general mindreading capacity, or exploitation of cheaper statistical shortcuts (Gandhi, Fränken, Gerstenberg, & Goodman, 2023; Jones, Trott, & Bergen, to appear; Kim et al., 2023; Shapira et al., 2023; Ullman, 2023). Evidence that LLMs can exploit such shortcuts in instruments used to assess mindreading in humans is especially germane to the discussion of submentalizing above. We return to this issue in our general discussion.

In order to quantify the sufficiency of distributional information for recursive mindreading, we wanted to select a task that accurately reflected human performance. Early empirical evidence for recursive mindreading—mostly drawn from the Imposing Memory Task (IMT)—indicated that human accuracy deteriorates at around 4 levels of embedding. Kinderman, Dunbar, and Bentall (1998) found accuracy at the IMT dropped from 85% to below 50% when the level of recursion was increased from 4 to 5, despite the fact that performance on control questions remained above 90% up to 6 levels of embedding. Stiller and Dunbar (2007) ran an extended version of the IMT, including questions up to 8 levels of embedding, and found that the level at which participants fail was normally distributed around a mean of 4.

O'Grady et al. (2015), however, identified various problems with the IMT—including unanswerable questions and uncontrolled variance in syntactic complexity— and designed a novel recursive mindreading task that addresses these problems. In their study, participants watched videos which had a plot involving 7 levels of recursively-embedded mental representation, and 7 levels of a non-mental recursive concept, such as possession. For each of the levels of mental and non-mental recursion, the authors also created two scenes to follow the main story, only one of which was consistent with it. Participants watched the scenes and had to select the one that was consistent with the story. O'Grady et al. found that human comprehenders were successful at answering questions that involved up to 7 levels of recursive embedding (the maximum level used in their task). Moreover, they did not find a negative effect of embedding level on accuracy. We selected this experiment as a model to measure recursive mindreading in humans and LLMs because it addresses problems with stimuli in previous studies and provides a strong human baseline to which we can compare models.

In Experiment 1, we conducted a replication of O'Grady et al. (2015) in a text-only format to check that their finding was robust and to obtain human data that would be more comparable to LLM responses. We also ran an extension with small modifications to test potential confounding explanations for participants' strong performance in this task. In Experiment 2, we elicited responses to the same stimuli from GPT-3, an LLM, and tested whether humans outperform models and the extent to which their responses are correlated.

Experiment 1

O'Grady et al. (2015)'s finding of 7th-order mindreading is striking and contrasts with previous literature: in itself providing motivation for replication. In addition, several features of the design suggest potential alternative explanations for participants' strong performance. First, stimuli were presented as videos rather than text. As one of several differences with previous work, it would be valuable to know whether video presentation was necessary for participants' success. Second, questions were presented in a two-alternative forced choice (2AFC) format, allowing participants to directly compare proposed responses. For instance, a participant could identify that the only difference between the choices in (1) (Story 1's level 5 mental question, emphasis ours) is whether Stephen knows or does not know what Elaine knows.

(1) a. Megan knows that Stephen *doesn't know* that Elaine knows that Bernard feels that she doesn't know him well enough to date

b. Megan knows that Stephen *knows* that Elaine knows that Bernard feels that she doesn't know him well enough to date

This could potentially simplify participants' reasoning process; they might only remember that Stephen didn't know something important about Elaine, prompting them to guess 'a' without fully processing the conceptual chain. Moreover, participants could assume that whatever followed the difference (and is stated in both answers) must be true.

Finally, questions were designed such that they often concerned different portions of the same conceptual chain. For instance, (2) is the level 3 mental question for Story 1:

(2) a. Elaine doesn't know that Bernard feels that she doesn't know him well enough to date.
b. Elaine knows that Bernard feels that she doesn't know him well enough to date.

The correct answer to (2), 'b', is contained in both versions of (1), as well as the questions for levels 4, 6, & 7. Because participants saw each story's questions in a random order, they were often exposed to the correct answer to (2) many times before encountering the question itself. Participants could have selected this answer through an explicit strategy to exploit this fact, or simply because it sounded more familiar.

In order to understand the extent to which these factors accounted for performance in O'Grady et al's original study, we ran a replication and extension (preregistered here). The experiment randomly assigned participants to two conditions. In one they completed a close replication of O'Grady et al. (using the 2AFC response format), except with all stimuli as text. We used the data from this study to test 4 hypotheses. We hypothesised (1) that we would replicate the main finding that participant accuracy on mental questions would be significantly above chance for all 7 embedding levels. However, we predicted that in the text-only format, there would be (2) a negative effect of embedding level on question accuracy, and (3) a negative interaction between mental question type and embedding level. Finally, we hypothesised that participants would use information from earlier questions within each block to answer later questions, so that (4) participant accuracy would increase across trials within each block.

The second response format condition (TF) was identical, except that participants only ever saw one possible answer to each question and responded true or false (i.e. whether the response was consistent with the story). In this condition, participants could not compare responses to simplify reasoning, so we hypothesized that they would (5) be less accurate overall and (6) show a larger negative effect of question level. In addition, it was more difficult for them to infer answers to later questions from earlier ones, so we hypothesized (7) that they would improve less within blocks. Finally, we hypothesized that even with these restrictions, participants would still (8) perform above chance up to 7 levels of embedding.

Methods

The original materials and method are described in more detail in O'Grady et al. (2015). All materials, data, and analysis code are available on OSF here.

Materials The stimuli comprised 4 stories, whose plots involved 7 levels of recursively embedded mental representation, and 7 levels of a recursively embedded non-mental control concept, such as possession. For each of the levels of mental and non-mental recursion, the authors also created two possible continuations, only one of which was consistent with the main story. All stories and continuations were written in two different formats: as dialogues and as narratives. In total there were 112 pairs of continuation passages. While the original study recorded actors reading scripts, we presented all stimuli as text. The dialogues were displayed with speaking characters' names in bold and stage directions in italics. The stimuli were adapted from U.K. to U.S. English for U.S. participants. In addition to the original 2AFC question format, we added a new condition (TF), in which participants were presented with only one continuation and judged if it was consistent or inconsistent with the main passage.

Participants We recruited 128 UC San Diego undergraduate students (97 female, 23 male, 6 nonbinary, 2 prefer not to say; mean age = 21.6, sd = 4.1), who completed the experiment for course credit. We excluded 9 participants who indicated they had taken notes, and 2 participants who failed 3/8 level 1 questions, indicating inattention. After exclusions, we retained 117 participants (54 2AFC, 63 TF), who completed 6252 trials (2889 2AFC, 3363 TF).

Procedure The study was designed in jsPsych (De Leeuw, 2015) and hosted online. Participants were randomly assigned to either the TF or 2AFC condition (between subjects). Participants first completed 2 practice questions, with feedback, to ensure task understanding. They then read 4 stories, each followed by 14 questions (7 mental, 7 control) with different levels of recursive embedding (1-7) in a random order. The format (implicit vs explicit) of the story and questions was fully crossed within participant. Finally, participants completed a debrief asking about strategies they employed.

Results

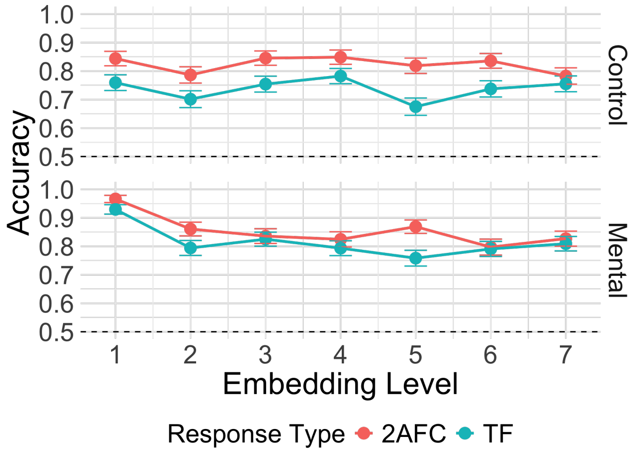

Figure 1: Mean accuracy by question type (control vs mental), response format (2AFC vs TF) and depth of recursive embedding (1-7). Dashed lines represent chance performance. Participants in the 2AFC condition performed slightly better ($z = -2.21$, $p = 0.027$), but participants in both response format conditions performed significantly above chance through all embedding levels (all $p < 0.004$).

We constructed linear mixed effects models with maximal random effects structures (Barr, Levy, Scheepers, & Tily, 2013). Where models failed to converge, we iteratively removed random effects in a prespecified order of theoretical importance. We used *lmertest* (Kuznetsova, Brockhoff, & Christensen, 2015) to estimate statistical significance.

Our first 4 hypotheses concerned the 2AFC data alone. (1) We replicated the original finding that accuracy on mental questions was significantly above chance at each embedding level (all $p < 0.004$, see Figure 1). (2) In contrast to the original study, we found a significant negative effect of embedding level on mental question accuracy ($\beta = -0.156$, $z = -4.05$, $p < 0.001$). (3) Contrary to our hypothesis, the effect of embedding level was more negative for control vs mental questions (i.e. there was a positive interaction of mental question type with question level). (4) Again contradicting our hypothesis, we found no effect of "block trial index"—the order in which participants completed items within blocks ($\beta = 0.070$, $z = 1.52$, $p = 0.129$)—suggesting that participants' high performance cannot be explained by their learning across blocks.

The next 4 hypotheses pertained to comparisons between the 2AFC and TF data. (5) Participants in the TF condition were significantly less accurate overall (81% vs 85%; $z = -2.21$, $p = 0.027$). (6) Contrary to our hypothesis, there was no negative interaction between response format and embedding level within mental questions. (7) Again, in contrast to our prediction, there was no interaction between response format and block trial index, suggesting that participants did not improve more within blocks in the 2AFC vs TF conditions. Mean accuracy of TF participants on the first trial of each block was 78.9%. (8) Finally, as predicted, TF participants achieved an accuracy that was significantly above chance at all embedding levels (all $p < 0.001$; see Figure 1).

Discussion

The results robustly confirm the primary result of (O'Grady et al., 2015): comprehenders were able to track characters' mental states up to 7 levels of embedding using text-based stimuli, and a more challenging response format that removed alternative routes to the correct answer. However, in contrast to the original study, we saw a negative effect of embedding level on accuracy within mental questions. This might be more consistent with the prior expectation that increased embedding is costly, but future work is needed to understand whether this difference in results stems from the participant population, video- versus text-based format, or other design differences.

We did not find support for our concerns that participants were using shortcuts to answer questions without processing recursive conceptual chains. Although there was a small effect of response format overall ($z = -2.21$, $p = 0.027$), TF participants did not improve more within blocks and were significantly above chance at all embedding levels. We therefore used the 2AFC format and human data for Experiment 2.

Experiment 2

At a high level, language models are probability distributions over sequences of words or word-parts called tokens. While early language models directly used token frequencies or static representations to learn transition probabilities between sequences, the self-attention mechanism of the transformer architecture allows LLMs to make predictions about upcoming tokens in a highly context-sensitive way. An LLM's representation of each token in the input is influenced by the token's relationship to every preceding token. This allows models to represent different meanings of polysemous words differently (Trott & Bergen, 2023). It also theoretically allows models to encode information about the state of a referent in an unfolding situation description (Li, Nye, & Andreas, 2021; Wang, Variengien, Conmy, Shlegeris, & Steinhardt, 2022).

This mechanism could be used to keep track of characters' mental states. For instance, the model could learn that if Bernard is described as overhearing that Elaine is upset, the token 'Bernard' is more likely to be followed by sequences describing his knowledge of this fact. Existing work suggests that LLM's contextual representations can track mental states up to 1 or 2 levels of recursive embedding (Gandhi et al., 2023; Trott et al., 2023). Here we investigate whether this sensitivity extends to more deeply recursive embedding.

We tested 4 hypotheses, pre-registered here, about whether LLMs could track mental states and the extent to which they could explain this behavior in humans. We hypothesized that GPT-3 accuracy on mental questions would be (1) significantly above chance up to 7 levels of embedding but (2) significantly below human accuracy overall. To test the stronger claim that humans and GPT-3 use similar information to answer questions, we asked (3) whether human and GPT-3 responses were correlated across items, and (4) whether human accuracy was significantly above chance after controlling for the likelihood that GPT-3 assigned to each response. In an additional exploratory analysis of the importance of model scale, we presented the stimuli to 4 smaller GPT-3 variants.

Methods

Materials We used the same story and question stimuli that were presented to human participants. To test whether the materials had been included in GPT-3's training data, we conducted a data contamination analysis using the guided instruction method (Golchin & Surdeanu, 2023). We used GPT-3 to generate completions for the first half of each story either with (guided) or without (unguided) a prompt prefix describing the origin of the data. We compared guided and unguided generations to the reference completions and found no significant difference in either BLEURT ($p = 0.91$) or ROUGE-L scores ($p = 0.91$), and no near-exact matches. These results suggest that the materials were not in GPT-3's training data.

Models For our main analysis, we used GPT-3 *text-davinci-002* (T. Brown et al., 2020), an autoregressive language model with 175 billion parameters, pre-trained on 300bn tokens, and additionally fine-tuned on instruction following data. Because we are interested in the sufficiency of distributional information specifically, we did not use more recent models (GPT-3.5 and GPT-4) that have been additionally fine-tuned using (RLHF). Additionally, to investigate the role of model

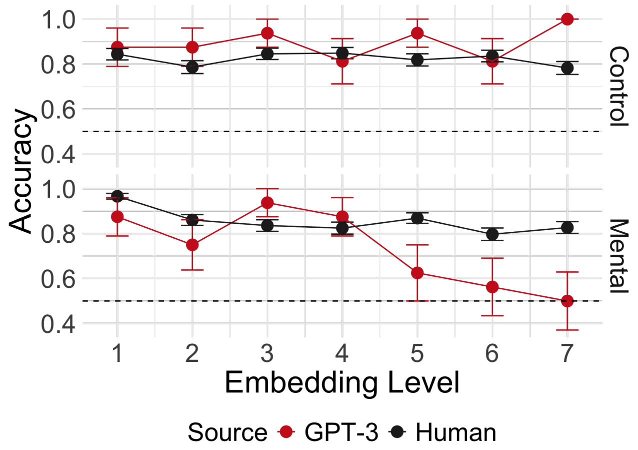

Figure 2: Mean accuracy by question type and embedding level for human (black) and GPT-3 (red) responses. GPT-3 accuracy on mental questions was significantly above chance for levels 1, 3, & 4, but deteriorated after level 5. GPT-3 accuracy remained high ($> 80\%$) up to 7 levels of recursive embedding for non-mental control questions.

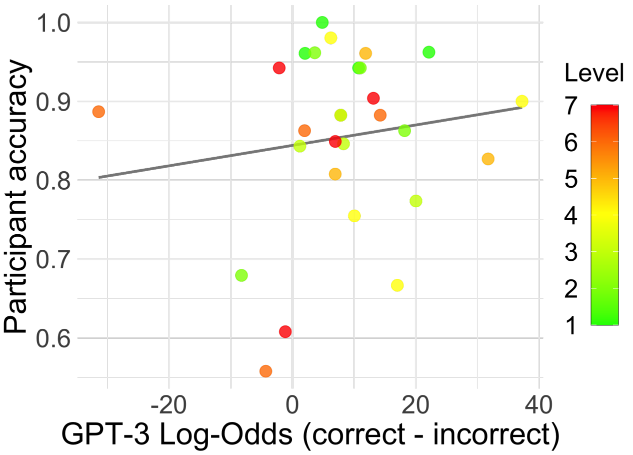

Figure 3: There was a small positive correlation between participant accuracy on mental questions and the log-odds ratio between the probabilities that GPT-3 assigned to correct and incorrect options (r=0.14, $z = 2.095$, $p = 0.036$). However, there was a large amount of variance in human responses that was not explained by GPT-3 predictions.

scale, we perform the analyses on 4 pre-trained base GPT-3 models with different numbers of parameters: *ada* (2.7B), *babbage* (6.7B), *curie* (13B), and *davinci* (175B).

Procedure We operationalized the experiment as a language modeling task, where the model predicted which of the completions would be more likely to follow the main story. For each question, we presented each potential response separately and measured the probability assigned to each response conditioned on the story. We presented all 4 combinations of story and question format (implicit vs explicit). Because continuations varied considerably in length and other surface features, we used PMI_{DC} to control for the probability of the continuation independent of the story (Holtzman, West, Shwartz, Choi, & Zettlemoyer, 2022). We scored the LLM as correct if it assigned a higher probability to the consistent response, and operationalized the LLM's preference for the correct response as the log-odds ratio $(log(p([correct])) - log(p([incorrect]))$, corrected with PMI_{DC}.

Results

The results showed (1) that GPT-3 accuracy was significantly above chance overall ($z = 4.03$, $p < 0.001$), and was above chance for embedding levels 1, 3, & 4 (all $p < 0.01$). However, it was not significantly above chance for levels 2, 5, 6, & 7 (all $p > 0.3$). (2) Human accuracy on mental questions (85%) was significantly above GPT-3's (73%, $z = 3.42$, $p < 0.001$), with 87% of participants outperforming the LLM. In a post-hoc exploratory analysis, we found GPT-3 accuracy on level 1-4 questions was not significantly different from humans (86% vs 87%, $z = 0.299$, $p = 0.765$. (3) The log-odds ratio between correct and incorrect responses was a significant positive predictor of human accuracy ($z = 2.095$, $p = 0.036$, Figure 3). (4) Human accuracy was significantly above chance at each embedding level, even after controlling for the predictive effect of GPT-3 log-odds (all $p < 0.03$).

We also conducted a scaling analysis, presenting the same stimuli to smaller GPT-3 variants. A model predicting mental question accuracy on the basis of the log of the number of parameters showed a slight increase in scale from *ada* (63%) to *davinci* (65%) ($z = 3.06, p = .002$). There was a much larger gap in performance between the base *davinci* model and the *text-davinci-002* model used in our main analysis (73%, see Figure 4). There was a more marked effect of scale on non-mental control questions from *ada* 73% to *davinci* (95%).

Discussion

GPT-3 performed above chance at recursive mindreading overall, specifically below 5 levels of embedding, where it performed comparably to humans. To our knowledge, this is the first result suggesting that LLMs exhibit behavior consistent with recursive mindreading beyond 2 levels of embedding. Accuracy at level 2 specifically was not significantly above chance. However, given that accuracy was 75% and that the model was more accurate at levels 3 and 4, this null result could be due to insufficient power (16 observations).

The model's performance dropped sharply at level 5, suggesting that it is poor at tracking more deeply embedded mental representations. Importantly, however, GPT-3 accuracy remained above 80% through to 7 levels of recursion on non-mental control concepts, implying that the model does not struggle with complex recursive chains *per se* but that recursive mentalizing specifically is particularly challenging.

There was a weak positive inter-item correlation between GPT-3 predictions and human accuracy. However, a large

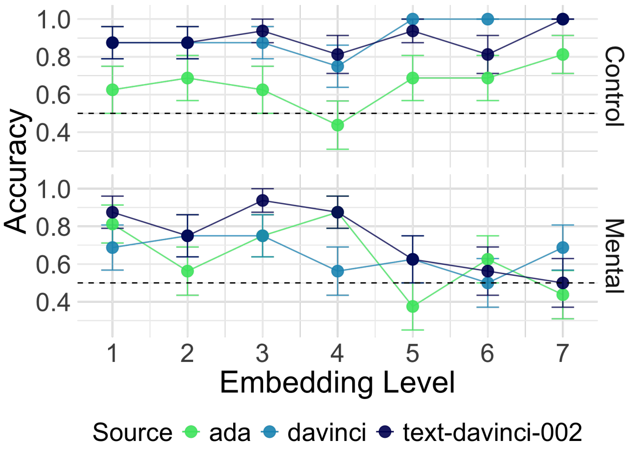

Figure 4: Accuracy for a subset of GPT-3 models. *davinci* and the fine-tuned *text-davinci-002* perform near ceiling on control questions at all embedding levels, outperforming *ada*. All models perform worse on Mental questions, especially after 5 levels of embedding. However, the difference between the smallest and largest models is less pronounced.

proportion of variation in human responses was not explained by the model (see Figure 3), suggesting that humans and LLMs used different information to respond to questions. This was confirmed by the distributional baseline analysis; human comprehenders were significantly above chance after controlling for the distributional likelihood of responses as measured by GPT-3. Even for question levels where humans and GPT-3 had similar accuracy, human behavior was not explained by the probability that GPT-3 assigned to responses.

The scaling analysis showed a weak positive effect of model scale, with a much more pronounced difference between *davinci* and *text-davinci-002* on mental questions. These models are the same size, but the latter was fine-tuned on instruction following data. Unfortunately we do not know exactly what this data comprised, but it indicates that data quality may be more important than model scale for this task.

General Discussion

We ran a replication and extension of O'Grady et al. (2015) to compare recursive mindreading in humans and distributional language models. Our results provide a robust confirmation of O'Grady et al.'s result that human comprehenders can track mental states up to 7 levels of embedding. Not only did the result replicate with text-only stimuli, but we found that participants' mindreading abilities could not be explained by exploiting information in the 2AFC answer format or learning across blocks. Even in the first trial per block, mean accuracy in the TF condition was 78.9%. The result is consistent with claims that recursive mindreading is deeply engrained in human cultural and evolutionary history (Dunbar, 2009; Sperber, 2000), and provides support for the plausibility of pragmatic theories that require frequent recourse to deep mental recursion (Grice, 1975; Sperber & Wilson, 1986).

GPT-3 also performed significantly above chance on the recursive mindreading task overall, and performed comparably to humans up to 4 levels of embedding. As has been noted elsewhere (Trott et al., 2023), evidence that LLMs are successful at a given task can be interpreted in multiple ways: as evidence that the LLM has the construct which the task is designed to measure (Kosinski, 2023); that the task is a flawed measure of the construct (Bender & Koller, 2020); or that the task has differential construct validity for humans and LLMs (Ullman, 2023). The last interpretation relies on arguing that LLMs solve the task in a superficial way that obviates the processes implicated in theories of mentalizing. We argue that claims of differential construct validity must be supported by empirical evidence of systematic errors (McCoy, Yao, Friedman, Hardy, & Griffiths, 2023). Mechanistic interpretation could be especially helpful in resolving this debate (Li et al., 2021; Lepori, Serre, & Pavlick, 2023). Evidence that LLMs encode theoretically important intermediate representations (e.g. mental states) could suggest they are not using superficial shortcuts.

Finally, we were interested in quantifying the extent to which distributional information learned by LLMs could explain human mindreading behavior. This question differs from the previous one. LLMs could be found to solve a task in a theoretically interesting way that nevertheless differs from human cognitive mechanisms. We found scant evidence to support this stronger claim and some evidence against it. GPT-3 failed specifically at mental (but not control) questions beyond 4 levels of embedding, while humans performed similarly at both question types up to 7 levels. GPT-3 probabilities were predictive of human accuracy, yet the correlation was weak, and human accuracy was still high after accounting for this correlation. If humans were using the same kind of distributional information as GPT-3 to solve these problems, we would expect their responses to be better correlated. Therefore, even where GPT-3 accuracy is comparable to humans, the results suggest that it generates its responses in a different manner than human comprehenders. In turn, this suggests that humans draw on mechanisms beyond distributional language statistics to reason about mental states.

Although larger models not tested here might perform better, GPT-3 is arguably already overpowered with respect to linguistic training data—being exposed to more than 1000x as much text as a human (Warstadt & Bowman, 2022). In fact, the large performance difference between *text-davinci-002* and *davinci* suggests that quality of data may matter more for this task than scale. Many theories specify particular types of linguistic input thought to be most helpful for learning to represent mental states, whether sentential complements (de Villiers & de Villiers, 2014), mental state verbs (J. R. Brown et al., 1996), or dialogue (P. L. Harris, 2005). Future work could mimic training studies in humans (Hale & Tager-Flusberg, 2003) by fine-tuning LLMs on different types of language input and measuring the efficacy of each.

Acknowledgements

We would like to thank the anonymous reviewers and metareviewer for comments that strengthened this paper.

References

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. doi: 10.1080/17470218.2012.676055

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009, July). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences*, *106*(27), 11312–11317. doi: 10.1073/pnas.0900010106

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). doi: 10.18653/v1/2020.acl-main.463

Brown, J. R., Donelan-McCall, N., & Dunn, J. (1996). Why Talk about Mental States? The Significance of Children's Conversations with Friends, Siblings, and Mothers. *Child Development*, *67*(3), 836–849. doi: 10.1111/j.1467-8624.1996.tb01767.x

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Chang, T. A., & Bergen, B. K. (2023, August). *Language Model Behavior: A Comprehensive Survey* (No. arXiv:2303.11504). arXiv.

de Villiers, J. G., & de Villiers, P. A. (2014). The Role of Language in Theory of Mind Development. *Topics in Language Disorders*, *34*(4), 313–328. doi: 10.1097/TLD.0000000000000037

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12.

Dunbar, R. (2009, January). The social brain hypothesis and its implications for social evolution. *Annals of Human Biology*, *36*(5), 562–572. doi: 10.1080/03014460902960289

Firth, J. R. (1957). *A synopsis of linguistic theory*. Oxford: Blackwell.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023, December). *Understanding Social Reasoning in Language Models with Language Models* (No. arXiv:2306.15448). arXiv. doi: 10.48550/arXiv.2306.15448

Golchin, S., & Surdeanu, M. (2023, October). *Time Travel in LLMs: Tracing Data Contamination in Large Language Models* (No. arXiv:2308.08493). arXiv.

Gómez, J. C. (1994, May). Mutual awareness in primate communication: A Gricean approach. In S. T. Parker, R. W. Mitchell, & M. L. Boccia (Eds.), *Self-Awareness in Animals and Humans* (1st ed., pp. 61–80). Cambridge University Press. doi: 10.1017/CBO9780511565526.007

Goodman, N. D., & Frank, M. C. (2016, November). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi: 10.1016/j.tics.2016.08.005

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, *6*(3), 346–359. doi: 10.1111/1467-7687.00289

Harris, P. L. (2005). Conversation, Pretense, and Theory of Mind. In *Why language matters for theory of mind* (pp. 70–83). New York, NY, US: Oxford University Press. doi: 10.1093/acprof:oso/9780195159912.003.0004

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. doi: 10.1080/00437956.1954.11659520

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131–143. doi: 10.1177/1745691613518076

Holtzman, A., West, P., Shwartz, V., Choi, Y., & Zettlemoyer, L. (2022, November). *Surface Form Competition: Why the Highest Probability Answer Isn't Always Right* (No. arXiv:2104.08315). arXiv.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child development*, *76*(2), 356–370.

Jones, C. R., Trott, S., & Bergen, B. K. (to appear). Comparing Humans and Large Language Models on an Experimental Protocol Inventory for Theory of Mind Evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*.

Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., & Sap, M. (2023, October). *FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions* (No. arXiv:2310.15421). arXiv.

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998, May). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204. doi: 10.1111/j.2044-8295.1998.tb02680.x

Kosinski, M. (2023, March). *Theory of Mind May Have Spontaneously Emerged in Large Language Models* (No. arXiv:2302.02083). arXiv. doi: 10.48550/arXiv.2302.02083

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R package version*, *2*(0), 734.

Lepori, M. A., Serre, T., & Pavlick, E. (2023). Uncovering

Intermediate Variables in Transformers using Circuit Probing. *arXiv e-prints*, arXiv–2311.

Li, B. Z., Nye, M., & Andreas, J. (2021, June). *Implicit Representations of Meaning in Neural Language Models* (No. arXiv:2106.00737). arXiv.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (No. arXiv:2309.13638). arXiv.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*.

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015, July). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, *36*(4), 313–322. doi: 10.1016/j.evolhumbehav.2015.01.004

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., ... Shwartz, V. (2023, May). *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (No. arXiv:2305.14763). arXiv. doi: 10.48550/arXiv.2305.14763

Shintel, H., & Keysar, B. (2009). Less Is More: A Minimalist Account of Joint Action in Communication. *Topics in Cognitive Science*, *1*(2), 260–273. doi: 10.1111/j.1756-8765.2009.01018.x

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. *Metarepresentations: A multidisciplinary perspective*, *10*, 117–137. doi: 10.1093/oso/9780195141146.003.0005

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press Cambridge, MA.

Stiller, J., & Dunbar, R. (2007, January). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*(1), 93–104. doi: 10.1016/j.socnet.2006.04.001

Trott, S., & Bergen, B. (2023, March). Word meaning is both categorical and continuous. *Psychological Review*. doi: 10.1037/rev0000420

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. doi: 10.1111/cogs.13309

Ullman, T. (2023, March). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (No. arXiv:2302.08399). arXiv.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022, November). *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small* (No. arXiv:2211.00593). arXiv. doi: 10.48550/arXiv.2211.00593

Warstadt, A., & Bowman, S. R. (2022, August). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition* (No. arXiv:2208.07998). arXiv.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. doi: 10.1111/1467-8624.00304