EPITOME: Experimental Protocol Inventory for Theory Of Mind Evaluation

Cameron Jones [1] **Sean Trott** [1] **Benjamin Bergen** [1]

Abstract

We address a growing debate about the extent to which large language models (LLMs) produce behavior consistent with Theory of Mind (ToM) in humans. We present EPITOME: a battery of six experiments that tap diverse ToM capacities, including belief attribution, emotional inference, and pragmatic reasoning. We compare performance of five LLMs to a baseline of responses from human comprehenders. Results are mixed. LLMs display considerable sensitivity to mental states and match human performance in several tasks. Yet, they commit systematic errors in others, especially those requiring pragmatic reasoning on the basis of mental state information. Such uneven performance indicates that attributing ToM to LLMs might be premature.

1. Introduction

Theory of Mind (ToM) is a broad construct encompassing a range of social behaviors from reasoning about others' beliefs and emotions to understanding non-literal communication (Apperly, 2012; Beaudoin et al., 2020). These *mentalizing* or *mindreading* capacities underpin social intelligence (Frith & Frith, 2012), allowing us to anticipate others' actions (Tomasello et al., 2005), solve social coordination problems (Sebanz et al., 2006) and understand communicative intent (Grice, 1975; Sperber & Wilson, 2002).

There is growing interest whether artificial intelligence (AI) agents could display ToM abilities (Johnson & Iziev, 2022; Langley et al., 2022; Rabinowitz et al., 2018). Many desirable AI applications require something akin to ToM, including recognizing users' intents (Wang et al., 2019), displaying empathy toward users' emotions (Sharma et al., 2021), and interpreting requests in the context of users' goals (Dhelim et al., 2021). Equally, improved mentalizing could improve models' capacity for deception (Sarkadi et al., 2019), leading to safety concerns (Ngo et al., 2023). The recent success of Large Language Models (LLMs) has further intensified interest and optimism in the potential for artificial ToM. Although their pre-training regime does not explicitly include social interaction or communicative intent (Bender & Koller, 2020), LLMs produce text which superficially bears many hallmarks of mentalizing (Shevlin, under review; Y Arcas, 2022). However, studies evaluating LLM performance on ToM tasks have yielded inconsistent findings, sparking debates on models' capacities (Kosinski, 2023; Sap et al., 2022; Ullman, 2023). Here, we collate six diverse tasks to investigate the consistency of LLMs' ToM capabilities.

A variety of tasks have been designed to measure different facets of mentalizing (Happé, 1994; Premack & Woodruff, 1978; Wimmer & Perner, 1983). Unfortunately, these measures exhibit poor convergent validity—performance in one task does not necessarily correlate with any other—and limited predictive validity, with task performance failing to consistently predict socioemotional functioning (Gernsbacher & Yergeau, 2019; Hayward & Homer, 2017; Warnell & Redcay, 2019). This limits the extent to which performance on a single task can be taken as evidence of ToM more generally (Schaafsma et al., 2015), and underscores the need for running varied, tightly controlled experiments each measuring distinct aspects of mentalizing. We select six experimental psychology tasks that collectively measure a diverse set of ToM-related abilities including belief attribution, emotional reasoning, non-literal communication, and pragmatic inference.

Beyond measuring LLMs' ToM performance, these models can provide insights into debates on human ToM's evolutionary and developmental origins (Krupenye & Call, 2019; Premack & Woodruff, 1978). Researchers disagree about whether ToM is an innate, evolutionary adaptation (Bedny et al., 2009; Surian et al., 2007) or learned via social interaction (Harris, 2005; Hughes et al., 2005) and language (Brown et al., 1996; de Villiers & de Villiers, 2014; Hale & Tager-Flusberg, 2003; Milligan et al., 2007). If language exposure is sufficient for human ToM, then the statistical information learned by LLMs could account for variability in human responses. We collate human responses to each task for comparison with LLM performance, using identical materials for both. This approach allows us to ask

[1]Department of Cognitive Science, University of California, San Diego, CA, U.S.A.. Correspondence to: Cameron Jones <cameron@ucsd.edu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.

where LLMs sit in the distribution of human scores; whether their accuracy is significantly different from humans; and whether their predictions explain the effects of mental state variables on human responses.

2. Related Work

Several recent studies have directly investigated ToM abilities in LLMs. Sap et al. (2022) evaluated GPT-3 *davinci* on SocialIQA—a crowdsourced dataset of multiple choice questions about social reactions to events (Sap et al., 2019)—and ToMi—a synthetically generated dataset of False Belief Task passages (Le et al., 2019). GPT-3 achieved 55% accuracy on SocialIQA, well below a baseline of 84% set by three human participants. While ToMi lacks a specific human baseline, GPT-3 performed poorly (60%) at belief questions, despite being near ceiling on factual questions.

Kosinski (2023) similarly found that GPT-3 *davinci* performs poorly (40% accuracy) on a range of novel False Belief stimuli (Perner et al., 1987; Wimmer & Perner, 1983). However, later models in the series performed much better. GPT-3 *text-davinci-002*, fine-tuned to follow instructions, achieved 70% accuracy. GPT-3 *text-davinci-003* and GPT-4—fine-tuned using reinforcement learning—achieve 90% and 95% respectively. Although the paper does not establish a human baseline for the novel stimuli, this compares favorably to meta-analyses suggesting typical accuracy of 90% for 7-year olds (Wellman et al., 2001).

Ullman (2023), however, showed that 8 simple perturbations to Kosinski's stimuli cause GPT-3 *text-davinci-003* to fail, suggesting that LLMs exploit shallow statistical patterns rather than deploying a deep, emergent ToM ability. Though these perturbations were not tested with humans or generalized to a larger sample of items, Ullman argues that "outlying failure cases should outweigh average success rates." Most recently, Shapira et al. (2023) evaluated 15 LLMs across 6 tasks incorporating belief attribution (ToMi, False Belief), epistemic reasoning, and social reactions (SocialIQa and Faux Pas), and found that none performed robustly. Moreover, they showed that models were vulnerable to systematic adversarial perturbations in the style of Ullman (2023). It is not clear how humans would perform on these tasks, especially synthetic and adversarial examples that might contain less naturalistic language. However, the authors' comprehensive and diagnostic approach proves very valuable, and we hope to extend and complement it in the present work.

Our contribution differs from existing studies in several ways. First, we include tasks that incorporate a broader range of ToM capacities (including emotional reasoning, pragmatic inference, and non-literal communication), and evaluation criteria (including human evaluation of free-text completions). Second, we use experimental stimuli originally designed to measure ToM in humans. While some researchers are concerned that human experiments may be poorly designed for LLMs (Mitchell & Krakauer, 2023; Shapira et al., 2023; Ullman, 2023), it is unclear that novel tasks (especially those generated synthetically or via crowdsourcing) overcome underlying problems. Moreover, experimental stimuli have the advantage of being carefully designed and validated to control for confounds and measure specific latent constructs (Binz & Schulz, 2023). Third, to minimize the risk of selecting items or analyses that would lead to a given result, we preregistered materials and analyses for four of the six studies. Fourth, to allow direct item-level comparison between model and human performance, for each study we elicit an appropriately powered human baseline for all items. Finally, to test whether experimental variables explain variance in human responses when controlling for language model predictions, we perform pre-planned hierarchical model comparisons.

3. The present study

We assemble EPITOME: a battery of six experiments designed to measure distinct aspects of ToM in humans (see Figure 1). The **False Belief Task (FB)** tests whether participants can maintain a representation of someone else's belief, even if it differs from their own (Wimmer & Perner, 1983). **Recursive Mindreading (RM)** tests whether participants can recursively represent mental states up to seven levels of embedding, e.g. "Alice knows that Bob believes that Charlie..." (O'Grady et al., 2015). The **Short Story Task (ShS)** measures the ability to infer and explain emotional states of characters (Dodell-Feder et al., 2013), while the **Strange Stories Task (StS)** (Happé, 1994) asks participants to explain why characters might say things they do not mean literally. The final two tasks measure sensitivity to speaker knowledge during pragmatic inference. The **Indirect Request Task (IR)** asks whether participants are less likely to interpret an utterance as a request if the speaker knows that the request can't be fulfilled (Trott & Bergen, 2020). The **Scalar Implicature (SI)** task tests whether comprehenders are less likely to interpret *some* to mean *not all* when the speaker does not know enough to make the stronger claim (Goodman & Stuhlmüller, 2013).

We pre-registered our analyses for four of the six tasks, and provide code, data, and materials for all six.[1] To elicit a performance baseline for models, we collect human responses for each task. We ask three types of question: (1) Where do LLMs sit in the distribution of human performance? (2) Are LLMs sensitive to experimental variables that alter characters' mental states? (3) Do LLMs fully explain hu-

[1] https://osf.io/sn7gj/?view_only=a793639cceda492f9020705b89045a31

Figure 1. Truncated examples of materials from each of the 6 Theory of Mind tasks. Participants read a context passage (light text) and then answered a question using the response type indicated in the top-right of each box. For unabbreviated examples, see Appendix B.

man behavior, or is there a residual effect of mental state variables on human comprehenders after controlling for distributional likelihood as measured by LLM predictions?

Our main analysis focuses on GPT-3 *text-davinci-002* (henceforth, GPT-3): one of the best-performing models which has not been trained using Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022). While RLHF has been found to improve performance at a many tasks, it introduces an additional training signal, which complicates inferences about the sufficiency of distributional information. We make our code and materials available to facilitate addressing further questions, including whether RLHF improves performance at ToM tasks. Finally, we perform analyses with smaller GPT-3 variants, to measure the extent to which model performance changes with scale.

4. Methods

We accessed models through the OpenAI API with temperature $= 0$. When measuring the probability assigned to a multi-token string, we summed the log probabilities of each token. The number of human participants in each study varied based on the types of statistical analysis being run, the number of items, and the number of observations per participant. For tasks without explicit correct answers, 'accuracy' is defined as the total score on questions measuring sensitivity to mental states. We use data and analysis from Trott et al. (2023) for the FB component of our battery. LLM data and analyses for all other tasks, as well as human data for RM, StS, and SI are novel contributions.

4.1. False Belief Task

Materials Trott et al. (2023) constructed 12 passage templates, in which a main character puts an object in a Start location, and a second character moves it to an End location. The last sentence states that the main character believes the object is in some (omitted) location. There are 16 versions of each item (192 passages) varied across 4 dimensions: i) Knowledge State: whether the main character knows (True Belief) or not (False Belief) that the object has changed location; whether (ii) the First Mention and (iii) the most Recent Mention of a location is the Start or End location; and (iv) Knowledge Cue: whether the main character's belief is implicit or explicit ("X goes to get the book from the ___", vs "thinks the book is in the ___").

Human Responses 1156 participants from Amazon's Mechanical Turk were compensated $1 to complete a single trial. Each read a passage (except the final sentence), and on a new page, produced a single word free-response completion of the final sentence. Participants then completed two free-response attention check questions that asked for the true location of the object at the start and the end of the passage. Responses were preprocessed by lowercasing and removing punctuation, stopwords, and trailing whitespace. Participants were excluded if they were non-native English speakers (13), answered ≥ 1 attention check incorrectly (513), or answered the sentence completion with a word that was not the start or end location (17), retaining 613 trials.

LLM Responses LLM responses were operationalized as the probability assigned to each possible location (Start vs. End) conditioned on each of the passage versions. Using the Log-Odds Ratio, $log(p(Start)) - log(p(End))$, higher values indicate larger relative probabilities of the Start location. We score model responses as correct if $p(Start) > p(End)$ in False Belief trials and vice versa in True Belief Trials.

4.2. Recursive Mindreading

Materials We adapted stimuli from O'Grady et al. (2015) for U.S. participants. The stimuli comprised 4 stories, each of which had a plot involving seven levels of recursively-embedded mental representation, and seven 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 the main story. All of the stories and continuations were written in two different formats: as scripts (dialogue) and as narratives. In total there were 112 pairs of continuation passages. While the original study recorded actors reading scripts, we presented the materials in text format to both LLMs and human participants.

Human Responses We recruited 72 undergraduates who participated in the experiment online. Each read all four stories in a randomized order. After each story, they responded to 14 questions (2 conditions $\times$ 7 embedding levels); each asked which of a pair of story continuations was consistent with the main story. The format of the story and continuations (narrative vs dialogue) was fully crossed. We excluded 6 participants who scored $< 62\%$ on level 1 questions, and trials in which the participant read the story in < 65ms/word (322), or responded to the question in < 300ms (45).

LLM Responses We measured the probability assigned by LLMs to each continuation following the story. We presented all four combinations of story and question format to the LLM. Because continuations varied considerably in length and other surface features, we used PMI_{DC} to control for the probability of the continuation in the absence of the story (Holtzman et al., 2022). We operationalize the LLM's preference for one option over another as the log-odds $(log(p([A])) - log(p([B]))$, corrected with PMI_{DC}. We scored the LLM as correct if it assigned a higher probability to the consistent continuation.

4.3. Short Story Task

Materials Dodell-Feder et al. (2013) designed a set of 14 questions about Ernest Hemingway's short story *The End of Something*. The story describes an argument between a couple, culminating in their breakup. The mental lives of the characters are not explicitly described and must be inferred from their behavior. There are 5 Reading Comprehension (RC) questions; 8 Explicit Mental State Reasoning (EMSR) questions, and 1 Spontaneous Mental State Inference (SMSI) question that asks whether participants make mental state inferences when summarizing the passage.

Human Responses Human response data came from Trott & Bergen (2018). 240 participants completed a web version of the Short Story Task, in which they read *The End of Something* and then answered all 14 questions. Participants who indicated that they had read the story before were excluded, and there were 227 subjects retained after exclusions. All responses were scored by two independent coders using the rubric provided by Dodell-Feder et al. (2013). A third coder acted as tiebreaker for cases where these coders disagreed.

LLM Responses LLMs generated completions for prompts that comprised the passage and a question. Each question was presented separately. A third coder scored LLM responses and a subset of human responses in a single batch. The scores assigned to the human participant responses by this third rater were consistent with the original assigned scores across all three components, (RC: $r = 0.98$; EMSR: $r = 0.90$; SMSI: $r = 0.76$).

4.4. Strange Story Task

Materials Happé (1994) designed 24 passages in which a character says something they do not mean literally. Each story is accompanied by a comprehension question ("Was it true, what X said?") and a justification question ("Why did X say that?"). 6 non-mental control stories measured participants' general reading comprehension skill.

Human Responses We recruited 44 undergraduates who participated online. Participants saw a non-mentalistic example passage, and example responses to both question types. Participants read each passage and answered the associated questions using a free-response input. We removed 95 trials (7%) in which the participant answered the comprehension question incorrectly. Responses to the justification question were scored by two naive raters using the rubric provided by Happé (1994), with a third coder acting as tiebreaker. We excluded 16 participants for scoring $< 66\%$ on the control stories, indicating inattention.

LLM Responses We generated completions from LLMs for a prompt which consisted of the same instructions and examples that human participants saw, a passage, and the relevant question. For the justification question, the prompt additionally contained the first question along with the correct answer (i.e. "No"). LLM responses were scored in the same batch as the human responses. Coders were not aware that any of the responses had been generated by LLMs.

4.5. Indirect Request

Materials Trott & Bergen (2020) created 16 pairs of short passages, each ending with an ambiguous sentence that could be interpreted as either an indirect request or a direct speech act (e.g. "it's cold in here" could be a request to turn on a heater, or a complaint about the temperature of the room). In each passage, the participant learns about an obstacle that would prevent fulfilment of the potential request (e.g. the heater being broken). The authors manipulated Speaker Awareness—whether the speaker was aware of the obstacle or not— and Knowledge Cue: whether the speaker's knowledge about the obstacle was indicated explicitly ("Jonathan doesn't know about the broken heater") or implicitly (Jonathan being absent when the heater breaks).

Human Responses Human response data came from Trott & Bergen (2020) Experiment 2. 69 participants from Amazon Mechanical Turk read 8 passages. Condition (Speaker Aware vs Speaker Unaware) was randomized within subjects. After each passage, participants were asked: "Is X making a request?" and responded "Yes" or "No."

LLM Responses We presented each version of each passage to GPT-3 followed by the critical question "Do you think [the speaker] is making a request?" and measured the probability assigned by the model to the tokens "Yes" and "No." We calculate the log odds ratio $log(p(Yes)) - log(p(No))$ and score answers as correct if this is positive when the speaker is unaware of the obstacle, and negative when the speaker is unaware.

4.6. Scalar Implicature

Materials We designed 40 novel passage templates based on the 6 items in Goodman & Stuhlmüller (2013). The first section of each passage introduces three objects that almost always have some property (e.g. "David orders 3 pizzas that almost always have cheese in the crust."). The next section contains an utterance about the speaker's knowledge state ("David says: 'I have looked at [a] of the 3 pizzas. [n] of the pizzas have cheese in the crust.", where $1 \leq a \leq 3$, $n =$"Some" in Experiment 1, and $1 \leq n \leq a$ in Experiment 2. After each of the two passage sections, participants are asked "How many of the 3 pizzas do you think have cheese in the crust? (0, 1, 2, or 3)", probing participants' beliefs both before and after the utterance. A third question asks if the speaker knows how many objects have the property ("Yes" or "No").

Human Responses We randomly assigned 242 undergraduate student participants to either Experiment 1 (126) or Experiment 2 (116).[2] For each question, participants were instructed to divide "\$100" among the options, betting to indicate their confidence in each option. Participants completed 3 trials in E1 (each with different values of *a*) and 6 trials in E2 (with all possible combinations of *a* and *n*). Following Goodman & Stuhlmüller (2013), we excluded 410 trials (143 in E1, 247 in E2) in which the knowledge judgement was less than 70 in the expected direction (i.e. $< \$70$ on "Yes" when $a = 3$; $< \$70$ on "No" when $a < 3$). We measured accuracy by testing whether the relationships between bets before and after the speaker's utterance reflect the fact that a scalar implicature should only be drawn when the speaker has complete access (see Appendix C).

LLM Responses For each question, we constructed a prompt consisting of the relevant sections of the story, followed by the question (marked by 'Q:'), then by an answer prompt, 'A:'. We found the probability assigned by the model to each response option (0, 1, 2, and 3), normalized by the total probability assigned to all response options. We did not use the knowledge check filtering criterion for model responses as this would amount to removing entire items.

5. Results

For each task we ask how GPT-3 and human accuracy compare by testing whether data source (human vs GPT-3) improves the fit of a regression model predicting accuracy. We also test whether log model scale predicts accuracy across four base GPT-3 models (*ada* to *davinci*). We exclude *text-davinci-002* from the scaling analysis to avoid conflating contributions of scale and training data. For 4 of the tasks (FB, RM, IR, SI), we additionally ask whether mental state variables have significant effects on GPT-3 metrics, analogous to the statistical analyses from the original experiments. Finally, we use hierarchical model comparisons to test whether the effects of these variables on humans are robust to controlling for LLM predictions.

5.1. False Belief Task

GPT-3 accuracy was 74%, significantly below the human mean of 83% ($\chi^2(1) = 6.97, p = .008$, see Figure 2). Accuracy increased with model size from *ada* (51%) to *davinci* (60%) ($\chi^2(1) = 7.51, p = .006$, see Figure 3).

Knowledge State—whether the character knew that the

[2] We originally ran this study on Mechanical Turk. An unusually high exclusion rate of 70% indicated unreliable data and we re-ran the study with undergraduate students.

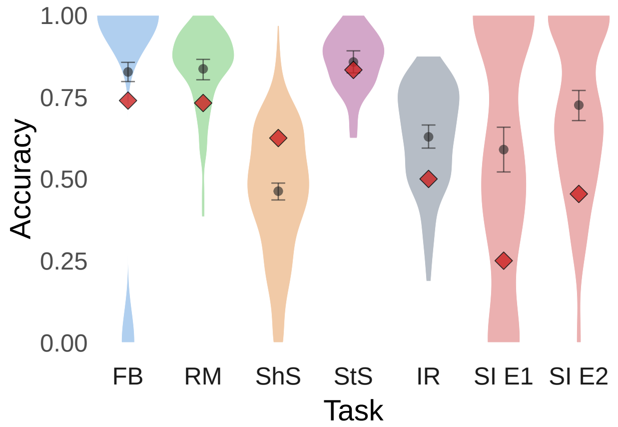

Figure 2. Distribution of human accuracy by participant (grey circles with 95% bootstrap confidence intervals) compared to mean GPT-3 *text-davinci-002* accuracy (red diamonds). GPT-3 accuracy was not significantly different from human accuracy across 3 tasks (ShS, StS, IR), but was significantly lower in others (FB, RM, SI).

object had been moved—had a significant effect on the log-odds that GPT-3 assigned to each location ($\chi^2(1) = 18.6, p < .001$). Concretely, GPT-3 assigned a higher probability to the true (end) location of the object when the character was in a position to observe the object having moved to that location. Human comprehenders also showed an effect of Knowledge State on the likelihood that they completed a probe sentence with the end location ($\chi^2(1) = 31.7, p < .001$). Crucially, this effect on human comprehenders was robust to controlling for the predictions of GPT-3 ($\chi^2(1) = 30.4, p < .001$), suggesting that Knowledge State influenced human responses in a way that was not captured by the LLM.

5.2. Recursive Mindreading

GPT-3's mean accuracy on mental questions was 73%, significantly lower than the human mean of 85% ($\chi^2(1) = 9.12, p = .003$). GPT-3 was in the 16th percentile of human accuracy scores, aggregated by participant. Model accuracy increased slightly with scale, from *ada* (63%) to *davinci* (65%) ($z = 3.06, p = .002$, see Figure 3).

Human accuracy on mental questions was significantly above chance up to 7 levels of embedding ($z = 5.56, p < .001$), though there was a negative effect of embedding level ($z = -4.12, p < .001$). GPT-3 accuracy on mental questions decreased after level 4 and was not significantly different from chance beyond level 5 ($z = -0.06, p = 0.949$). However, there was no such drop for control questions (see Figure 4). The difference in log-probability assigned to correct and incorrect continuations did not significantly predict human accuracy ($z = 1.78, p = 0.075$), indicating that human comprehenders are using different information from the LLM to select responses. Human accuracy was significantly above chance at all embedding levels when controlling for GPT-3 log probabilities (all p values < 0.022).

5.3. Short Story Task

GPT-3 scored 100% on both the RC and SMSI questions, and 62% on EMSR. Mean human performance was 83%, 42%, and 46% for these components respectively. GPT-3's EMSR score was better than 73% of human subjects, but not significantly greater than the human mean ($\chi^2(1) = 0.997, p = .318$). It was not possible to perform scaling analysis on the StS as the story did not fit in the context window of smaller models. In order to test whether GPT-3's EMSR performance could be attributable to general comprehension performance, we performed a follow-up analysis on the 55 participants (25%) who matched GPT-3's Reading Comprehension score. Mean EMSR performance among this group was 57% and GPT-3 fell in the 50th percentile of this distribution.

5.4. Strange Story Task

GPT-3 *text-davinci-002*'s mean accuracy on critical trials was 83%, below mean human accuracy of 86%, however the difference was not significant ($\chi^2(1) = 0.119, p = .73$). GPT-3 performed better than 36% of human participants. Model performance increased monotonically with scale, from *ada* (18%) to *davinci* (75%) ($t(71) = 6.02, p < .001$, see Figure 3). GPT-3's accuracy on the control questions (83%) was very similar to the mean accuracy of retained participants (80%).

5.5. Indirect Request

GPT-3 interpreted all statements as requests (*i.e.* it assigned a higher probability to 'Yes' vs 'No'), yielding an accuracy of 50%. Human mean accuracy was 62% and there was no significant difference in accuracy between Human and LLM responses ($\chi^2(1) = 0.666, p = .414$). GPT-3's accuracy placed it in the 11th percentile of humans, aggregated by subject. No consistent relationship held between model scale and performance, with all smaller models performing at around 50% accuracy ($z = -1.13, p = .260$).

There was a significant effect of Speaker Awareness on human responses ($\chi^2(1) = 23.557, p < .001$). Human participants were less likely to interpret a statement as a request if the speaker was aware of an obstacle preventing the request's fulfillment. There was no significant effect of Speaker Awareness on the log-odds ratio between the probabilities assigned to 'Yes' and 'No' by GPT-3, suggesting that the model was not sensitive to this information when interpreting the request ($\chi^2(1) = 1.856, p = .173$).

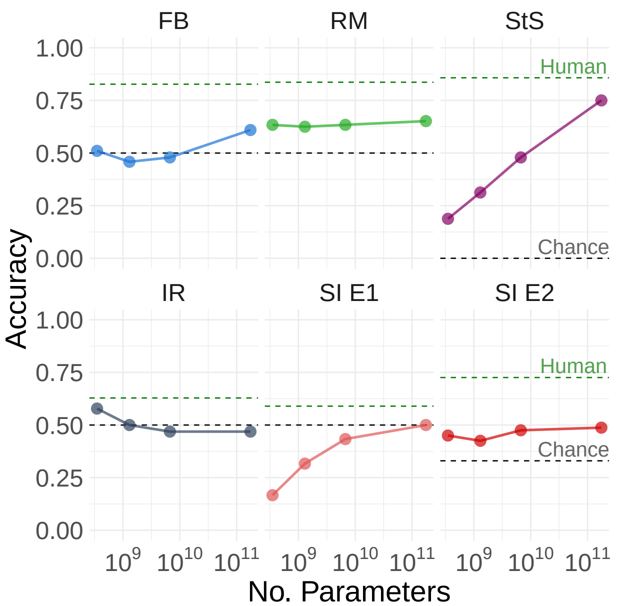

Figure 3. ToM task accuracy vs model scale across four base GPT-3 models (*ada*, *babbage*, *curie*, and *davinci*). FB, StS, RM, and SI E1 show positive scaling, with higher-parameter models achieving increased accuracy. IR and SI E2 show relatively flat scaling, with no significant increase in accuracy for larger models. GPT-3 *text-davinci-002* was excluded from the scaling analysis.

5.6. Scalar Implicature

In Experiment 1, GPT-3 accuracy was 25%, significantly lower than the human mean of 56% ($\chi^2(1) = 28.0, p < .001$), and outperforming only 19% of human participants. Accuracy increased with scale from *ada* (17%) to *davinci* (50%) ($z = 3.93, p < .001$, see Figure 3). In line with the original results, human participants bet significantly more on 2 vs 3 when access = 3 ($t(1) = -13.07, p < .001$). However, in contrast with the original results we also find this effect when the speaker has incomplete access and the implicature ought to be cancelled ($t(1) = -5.881, p < .001$). This could be due to the ambiguity of whether 'some' refers to some of the observed objects or some of the total set of objects (Zhang et al., 2023). GPT-3's predictions were inconsistent with the rational model in both cases. It assigned a *higher* probability to 3 vs 2 in the complete access condition— inconsistent with the scalar implicature—and a *lower* probability to 3 vs 2 in the incomplete access conditions—inconsistent with cancelling the implicature.

In Experiment 2, GPT-3 achieved 45% accuracy, placing it in the 12th percentile of the human distribution and significantly below the human mean of 72% ($\chi^2(1) = 37.0, p < .001$). There was no significant relationship between model scale and performance ($z = 1.04, p = .300$). GPT-3 failed to show the scalar implicature effect in the complete access condition (see Figure 5). The model assigned a higher probability to 2 vs 1 when $n = 1$ ($t(1) = 29.3, p < .001$), and there was no difference between $p(2)$ and $p(3)$ when $n = 2$ ($t(1) = 0.39, p < .697$). The probabilities reflected cancellation of the implicature in all of the incomplete access conditions: $p(2) \geq p(1)$ when $a = 1$ and $n = 1$ ($t(1) = 216, p < .001$) and when $a = 2$ and $n = 1$ ($t(1) = 71.4, p < .001$), and $p(3) \geq p(2)$ when $a = 2$ and $n = 2$ ($t(1) = 13.256, p < .001$). The pattern of human responses replicated all of the planned comparison effects from Goodman & Stuhlmüller (2013), and all effects persisted when controlling for GPT-3 predictions.

6. Discussion

We compared GPT-3 and human performance on EPITOME: a battery of 6 ToM experiments. LLM performance varied considerably by task, achieving parity with humans in some cases and failing to show sensitivity to mental states at all in others. There was also significant variation in human performance within and between tasks—with close to baseline performance on SI E1 and IR—highlighting the importance of establishing human baselines to contextualise LLM performance. While previous work has shown isolated successes (Kosinski, 2023) and failures (Sap et al., 2022; Ullman, 2023) of LLMs at specific tasks, the breadth of tasks presented here provide a more systematic basis for understanding model performance on diverse aspects of ToM. We make the code, materials, and human data from EPITOME available to facilitate further research into differences in ToM between humans and LLMs.

In some respects, GPT-3 showed striking sensitivity to mental state information. For three of the tasks (ShS, StS, and IR), GPT-3 accuracy was not significantly different from the human mean. For the ShS and StS tasks, this means that GPT-3's free-text explanations of character's mental states were rated as equivalent to humans' by naive raters. In others tasks, GPT-3 was sensitive to mental states, with above chance performance in RM up to 5 levels of embedding, and significant effects of knowledge state in FB.

However, other aspects of the current results suggest crucial differences between human and LLM performance. First, GPT-3 was insensitive to knowledge state in the IR task, interpreting every statement as a request. Second, GPT-3 failed to show effects of speaker knowledge in SI (although poor human performance indicates the wording of E1 may be ambiguous). Third, GPT-3 failed to perform above chance at Recursive Mindreading beyond 5 levels of embedding, suggesting that distributional information may be insufficient for more complex mentalizing behavior. Finally, across 4 tasks (FB, RM, IR, and SI) there were residual effects of mental state variables on human responses after controlling for GPT-3 predictions, indicating that humans are sensitive to mental state information in a way that is not captured by models.

Consistent with the hypothesis that an LLM's performance is positively correlated with its size (Kaplan et al., 2020), we found positive scale-accuracy relationships for 4 tasks (FB, RM, and StS, SI E1). However, IR and SI E2 showed flat or even negative scaling. This could indicate that models will require information beyond distributional statistics to achieve human parity.

GPT-3 performed worst on IR and SI, the two tasks requiring pragmatic inferences from mental state information. These showed the largest gaps in accuracy, insensitivity to mental states, and the flat scaling relationships noted above. Given existing work showing LLM sensitivity to pragmatic inference (Hu et al., 2022), this trend could indicate a specific difficulty for LLMs in varying pragmatic inferences on the basis of mental state information. These tasks require a complex multi-step process of sampling, maintaining, and deploying mental-state information (Trott & Bergen, 2020), increasing the chances of information loss.

The results also bear on the origins of mentalizing abilities in humans. LLMs' sensitivity to mental state variables suggests that domain-general learning mechanisms and exposure to language could be sufficient to produce ToM-consistent behavior. But LLMs also performed relatively better at non-mental control questions (in RM and ShS). This could imply that distributional information is *less* useful for predicting human performance in mentalistic than non-mentalistic tasks, supporting the view that humans recruit other resources for mental reasoning specifically.

6.1. Limitations

The current work has several important limitations. First, the tasks were designed to test specific hypotheses about human comprehenders and may not be well suited to comparing mentalizing performance of humans and LLMs. The performance score for the SI tasks, for instance, was not proposed by the original authors and may not reliably track mentalizing ability. Second, some aspects of ToM are not measured by the tasks in this inventory, including recognizing intentions, perspective taking, and inferring emotions from visual cues (Beaudoin et al., 2020). Third, several tasks require abilities beyond mentalizing (Bloom & German, 2000), for instance infrequent vocabulary (ShS) and probabilistic reasoning (SI). Fourth, many differences between LLMs and human comprehenders complicate comparisons between them. In particular, LLMs are exposed to orders of magnitude more words than humans in a lifetime (Warstadt & Bowman, 2022), which undermines claims that LLM performance indicates the practical viability of distributional learning in humans. Fifth, although we tried to closely align experimental procedures between LLMs and humans, there are inevitably differences. For instance, while humans could not look back at context passages, LLMs can attend to any previously presented token in their context window. Finally, although stimuli for 2 tasks were novel (FB and SI), stimuli for other tasks could theoretically have appeared in GPT-3's training data. This exposure could artificially inflate LLM performance, and so results might not generalize to novel examples.

6.2. Does the LLM have a Theory of Mind?

Do the results suggest that LLMs have ToM-like abilities? One interpretation argues that these tasks, which are used to measure mentalizing in humans, should be equally persuasive for artificial agents (Hagendorff, 2023; Schwitzgebel, 2013; Y Arcas, 2022). On this view, LLMs demonstrably learn to implicitly represent mental states to some degree, and we should attribute ToM-like abilities to them insofar as it helps to explain their behavior (Dennett, 1978; Sahlgren & Carlsson, 2021). An alternative view proposes that we should deny *a priori* that LLMs can mentalize, due to their lack of grounding and social interaction (Bender & Koller, 2020; Searle, 1980). On this view, successful LLM performance undermines the validity of the tasks themselves, revealing unidentified confounds that allow success in the absence of the relevant ability (Niven & Kao, 2019; Raji et al., 2021). While some argue these tests can be valid for humans in a way that they are not for LLMs (Mitchell & Krakauer, 2023; Ullman, 2023), it is unclear how well these arguments apply in an unsupervised, zero-shot setting, where models are not trained on specific dataset artifacts. Moreover, growing evidence suggests that humans are also sensitive to distributional information (Michaelov et al., 2022; Schrimpf et al., 2021) and therefore could be exploiting the same statistical confounds in materials.

An analogous debate revolves around attributing ToM to non-human animals on the basis of behavioral evidence. Chimpanzees produce behavior that is consistent with them representing mental states, (Krupenye et al., 2016; Krupenye & Call, 2019), but can also be explained by low-level, domain-general mechanisms operating on observable behavioral regularities (Heyes, 2014; Penn & Povinelli, 2007). One integrative proposal to resolve this debate is to test behavior in a wide variety of conditions: if mentalizing explanations predict behavior in diverse situations they may be more useful than equivalent deflationary accounts (Halina, 2015). The current work is intended in this vein and presents mixed evidence. While GPT-3 performance is impressive and humanlike in several ToM tasks, it lags behind humans in others and makes errors that would be surprising for an agent with a general and robust theory of mind. Even if GPT-3s don't appear to represent mental states of others in a general sense, continued work along the lines described here may uncover such developments if and when they emerge.

References

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

Beaudoin, C., Leblanc, É., Gagner, C., and Beauchamp, M. H. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02905.

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

Bender, E. M. and Koller, A. 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, 2020. doi: 10.18653/v1/2020.acl-main.463.

Binz, M. and Schulz, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

Bloom, P. and German, T. P. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):B25–31, October 2000. ISSN 0010-0277. doi: 10.1016/s0010-0277(00)00096-2.

Brown, J. R., Donelan-McCall, N., and Dunn, J. Why Talk about Mental States? The Significance of Children's Conversations with Friends, Siblings, and Mothers. *Child Development*, 67(3):836–849, 1996. ISSN 1467-8624. doi: 10.1111/j.1467-8624.1996.tb01767.x.

de Villiers, J. G. and de Villiers, P. A. The Role of Language in Theory of Mind Development. *Topics in Language Disorders*, 34(4):313–328, 2014. ISSN 0271-8294. doi: 10.1097/TLD.0000000000000037.

Dennett, D. C. Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, 1(4):568–570, December 1978. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X00076664.

Dhelim, S., Ning, H., Farha, F., Chen, L., Atzori, L., and Daneshmand, M. IoT-Enabled Social Relationships Meet Artificial Social Intelligence. *IEEE Internet of Things Journal*, 8(24):17817–17828, December 2021. ISSN 2327-4662, 2372-2541. doi: 10.1109/JIOT.2021.3081556.

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., and Hooker, C. I. Using Fiction to Assess Mental State Understanding: A New Task for Assessing Theory of Mind in Adults. *PLOS ONE*, 8(11):e81279, November 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0081279.

Frith, C. D. and Frith, U. Mechanisms of Social Cognition. *Annual Review of Psychology*, 63(1):287–313, 2012. doi: 10.1146/annurev-psych-120710-100449.

Gernsbacher, M. A. and Yergeau, M. Empirical Failures of the Claim That Autistic People Lack a Theory of Mind. *Archives of scientific psychology*, 7(1):102–118, 2019. ISSN 2169-3269. doi: 10.1037/arc0000067.

Goodman, N. D. and Stuhlmüller, A. Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013. ISSN 1756-8765. doi: 10.1111/tops.12007.

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

Hagendorff, T. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods, April 2023.

Hale, C. M. and Tager-Flusberg, H. The influence of language on theory of mind: A training study. *Developmental Science*, 6(3):346–359, 2003. ISSN 1467-7687. doi: 10.1111/1467-7687.00289.

Halina, M. There is no special problem of mindreading in nonhuman animals. *Philosophy of Science*, 82(3):473–490, 2015. doi: 10.1086/681627.

Happé, F. G. E. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2):129–154, April 1994. ISSN 0162-3257, 1573-3432. doi: 10.1007/BF02172093.

Harris, P. L. Conversation, Pretense, and Theory of Mind. In *Why Language Matters for Theory of Mind*, pp. 70–83. Oxford University Press, New York, NY, US, 2005. ISBN 978-0-19-515991-2. doi: 10.1093/acprof:oso/9780195159912.003.0004.

Hayward, E. O. and Homer, B. D. Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, 35(3):454–462, 2017. ISSN 2044-835X. doi: 10.1111/bjdp.12186.

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

Holtzman, A., West, P., Shwartz, V., Choi, Y., and Zettlemoyer, L. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, November 2022.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., and Gibson, E. A fine-grained comparison of pragmatic language understanding in humans and language models, December 2022.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., and Moffitt, T. E. Origins of individual differences in theory of mind: From nature to nurture? *Child development*, 76(2):356–370, 2005. doi: 10.1111/j.1467-8624.2005.00850_a.x.

Johnson, S. and Iziev, N. AI is mastering language. should we trust what it says. *The New York Times*, 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*, 2020.

Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models, March 2023.

Krupenye, C. and Call, J. Theory of mind in animals: Current and future directions. *WIREs Cognitive Science*, 10 (6):e1503, 2019. ISSN 1939-5086. doi: 10.1002/wcs.1503.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. Great apes anticipate that other individuals will act according to false beliefs. *Science (New York, N.Y.)*, 354 (6308):110–114, October 2016. doi: 10.1126/science.aaf8110.

Langley, C., Cirstea, B. I., Cuzzolin, F., and Sahakian, B. J. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Frontiers in Artificial Intelligence*, 5, 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.778852.

Le, M., Boureau, Y.-L., and Nickel, M. 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)*, pp. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598.

Michaelov, J. A., Coulson, S., and Bergen, B. K. So Cloze yet so Far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, 2022. doi: 10.1109/TCDS.2022.3176783.

Milligan, K., Astington, J. W., and Dack, L. A. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2):622–646, 2007. ISSN 0009-3920. doi: 10.1111/j.1467-8624.2007.01018.x.

Mitchell, M. and Krakauer, D. C. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, March 2023. doi: 10.1073/pnas.2215907120.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective, February 2023.

Niven, T. and Kao, H.-Y. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459.

O'Grady, C., Kliesch, C., Smith, K., and Scott-Phillips, T. C. The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, 36(4):313–322, July 2015. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2015.01.004.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, December 2022.

Penn, D. C. and Povinelli, D. J. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):731–744, January 2007. doi: 10.1098/rstb.2006.2023.

Perner, J., Leekam, S. R., and Wimmer, H. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137, 1987. doi: 10.1111/j.2044-835X.1987.tb01048.x.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, December 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., and Botvinick, M. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4218–4227. PMLR, July 2018.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Sahlgren, M. and Carlsson, F. The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *Frontiers in Artificial Intelligence*, 4, 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.682578.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, October 2022.

Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., and Chapman, M. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, January 2019. ISSN 0921-7126. doi: 10.3233/AIC-190615.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., and Adolphs, R. Deconstructing and Reconstructing Theory of Mind. *Trends in cognitive sciences*, 19(2):65–72, February 2015. ISSN 1364-6613. doi: 10.1016/j.tics.2014.11.007.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. doi: 10.1073/pnas.2105646118.

Schwitzgebel, E. A dispositional approach to attitudes: Thinking outside of the belief box. *New essays on belief: Constitution, content and structure*, pp. 75–99, 2013. doi: 10.1057/9781137026521_5.

Searle, J. R. Minds, Brains, and Programs. *Behavioral and brain sciences*, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.

Sebanz, N., Bekkering, H., and Knoblich, G. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, February 2006. ISSN 13646613. doi: 10.1016/j.tics.2005.12.009.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models, May 2023.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pp. 194–205, 2021. doi: 10.1145/3442381.3450097.

Shevlin, H. Uncanny Believers: Uncanny believers: Chatbots, beliefs, and folk psychology. https://henryshevlin.com/wp-content/uploads/2021/11/Uncanny-Believers.pdf, under review.

Sperber, D. and Wilson, D. Pragmatics, Modularity and Mind-reading. *Mind & Language*, 17(1-2):3–23, 2002. ISSN 1468-0017. doi: 10.1111/1468-0017.00186.

Surian, L., Caldi, S., and Sperber, D. Attribution of beliefs by 13-month-old infants. *Psychological science*, 18(7): 580–586, 2007. doi: 10.1111/j.1467-9280.2007.01943.x.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5): 675–691, 2005. doi: 10.1017/S0140525X05000129.

Trott, S. and Bergen, B. Individual Differences in Mentalizing Capacity Predict Indirect Request Comprehension. *Discourse Processes*, 56(8):675–707, December 2018. ISSN 0163-853X, 1532-6950. doi: 10.1080/0163853X.2018.1548219.

Trott, S. and Bergen, B. When Do Comprehenders Mentalize for Pragmatic Inference? *Discourse Processes*, 57 (10):900–920, November 2020. ISSN 0163-853X, 1532-6950. doi: 10.1080/0163853X.2020.1822709.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. Do Large Language Models know what humans know?, May 2023.

Ullman, T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. 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*, pp. 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566.

Warnell, K. R. and Redcay, E. Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, 191:103997, October 2019. ISSN 0010-0277. doi: 10.1016/j.cognition.2019.06.009.

Warstadt, A. and Bowman, S. R. What Artificial Neural Networks Can Tell Us About Human Language Acquisition, August 2022.

Wellman, H. M., Cross, D., and Watson, J. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, 72(3):655–684, 2001. ISSN 1467-8624. doi: 10.1111/1467-8624.00304.

Wimmer, H. and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, January 1983. ISSN 0010-0277. doi: 10.1016/0010-0277(83)90004-5.

Y Arcas, B. A. Do Large Language Models Understand Us? *Daedalus*, 151(2):183–197, May 2022. ISSN 0011-5266, 1548-6192. doi: 10.1162/daed_a_01909.

Zhang, Z., Bergen, L., Paunov, A., Ryskin, R., and Gibson, E. Scalar Implicature is Sensitive to Contextual Alternatives. *Cognitive Science*, 47(2):e13238, 2023. ISSN 1551-6709. doi: 10.1111/cogs.13238.

A. Supplementary Results

A.1. Accuracy Summary Table

Table 1. Accuracy of models and human participants across tasks.

Model	FB	RM	ShS	StS	IR	SI E1	SI E2
ada	0.51	0.63		0.19	0.58	0.17	0.45
babbage	0.46	0.62		0.31	0.50	0.32	0.42
curie	0.48	0.63		0.48	0.47	0.43	0.47
davinci	0.61	0.65		0.75	0.47	0.50	0.49
t-d-002	0.74	0.73	**0.62**	0.83	0.50	0.25	0.45
Human	**0.83**	**0.84**	0.46	**0.86**	**0.63**	**0.59**	**0.73**

A.2. Supplementary Figures

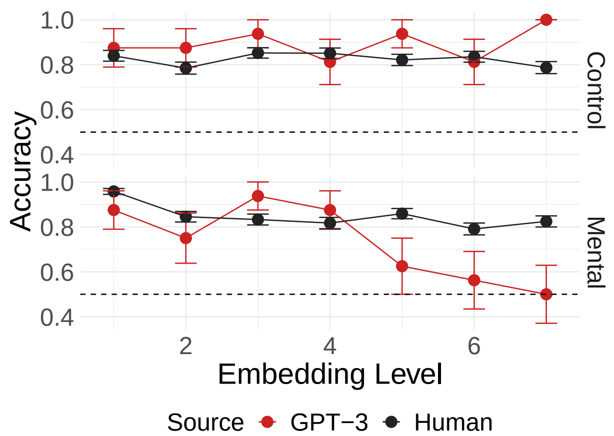

Figure 4. Recursive mindreading accuracy by embedding level and question type for GPT-3 and human participants.

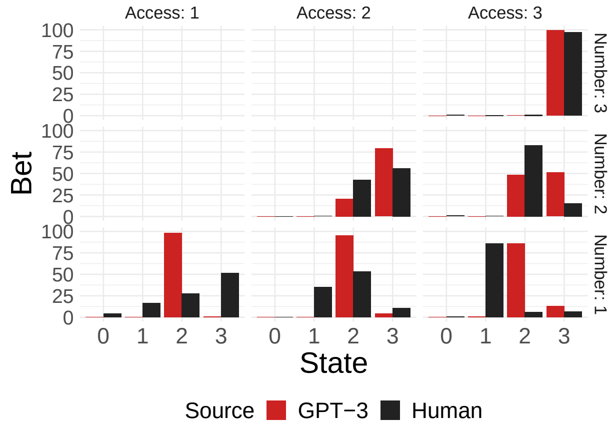

Figure 5. GPT-3 and human bets on each state (the number of objects that meet the property) for all conditions in SI E2.

B. Example Materials

B.1. False Belief

Context: Sean is reading a book. When he is done, he puts the book in the box and picks up a sweater from the basket. Then, Anna comes into the room. Sean leaves to get something to eat in the kitchen. While he is away, Anna moves the book from the box to the basket. Sean comes back into the room and wants to read more of his book.

Question: Sean thinks the book is in the...

Response type: Free-text completion

Scoring: 1 — *box*, 0 — *basket* or other

B.2. Recursive Mindreading

Context: One evening, Megan finds out that her sister Lauren wants to go out with a boy in her Biology class, Stephen. Megan tells Lauren that Stephen used to be best friends with a boy called Chris, who is now Megan's best friend. Lauren tells Megan that she saw Stephen smiling and flirting with their cousin, Elaine, and so she thinks Stephen might want to go out with Elaine. Because Lauren thinks Stephen likes someone else, she is too nervous to ask him out.

Megan talks to Elaine at school and finds out that Elaine actually wants to go out with Bernard, whom Megan knows from the school play. Megan learns that Elaine and Bernard are next-door neighbours, and that Bernard thinks that Elaine doesn't know him well enough to date. Elaine tells Megan that Stephen knows how Elaine feels about Bernard and how Bernard feels about Elaine.

Megan later talks to her friend Chris about the situation, realizing that if Lauren knew about Elaine's situation, and knew that Stephen knows about it too, Lauren would realize that Stephen doesn't want to go out with Elaine, and might work up the courage to ask him out. Megan plans to tell Lauren about everything that evening.

Question: Which of the following sentences do you think is consistent with the story you read earlier?

A: Stephen knows that Elaine knows that Bernard feels she doesn't know him well enough to date.

B: Stephen doesn't know that Elaine knows that Bernard feels she doesn't know him well enough to date

Response type: 2AFC

Scoring: 1 — A, 0 — B

B.3. Short Stories

Context: *The End of Something*, by Ernest Hemingway

Question: Why does Nick say to Marjorie, "You know everything"?

Response type: Free-text

Scoring: **2** – He's being sarcastic/cynical/intentionally mean AND wants to get Marjorie upset/sad/mad/annoyed; provoke a fight or provoke Marjorie so that she breaks up with him so he can blame the breakup on her; **1** – He's unhappy with the relationship; wants to end the relationship; He's annoyed/nervous about the situation/impending breakup; he's being sarcastic/cynical (no mention of consequences, i.e. what Marjorie's reaction will be); **0** – He thinks Marjorie is a know-it-all; He's just being mean; He's a mean person

B.4. Strange Stories

Context: One day Aunt Jane came to visit Peter. Now Peter loves his aunt very much, but today she is wearing a new hat; a new hat which Peter thinks is very ugly indeed. Peter thinks his aunt looks silly in it, and much nicer in her old hat. But when Aunt Jane asks Peter, "How do you like my new hat?", Peter says, "Oh, its very nice".

Question: Why does Peter say that?

Response type: Free-text completion

Scoring: **2** — reference to white lie or wanting to spare her feelings; some implication that this is for aunt's benefit rather than just for his, desire to avoid rudeness or insult; **1** — reference to trait (he's a nice boy) or relationship (he likes his aunt); purely motivational (so she won't shout at him) with no reference to aunt's thoughts or feelings; incomplete explanation (he's lying, he's pretending); **0** — reference to irrelevant or incorrect facts ⁄ feelings (he likes the hat, he wants to trick her)

B.5. Indirect Request

Stimulus: You and your friend Jonathan are taking a road trip. You began in California, and are now passing through Michigan. It's almost winter, so it's very cold outside - especially for Southern California dwellers like you and Jonathan. You see that you're almost out of gas, so you stop at a gas station in a small town.

You fill up the tank, and then the two of you go inside the gas station to buy some water and snacks. When you return to the car and start up the engine, you and Jonathan both notice with some dismay a blinking light, which indicates that the car's heating system is broken. You both bundle up.

As you leave the station, Jonathan shivers in his seat. He turns to you and says, "Man, it's really cold in here."

Question: Do you think he is making a request?

Response type: 2AFC

Scoring: 1 — no, 0 — yes

B.6. Scalar Implicature

Context: Pizzas from Luigi's Pizzeria almost always have cheese in the crust. David ordered 3 pizzas from Luigi's Pizzeria. David tells you on the phone: "I have looked at 3 of the 3 pizzas. 2 of the pizzas have cheese in the crust."

Question: How many of the 3 pizzas do you think have cheese in the crust?

Response type: Probability Distribution

Scoring: $\Delta bet3 < 0$

C. Scalar Implicature Scoring Criteria

We designed scoring rubrics for the SI tasks based on Δbet: the difference between bets on an outcome before and after the utterance. The scoring attempts to capture the intuition that scalar implicatures should only be drawn where the speaker has full access to the class of objects.

C.1. Experiment 1

For Experiment 1, we check that bets on 3 decrease when $access = 3$ (scalar implicature) and do not decrease when $access < 2$ (implicature cancelled).

Access	Criterion
3	$\Delta bet3 > 0$
≤ 2	$\Delta bet3 <= 0$

Table 2. Scoring criteria for Scalar Implicature Experiment 1.

C.2. Experiment 2

In Experiment 2, the speaker indicates a specific number of objects that have a given property. When $access = 3$, we expect the speaker to draw the scalar implicature and decrease bets on states $> n$. When $access \leq 2$ and $n = a$, the scalar implicature is cancelled, so bets on 3 ought not to decrease. When $access = 2$ and $n = 1$, the speaker can draw the partial implicature that fewer than 3 objects meet the condition.

Access	N	Criterion
3	3	$\Delta bet3 > 0$
3	2	$\Delta bet3 < 0$
3	1	$\Delta bet3 < 0$ and $\Delta bet2 < 0$
2	2	$\Delta bet2 > 0$ and $\Delta bet3 \geq 0$
2	1	$\Delta bet2 \geq 0$ and $\Delta bet3 < 0$
1	1	$\Delta bet2 \geq 0$ and $\Delta bet3 \geq 0$

Table 3. Scoring criteria for Scalar Implicature Experiment 2. $\Delta bet3$ and $\Delta bet2$ represent the change in bets on 3 and 2, respectively, between the prior and updated estimates.