Comparing Humans and Large Language Models on an Experimental Protocol Inventory for Theory of Mind Evaluation (EPITOME)

Cameron R. Jones* and **Sean Trott** and **Benjamin Bergen**
◇Department of Cognitive Science, UC San Diego
9500 Gilman Dr, La Jolla, CA 92093, USA

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 elicit a performance baseline from human participants for each task. We use the dataset to ask whether distributional linguistic information learned by LLMs is sufficient to explain ToM in humans. 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 human-level ToM may require resources beyond distributional information.

1 Introduction

Theory of Mind (ToM) is a broad construct encompassing a range of social behaviors from reasoning about others' mental states (internal psychological states such as beliefs and emotions) to understanding non-literal communication (Apperly, 2012; Beaudoin et al., 2020). These *mentalizing* or *mind-reading* capacities underpin social intelligence (Frith and 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 and Wilson, 2002).

There is growing interest in whether artificial intelligent (AI) agents could display ToM abilities (Johnson and 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).

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 and Koller, 2020), LLMs produce text which superficially bears many hallmarks of mentalizing (Shevlin, under review; Agüera y Arcas, 2022). However, previous studies evaluating LLM performance on ToM tasks have yielded inconsistent findings, sparking debates on LLMs' ToM capacities (Kosinski, 2023; Sap et al., 2022; Ullman, 2023). Here, we collect a battery of six diverse tasks, used to measure ToM in humans, to investigate the consistency of LLMs' ToM capabilities.

A variety of tasks have been designed to measure different facets of mentalizing (Happé, 1994; Premack and Woodruff, 1978; Wimmer and 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 socio-emotional functioning (Gernsbacher and Yergeau, 2019; Hayward and Homer, 2017). This limits the extent to which performance on a single task can be taken as evidence of ToM more generally, and underscores the need for running varied, tightly controlled experiments, each measuring distinct aspects of mentalizing. We select six tasks from the psychology literature which collectively measure a diverse set of ToM-related abilities including belief attribution, emotional reasoning, non-literal communication, and pragmatic inference.

* Corresponding author: cameron@ucsd.edu.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 803–819, 2024. https://doi.org/10.1162/tacl_a_00674
Action Editor: Dilek Hakkani-Tur. Submission batch: 2/2024; Revision batch: 3/2024; Published 6/2024.

Beyond measuring LLMs' ToM performance, these models can provide insights into debates on human ToM's evolutionary and developmental origins (Krupenye and Call, 2019; Premack and 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 and de Villiers, 2014; Hale and Tager-Flusberg, 2003). 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 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

Early work in machine ToM (Rabinowitz et al., 2018) found that neural language models could learn to coordinate actions using language (Zhu et al., 2021), but struggled with explicit mental state reasoning (Nematzadeh et al., 2018). Several recent studies have directly investigated ToM abilities in LLMs. Sap et al. (2022) evaluated GPT-3 *davinci* (Brown et al., 2020) 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 (Sap et al., 2019). While ToMi lacks a specific human baseline, GPT-3 performed poorly (60% accuracy) 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 and 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.''

More recently, Gandhi et al. (2023) used LLMs to construct a synthetic false belief benchmark from causal graphs, on which GPT-4 performs similarly to humans. Kim et al. (2023) used a similar approach to generate a belief attribution benchmark composed of naturalistic conversational dialogues. However, the best performing LLMs perform as low as 26.6% on their most challenging measures, lagging far behind a human baseline of 87.5%. Finally, 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). They found that no model performed robustly, and that all models were vulnerable to adversarial perturbations in the style of Ullman (2023).

Our contribution differs from existing studies in several ways. First, we incorporate tasks that evaluate a broader range of ToM capacities. While most studies focus primarily on belief attribution or social appropriateness, we additionally evaluate models on emotional reasoning, non-literal communication, and pragmatic reasoning from mental state inferences. Additionally, we test belief attribution up to 7 levels of embedding, and use a range of evaluation criteria (including human evaluation of free-text completions). Second, we intentionally use experimental stimuli originally designed to measure ToM in humans. Some researchers are rightly concerned that these tasks may not have the same construct validity for LLMs as they do for humans (Mitchell and Krakauer, 2023; Shapira et al., 2023; Ullman, 2023). We agree that successful performance on these tasks does not imply an agent has ToM. However, this objection is not overcome by designing novel tasks that have not been validated on human participants. The proposition that an LLM displays ToM must be supported by a range of empirical,

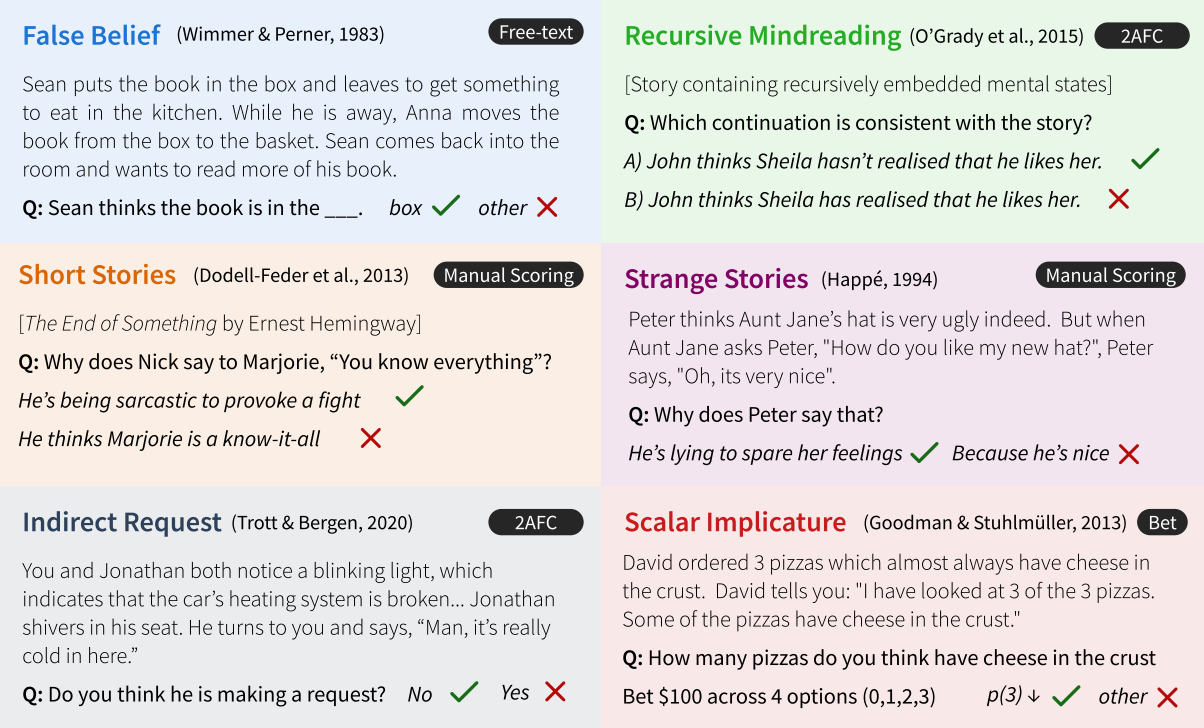

Figure 1: Truncated examples of materials from each of the 6 Theory of Mind tasks. Participants read a context passage (light text) and then answered a question using the response type indicated in the top-right of each box. Checks and crosses indicate examples of answers that would be scored as correct or incorrect (see §4 for details on how accuracy was measured in each task).

theoretical, and probably mechanistic evidence. Moreover, we believe that existing experimental stimuli have several advantages which complement contemporary work with synthetic or crowdsourced benchmarks: They have been carefully designed to control for confounds and they have been validated as measures of specific latent constructs in humans. Third, to allow direct item-level comparison between model and human performance, for each study we elicit an appropriately powered human baseline for all items and make all human data available. Fourth, we preregistered four of the six studies in order to minimize the risk of selecting materials or analyses that would bias results. Finally, to test whether distributional information learned by LLMs can *fully* account for human behavior, we run a *distributional baseline analysis* (Jones et al., 2022; Trott et al., 2023): testing whether mental state variables explain residual variance in human responses beyond the variance explained by the LLM responses.

3 The Present Study

We assemble EPITOME—a battery of six experiments designed to measure distinct aspects of ToM in humans (see Figure 1). We selected these six experiments in order to provide broad coverage of the theorized components of ToM (Beaudoin et al., 2020). 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 and 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 and 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 and Stuhlmüller, 2013).

We used this battery of tasks to address a longstanding debate about the origins of ToM in humans: namely, the extent to which language exposure is sufficient to account for human mentalizing ability. The *distributional hypothesis* (Firth, 1957; Harris, 1954) suggests that human comprehenders use statistical information about the co-occurrence frequency of words to understand language. The rapid advance of LLMs—that learn exclusively from such information—has galvanized interest in the distributional hypothesis, with many recent studies showing that LLMs can accurately predict human linguistic behavior (Chang and Bergen, 2023) and neural activity (Schrimpf et al., 2021; Michaelov et al., 2022). A more specific instantiation of this broader debate concerns the role of language exposure in human ToM development (de Villiers and de Villiers, 2014; Trott et al., 2023). We address this question by comparing the responses of LLMs and humans on EPITOME.

Crucially, in order to test the sufficiency of distributional information *per se*, we restrict our analysis to models that have not been fine-tuned on other objectives such as Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022). While RLHF is theorized to improve ToM performance (Moghaddam and Honey, 2023), it exposes models to an additional training signal, making it hard to draw inferences about the sufficiency of language exposure alone. 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 RLHF.[1] We make our code and materials available to facilitate addressing further questions, including whether RLHF improves ToM performance.

We ask four types of question: (1) Where does GPT-3 sit in the distribution of human performance? (2) How does GPT-3 performance vary with model scale? (3) Is GPT-3 sensitive to experimental variables that alter characters' mental states? (4) Does GPT-3 fully explain human mentalizing behavior? Or is there a residual effect of mental state variables on human comprehenders after controlling for distributional likelihood (as measured by GPT-3 predictions)? We pre-registered our analyses for four tasks, and provide code, data, and materials for all six.[2]

4 Methods

We accessed models through the OpenAI API. For tasks that involved generating text (ShS, StS), we set temperature to 0. For the remaining tasks, we measured the probability assigned by the model to a given string. When measuring the probability assigned to a multi-token string, we summed the log probabilities of each token. We used the same instructions and stimulus wording for both humans and LLMs. We avoided using any kind of prompt engineering with LLMs to ensure a fair comparison. We generated novel stimuli for the Scalar Implicature task and we conducted a contamination analysis following Golchin and Surdeanu (2023), which indicated that none of the other datasets were contained in the model's training data (see Appendix B).

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 publicly available data from Trott et al. (2023) for FB, and use their analysis as a model for other tasks. LLM data and analyses for all other tasks, as well as human data for RM, StS, and SI, are novel contributions. All novel human data was collected from undergraduate students, while existing data for FB, ShS and IR was collected via Amazon Mechanical Turk.

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 (e.g., ''X thinks the book is in the ___''). There are 16 versions of each item (192 passages in total) which varied across 4 dimensions: (i) Knowledge State: whether the main character knows (True Belief) or does not know (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

[1] https://platform.openai.com/docs/models/gpt-base.

[2] Available on OSF https://osf.io/sn7gj/.

or End location; and (iv) Knowledge Cue: whether the main character's belief is stated implicitly (''X goes to get the book from the ___'') or explicitly (''X 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 (e.g., ''Anne knows that Bob belives that Charlie saw...''), and seven levels of a non-mental recursive concept, such as relation (e.g., ''Stephen has Biology with Megan's sister Lauren''). 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 two-alternative forced-choice (2AFC) 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 answered fewer than 5/8 level 1 questions correctly, 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} (Holtzman et al., 2022) to control for the probability of the continuation in the absence of the story. 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 and Bergen (2018). A total of 240 participants recruited from Amazon Mechanical Turk 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 independently by two research assistants using the rubric provided by Dodell-Feder et al. (2013), with a third evaluator acting as a tiebreaker.

LLM Responses LLMs generated completions for prompts that comprised the passage and a question. Each question was presented separately. A research assistant scored LLM responses and a subset of human responses in a single batch. They were unaware that any of the responses had been generated by LLMs. In order to ensure consistent scoring, we checked the correlation between this evaluator's scores on the subset of human data and the scores assigned by the original evaluators of the human data (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 (e.g., being sarcastic or telling a white lie). Each story is accompanied by a comprehension question (''Was it true, what X said?'') and a justification question (''Why did X say that?''). Six 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. 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''). Human and LLM responses to the justification question were evaluated by two research assistants—unaware that any responses were generated by LLMs—in a single batch using the rubric provided by Happé (1994). A third evaluator acted as a tiebreaker.

4.5 Indirect Request

Materials Trott and 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 and Bergen (2020) Experiment 2. A total of 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 and 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''). The scoring criteria for the Scalar Implicature experiment can be found in Appendix A, Tables 2 and 3.

Human Responses We randomly assigned 242 undergraduate student participants to either Experiment 1 (126) or Experiment 2 (116).[3] 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 and Stuhlmüller (2013), we excluded 410 trials (143 in E1, 247 in E2) in which the knowledge judgment 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 A).

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 all 6 tasks, we asked the following 2 types of question:

(1) Is GPT-3 accuracy significantly different from humans? We ran a logistic regression:

$$\text{accuracy} \sim \textbf{data\_source}$$

where the source of the data is either human participants or GPT-3 (*text-davinci-002*).

[3]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.

(2) Does model scale predict accuracy? We ran a logistic regression:

$$\text{accuracy} \sim \textbf{log(n\_parameters)}$$

where n_parameters is the number of parameters in one of four base GPT-3 models (*ada* to *davinci*). In addition, for the experiments that manipulated a mental state variable (FB, RM, IR, SI), we asked two additional questions:

(3) Does GPT-3 show effects of mental state variables? We conducted statistical tests analogous to the tests run in the original human experiments, but using GPT-3 responses as the dependent variable. For example, in the False Belief study, we ran a linear regression

$$\text{log\_odds} \sim \textbf{knowledge\_state}$$

where log_odds is the log-odds ratio of the probabilities GPT-3 assigned to the Start and End tokens (see Section 4.1) and knowledge_state is either True Belief or False Belief. More detail on the specific variables tested in each study is contained in the relevant result sections.

(4) Does GPT-3 account for effects of mental state variables on human comprehenders? In order to test whether GPT-3 can fully account for mentalizing effects in humans, we ran linear regressions predicting human responses on the basis of mental state variables while controlling for the effect of GPT-3 predictions. For example, in the FB task we ran:

$$\text{prop\_start} \sim \text{log\_odds} + \textbf{knowledge\_state}$$

where prop_start is the proportion of human participants who responded with the Start location. If the addition of knowledge_state improves the fit of a base regression model using only log_odds, it suggests that knowledge_state explains unique variance, over and above GPT-3 predictions.

In each case, we use a Chi-Squared test to compare the fit of a full model (indicated above) with a base model (with boldface variables removed). For the fourth question, this allows us to test whether mental state variables explain significant variance in human responses once the effect of distributional likelihood (measured by GPT-3 predictions) has been controlled for. We used mixed effects models with random intercepts

Model	FB	RM	ShS	StS	IR	SI1	SI2
ada	51	63		19	58	17	45
babbage	46	62		31	50	32	42
curie	48	63		48	47	43	47
davinci	61	65		75	47	50	49
t-d-002	74	73	**62**	83	50	25	45
Human	**83**	**84**	46	**86**	**63**	**59**	**73**

Table 1: LLM and human accuracy (%) across tasks. Humans outperform models in all tasks except ShS.

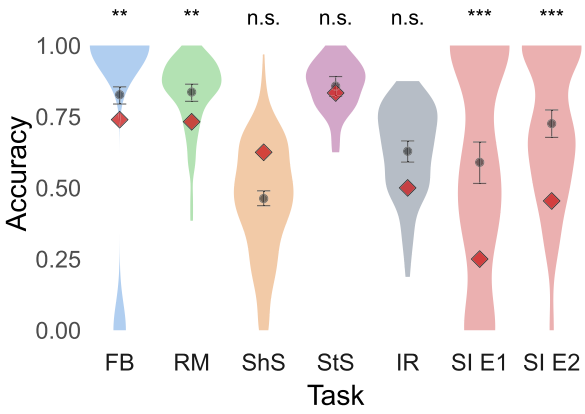

Figure 2: Distribution of human accuracy by participant (violins and gray circles with 95% CI) compared to mean GPT-3 *text-davinci-002* accuracy (red diamonds). GPT-3 accuracy was not significantly different from human accuracy across 3 tasks (ShS, StS, IR), but was significantly lower in others (FB,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

by item. Table 1 contains raw accuracies for all models and tasks.

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 4).

Knowledge State—whether the character knew that the 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 the critical sentence

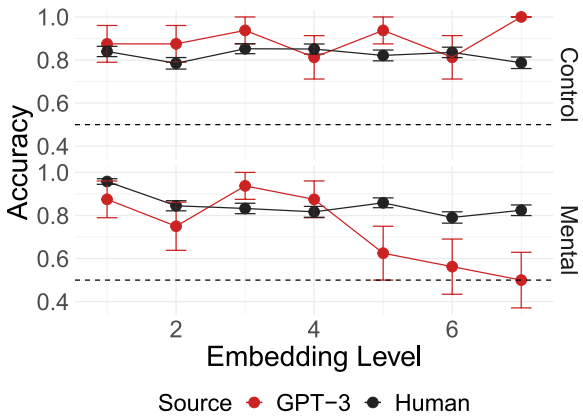

Figure 3: RM accuracy by embedding level and question type for GPT-3 and human participants. Humans maintain high accuracy across all levels in both question types. GPT-3 performance drops beyond level 5 for mental questions specifically.

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%) ($z = 3.06, p = .002$).

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 3). 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 types of 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).

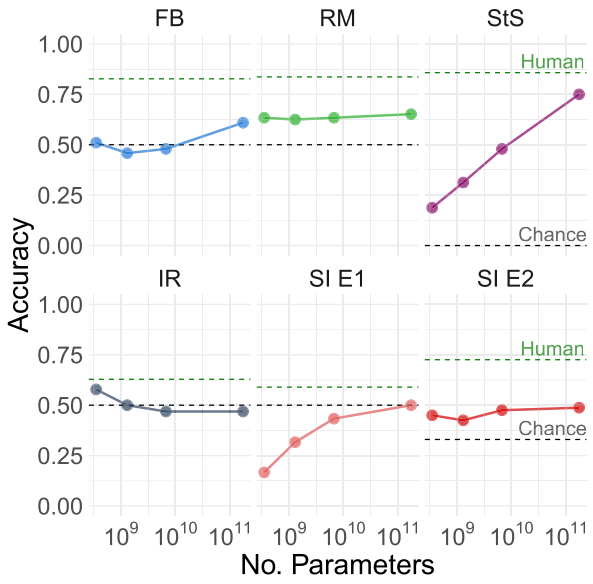

Figure 4: ToM task accuracy vs model scale across four GPT-3 models (*ada*, *babbage*, *curie*, and *davinci*). FB, StS, RM, and SI E1 show positive scaling, with higher-parameter models achieving increased accuracy. IR and SI E2 show relatively flat scaling, with no significant increase in accuracy for larger models.

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 ($\chi^2(1) = 0.997, p = .318$). 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, consistent with the theory that GPT-3's improved reading comprehension accounts for its high ESMR performance.

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%) ($t(71) = 6.02$, $p < .001$). 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 ($\chi^2(1) = 1.856, p = .173$).

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%) ($z = 3.93, p < .001$). In line with the original results, human participants make the scalar implicature that 'some' implies 'not all' when the speaker has complete access to the objects, i.e., they bet significantly more on 2 vs 3 when $a = 3$ ($t(1) = -13.07, p < .001$). However, in contrast with the original results we also find this effect when the speaker has incomplete access ($a < 3$) 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

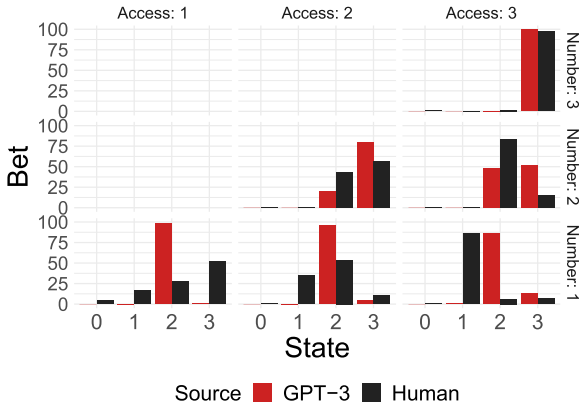

Figure 5: GPT-3 and human bets on each state (n objects with property) for all conditions in SI E2. Unlike humans, GPT-3 often fails to make a scalar implicature when access = 3.

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 (where $a = 3$, 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 and Stuhlmüller (2013), and all effects persisted when controlling for GPT-3 predictions.

6 Discussion

We assembled EPITOME—a battery of six ToM experiments that tap diverse aspects of ToM—and provided a human baseline for each task. We used the dataset to assess the extent to which distributional information learned by an LLM (GPT-3) was sufficient to reach human-level performance on these tasks. 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 characters' mental states were rated as equivalent to humans' by human evaluators. In others tasks, GPT-3 was sensitive to mental states, with above chance performance in RM up to 4 levels of embedding, and significant effects of knowledge state in FB. This provides an important demonstration that distributional information alone is sufficient to generate approximately humanlike behavior on several tasks that have been used to measure ToM in humans.

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. However, it's possible that more or better distributional data could enable progress on this task. 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. In other words, even after accounting for any variance in human responses that could be

explained by the distributional language statistics learned by GPT-3, there was still a significant effect of mental state variables on human responses. This indicates that humans are sensitive to mental state information in a way that is not captured by the model.

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 making 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 and Bergen, 2020), increasing the chances of information loss.

These results 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, for instance knowledge of infrequent words (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 and 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, transformer-based LLMs can attend to any previously presented token in their context window. In many cases, LLMs were exposed to each item independently, whereas humans completed multiple items. Sixth, we used attention checks in order to exclude participants who were not attending to the experiment, however, this could also artificially inflate our estimates of human performance. Finally, some of the datasets contain a relatively small number of items, and so non-significant effects of mental state variables could be due to a lack of power.

6.2 Does the LLM have a Theory of Mind?

Do the results suggest that GPT-3 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; Agüera 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 and 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 and 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 and Kao, 2020; Raji et al., 2021). While some argue these tests can be valid for humans in a way that they are not for LLMs (Mitchell and

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 and Call, 2019), but can also be explained by low-level, domain-general mechanisms operating on observable behavioral regularities (Heyes, 2014; Penn and 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.

Acknowledgments

We would like to thank Owen Pi, Alice Zhang, and Christy Auyeung for their help in evaluating the Strange Story and Short Story responses; James Michaelov, Tyler Chang, Seana Coulson, and Federico Rossaon for helpful discussions; and several anonymous reviewers and TACL action editors Alexander Clark and Dilek Hakkani-Tur for thoughtful comments on earlier versions of this manuscript.

References

Ian A. Apperly. 2012. What is ''theory of mind''? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65(5):825–839. https://doi.org/10.1080/17470218.2012.676055, PubMed: 22533318

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.02905, PubMed: 32010013

Marina Bedny, Alvaro Pascual-Leone, and Rebecca R. Saxe. 2009. Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences*, 106(27):11312–11317. https://doi.org/10.1073/pnas.0900010106, PubMed: 19553210

Emily M. Bender and Alexander Koller. 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*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Jane R. Brown, Nancy Donelan-McCall, and Judy Dunn. 1996. Why talk about mental states? The significance of children's conversations with friends, siblings, and mothers. *Child Development*, 67(3):836–849. https://doi.org/10.1111/j.1467-8624.1996.tb01767.x, PubMed: 8706529

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey.

Jill G. de Villiers and Peter A. de Villiers. 2014. The role of language in theory of mind development. *Topics in Language Disorders*, 34(4):313–328. https://doi.org/10.1097/TLD.0000000000000037

Daniel C. Dennett. 1978. Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, 1(4):568–570. https://doi.org/10.1017/S0140525X00076664

Sahraoui Dhelim, Huansheng Ning, Fadi Farha, Liming Chen, Luigi Atzori, and Mahmoud Daneshmand. 2021. IoT-enabled social relation ships meet artificial social intelligence. *IEEE Internet of Things Journal*, 8(24):17817–17828. https://doi.org/10.1109/JIOT.2021.3081556

David Dodell-Feder, Sarah Hope Lincoln, Joseph P. Coulson, and Christine I. Hooker. 2013. Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PLOS ONE*, 8(11):e81279. https://doi.org/10.1371/journal.pone.0081279, PubMed: 24244736

J. R. Firth. 1957. *A Synopsis of Linguistic Theory*. Blackwell, Oxford.

Chris D. Frith and Uta Frith. 2012. Mechanisms of social cognition. *Annual Review of Psychology*, 63(1):287–313. https://doi.org/10.1146/annurev-psych-120710-100449, PubMed: 21838544

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models.

Morton Ann Gernsbacher and Melanie Yergeau. 2019. Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, 7(1):102–118. https://doi.org/10.1037/arc0000067, PubMed: 31938672

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in LLMs: Tracing data contamination in large language models.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184. https://doi.org/10.1111/tops.12007, PubMed: 23335578

Herbert P. Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill. https://doi.org/10.1163/9789004368811_003

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods.

Courtney Melinda Hale and Helen Tager-Flusberg. 2003. The influence of language on theory of mind: A training study. *Developmental Science*, 6(3):346–359. https://doi.org/10.1111/1467-7687.00289, PubMed: 16467908

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

Francesca G. E. Happé. 1994. 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. https://doi.org/10.1007/BF02172093, PubMed: 8040158

Paul L. Harris. 2005. Conversation, pretense, and theory of mind. In *Why Language Matters for Theory of Mind*, pages 70–83. Oxford University Press, New York, NY, US. https://doi.org/10.1093/acprof:oso/9780195159912.003.0004

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

Elizabeth O. Hayward and Bruce D. Homer. 2017. Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, 35(3):454–462. https://doi.org/10.1111/bjdp.12186, PubMed: 28464376

Cecilia Heyes. 2014. Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, 9(2):131–143. https://doi.org/10.1177/1745691613518076, PubMed: 26173251

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2022. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. https://doi.org/10.18653/v1/2021.emnlp-main.564

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models.

Claire Hughes, Sara R. Jaffee, Francesca Happé, Alan Taylor, Avshalom Caspi, and Terrie E. Moffitt. 2005. Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, 76(2):356–370. https://doi.org/10.1111/j.1467-8624.2005.00850_a.x, PubMed: 15784087

Steven Johnson and Nikita Iziev. 2022. AI is mastering language. Should we trust what it says. *The New York Times*.

Cameron R. Jones, Tyler A. Chang, Seana Coulson, James A. Michaelov, Sean Trott, and Benjamin Bergen. 2022. Distrubutional semantics still can't account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

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

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models.

Christopher Krupenye and Josep Call. 2019. Theory of mind in animals: Current and future directions. *WIREs Cognitive Science*, 10(6):e1503. https://doi.org/10.1002/wcs.1503, PubMed: 31099977

Christopher Krupenye, Fumihiro Kano, Satoshi Hirata, Josep Call, and Michael Tomasello. 2016. Great apes anticipate that other individuals will act according to false beliefs. *Science (New York, N.Y.)*, 354(6308):110–114. https://doi.org/10.1126/science.aaf8110, PubMed: 27846501

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J. Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.778852, PubMed: 35493614

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1598

James A. Michaelov, Seana Coulson, and Benjamin K. Bergen. 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*. https://doi.org/10.1109/TCDS.2022.3176783

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. https://doi.org/10.1073/pnas.2215907120, PubMed: 36943882

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. *arXiv:1808.09352 [cs]*. https://doi.org/10.18653/v1/D18-1261

Timothy Niven and Hung-Yu Kao 2020. Probing neural network comprehension of natural language arguments. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664. https://doi.org/10.18653/v1/P19-1459

Cathleen O'Grady, Christian Kliesch, Kenny Smith, and Thomas C. Scott-Phillips. 2015. The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, 36(4):313–322. https://doi.org/10.1016/j.evolhumbehav.2015.01.004

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Derek C. Penn and Daniel J. Povinelli. 2007. 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. https://doi.org/10.1098/rstb.2006.2023, PubMed: 17264056

Josef Perner, Susan R. Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2):125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526. https://doi.org/10.1017/S0140525X00076512

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4218–4227. PMLR.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Magnus Sahlgren and Fredrik Carlsson. 2021. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4. https://doi.org/10.3389/frai.2021.682578, PubMed: 34557662

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1454

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118. https://doi.org/10.1073/pnas.2105646118, PubMed: 34737231

Eric Schwitzgebel. 2013. A dispositional approach to attitudes: Thinking outside of the belief box. *New Essays on Belief: Constitution, Content and Structure*, pages 75–99. https://doi.org/10.1057/9781137026521_5

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424. https://doi.org/10.1017/S0140525X00005756

Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76. https://doi.org/10.1016/j.tics.2005.12.009, PubMed: 16406326

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021.

Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205. https://doi.org/10.1145/3442381.3450097

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

Dan Sperber and Deirdre Wilson. 2002. Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1–2):3–23. https://doi.org/10.1111/1468-0017.00186

Luca Surian, Stefania Caldi, and Dan Sperber. 2007. Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7):580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x, PubMed: 17614865

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691. https://doi.org/10.1017/S0140525X05000129, PubMed: 16262930

Sean Trott and Benjamin Bergen. 2018. Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, 56(8):675–707. https://doi.org/10.1080/0163853X.2018.1548219

Sean Trott and Benjamin Bergen. 2020. When do comprehenders mentalize for pragmatic inference? *Discourse Processes*, 57(10):900–920. https://doi.org/10.1080/0163853X.2020.1822709

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309. https://doi.org/10.1111/cogs.13309, PubMed: 37401923

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1566

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. https://doi.org/10.1201/9781003205388-2

Henry M. Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684. https://doi.org/10.1111/1467-8624.00304, PubMed: 11405571

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128. https://doi.org/10.1016/0010-0277(83)90004-5, PubMed: 6681741

Blaise Agüera y Arcas. 2022. Do Large Language Models Understand Us? *Daedalus*, 151(2):183–197. https://doi.org/10.1162/daed_a_01909

Zheng Zhang, Leon Bergen, Alexander Paunov, Rachel Ryskin, and Edward Gibson. 2023. Scalar implicature is sensitive to contextual alternatives. *Cognitive Science*, 47(2):e13238. https://doi.org/10.1111/cogs.13238, PubMed: 36739521

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind.

A 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 complete access to the class of objects (i.e. they have checked all of the objects to see whether they have the relevant property).

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

Table 2: Scoring criteria for Scalar Implicature E1.

A.1 Experiment 1

We check that bets on 3 decrease when $access = 3$ (scalar implicature) and do not decrease when $access < 2$ (implicature cancelled).

A.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.

B Contamination Analyses

We ran contamination analyses on the 4 pre-existing datasets to test if the items had appeared in the models' training set. We used the guided instruction method from Golchin and Surdeanu (2023), in which models generate completions for fragments of dataset items either with or without a prompt prefix describing the origin of the data. We measured the similarity of the generated and reference samples in three ways: using BLEURT scores (BLEURT-20), ROUGE-L scores, and using an In-Context Learning appraoch with GPT-4 to near-exact matches. There were no significant difference between guided and unguided scores (all p's > 0.16) and GPT-4 flagged no near-exact matches in any dataset. The results suggest that GPT-3 davinci-002's training data was not contaminated with any of the items used here to assess it.

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 E2.