Language Statistics and False Belief Reasoning: Evidence from 41 Open-Weight LMs

Anonymous ACL submission

Abstract

Research on mental state reasoning in language models (LMs) has the potential to inform theories of human social cognition—such as the theory that mental state reasoning emerges in part from *language exposure*—and our understanding of LMs themselves. Yet much published work on LMs relies on a relatively small *sample* of closed-source LMs, limiting our ability to rigorously test psychological theories and evaluate LM capacities. Here, we replicate and extend published work on the *false belief task* by assessing LM mental state reasoning behavior across 41 open-weight models (from distinct model families). We find *sensitivity* to implied knowledge states in 34% of the LMs tested; however, consistent with prior work, none fully "explain away" the effect in humans. Larger LMs show increased sensitivity and also exhibit higher psychometric predictive power. Finally, we use LM behavior to generate and test a novel hypothesis about human cognition: both humans and LMs show a bias towards attributing false beliefs when knowledge states are cued using a non-factive verb ("John thinks...") than when cued indirectly ("John looks in the..."). Unlike the primary effect of knowledge states, where human sensitivity exceeds that of LMs, the magnitude of the human knowledge cue effect falls squarely within the distribution of LM effect sizes—suggesting that distributional statistics of language can in principle account for the latter but not the former in humans. These results demonstrate the value of using larger samples of open-weight LMs to test theories of human cognition and evaluate LM capacities.

1 Introduction

Advances in the performance of language models (LMs) have raised questions about which aspects of human cognition can emerge in a learner trained purely on the distributional statistics of language. Recent debates have focused on a range of topics, including grammatical knowledge (Dentella et al., 2023; Hu et al., 2024), analogical reasoning (Lewis and Mitchell, 2025; Webb et al., 2023), and mental state reasoning (Nematzadeh et al., 2018; Sap et al., 2022; Kosinski, 2024; Trott et al., 2023; Hu et al., 2025; Shapira et al., 2024; Pang et al., 2025; Jones et al., 2024). These debates are relevant both for theories of human cognition (i.e., in terms of identifying which behaviors can be attributed to linguistic input) and for understanding the limits of current LMs (i.e., what they can and can't do).

Here, we focus on *false belief reasoning*: the human capacity to reason about the belief states of others, even when those beliefs are different from one's own—a capacity sometimes grouped under the broader umbrella of "Theory of Mind". This capacity has been widely studied in humans (Apperly, 2012), non-human animals (Call and Tomasello, 2008), and more recently, language models (Kosinski, 2023; Shapira et al., 2024). Recent work has generated debate about whether psychological tasks appropriately measure mental state reasoning in LMs (Trott et al., 2023; Ivanova, 2025), with demonstrations that LM behavior is not robust to small stimulus perturbations (Ullman, 2023; Shapira et al., 2024). One less discussed—though equally crucial—limitation of much published work on this topic concerns the *LM sample* typically used: empirical investigations often rely on a **relatively *small sample* of mostly *closed-source* LMs**—for instance, Trott et al. (2023) evaluate a single closed-source model (*text-davinci-002*), which has since been deprecated.

This limitation presents several major inferential challenges. First, as others have argued (Liesenfeld and Dingemanse, 2024; Hussain et al., 2024; Conde et al., 2025), the use of closed-source models makes it virtually impossible to detect *data contamination*, calling the validity of claims based on those evaluations into question. A related challenge is that most providers of closed-source models do not

make public even relatively coarse-grained properties such as the size of a model, the number of tokens on which it was trained, or whether (and how) it was instruction-tuned—making it difficult to draw inferences about *which properties* of a learner affect its downstream behavior. Moreover, closed-source models are vulnerable to deprecation, raising challenges for scientists attempting to *reproduce* previously published work. Finally, beyond the limitations of closed-source models specifically, the small sample size used in most published work makes it difficult to determine the *generalizability* of empirical findings obtained on a handful of model instances (Trott, 2025).

These concerns are not "merely" methodological: they represent fundamental *epistemological challenges* for a systematic study of LM capacities. Building generalizable scientific theories of LMs requires studying representative samples of model instances about which relevant properties (e.g., model size) can be verified. Further, the use of LMs as "model organisms" to test theories about human cognition depends on accurate knowledge about those LMs. For example, knowing the number of words on which an LM was trained is extremely relevant for testing theories about the role of *language exposure* in the emergence of particular capacities like false belief reasoning (Brown et al., 1996; Dove, 2020).

In the current work, we address this limitation by reproducing and extending a previously published analysis of human and LM behavior on the false belief task (Trott et al., 2023), using a larger sample of open-weight (and some open-source) LMs from five model families. We first survey related work on the *false belief task* (Section 2). We then address several research questions. First, we reproduce previously published analyses (Trott et al., 2023) with a larger sample of LMs, asking which open-weight LMs (if any) are sensitive to a character's knowledge state (Section 4.1), and whether any LMs fully "explain away" the equivalent human effect (Section 4.2). Second, we exploit the larger sample to ask which properties of an LM predict its sensitivity to knowledge states (Section 4.3), as well as its psychometric predictive power (Section 4.4). Finally, we use the sample of LMs to generate (and test) a *novel hypothesis* about human behavior on the false belief task involving the role of *non-factive verbs* (e.g., "think") in belief attribution (Section 4.5). These questions (and empirical findings) are summarized in Table 1 below.

Finally, we conclude by discussing the implications of these results for the study of mental state reasoning in LMs and humans (Section 5), as well as key limitations and directions for future work (Section 6). Note that all materials, data, and code necessary to reproduce these results will be published on GitHub when the anonymity period is over.

Research Question	**Finding**
RQ1: Are any open-weight LMs sensitive to manipulations of knowledge state?	✓
RQ2: Do any open-weight LMs "explain away" the human effect?	✗
RQ3: Which LM properties correlate with sensitivity to knowledge state?	Parameter count
RQ4: Which LM properties correlate with psychometric predictive power?	Parameter count
RQ5: Are LMs and humans more likely to attribute false beliefs in the presence of non-factive verbs like "thinks"?	✓

Table 1: Summary of research questions and findings.

2 Related Work

2.1 False Belief Reasoning in Human Populations

The capacity to reason about mental states has been extensively studied in humans, often using a version of the *false belief (FB) task* (Wimmer and Perner, 1983; Bradford et al., 2020). In the canonical FB task, a participant witnesses the transfer of an object from one location (the "start") to a new location (the "end"), with or without the knowledge of another character. The participant must then report (either directly or indirectly) where the character likely thinks the object is. While the validity of the FB task has been criticized on various grounds (Bloom and German, 2000; Gernsbacher and Yergeau, 2019; Hayward and Homer, 2017), it is widely used to assess mental state reasoning in humans (Bradford et al., 2020; Fairchild and Papafragou, 2021; Schneider et al., 2014), and has even been adapted to study mental state reasoning in non-human animals (Premack and Woodruff, 1978; Call and Tomasello, 2008; Krupenye et al., 2016; Halina, 2015) (see also Section 5). Most relevantly, there is empirical support for the theory that *language exposure* can augment mental state reasoning capacity (de Villiers and de Villiers, 2014; De Villiers, 2007; Brown et al., 1996; Gola, 2012;

Hale and Tager-Flusberg, 2003). While not all of this evidence is *causal*, it does point to a correlation between types of linguistic knowledge and the capacity to reason about belief states.

2.2 False Belief Reasoning in Language Models

A growing body of research has investigated the ability of language models (LMs) to produce responses that are *sensitive* to manipulations of a character's belief states (Sap et al., 2022), often using a version of the FB task (Kosinski, 2023; Trott et al., 2023; Shapira et al., 2024). In some cases, this has motivated the creation of Theory of Mind benchmarks (Wu et al., 2023; Xu et al., 2024; Jones et al., 2024; Gandhi et al., 2023; Gu et al., 2024). Although results vary across tasks and models, a common finding is that LMs exhibit *some* sensitivity to canonical belief-state manipulations (Kosinski, 2023), but this sensitivity falls short of human performance (Trott et al., 2023) and breaks down under minor stimulus perturbations (Ullman, 2023; Shapira et al., 2024). As discussed in Section 6, there is considerable debate about the interpretation of these findings (Hu et al., 2025), particularly with respect to the construct validity of the FB task for LMs (Trott et al., 2023; Ivanova, 2025).

An equally important limitation, however, is that it often relies on a small sample of closed-source LMs to draw claims about mental state reasoning capacity in LMs more generally. This is particularly problematic when the goal of research is identifying the properties of individual learners that predict FB task performance—or evaluating the *sufficiency* of distributional language statistics in accounting for human behavior on the FB task.

2.3 Motivation

To address this limitation, we selected Trott et al. (2023) as a target for replication and extension for several reasons: first, the original study collected novel human data as a comparison for LM behavior, and explicitly investigated the link between distributional statistics and human FB task performance; second, the study used entirely novel stimuli, reducing the possibility of data contamination; third, the study used only a single model instance (*text-davinci-002*), which has since been deprecated; and fourth, the study analyzed (and made public) individual human data, allowing us to predict trial-level results rather than solely comparing to an average.

Notably, the current work is also (to our knowledge) a novel approach to testing the distributional hypothesis as an explanation of human behavior on a psycholinguistic task: by evaluating a *range* of LMs, we can directly quantify the likelihood of obtaining a given human effect size relative to the distribution of LM effect sizes (see Figure 4b).

3 Method

3.1 Materials

We used the same written passages developed by Trott et al. (2023), which followed the general structure of the canonical False Belief (FB) Task (see Figure 5 for an example). Specifically, in each scenario, a character places an object in the Start location (e.g., a box), and a second character moves that object to the End location (e.g., a basket). The key manipulation of Knowledge State corresponds to whether the first character is present (True Belief) or not present (False Belief) when the object is moved. Each passage ends with a probe statement assessing the character's belief about the object's location, either Explicitly (e.g., "John believes the book is in the...") or Implicitly (e.g., "John looks for the book in the ..."). The authors also manipulated which location (e.g., the box vs. the basket) first in the passage, and which was mentioned most recently. Each location was always mentioned twice in the passage. There were 12 different scenarios, with 16 versions each (Knowledge State x Knowledge Cue x First Mention x Recent Mention), for a total of 192 passages.

3.2 Models

We aimed to sample a diversity of model sizes, training regimes, and providers. We selected a total of 41 models from 5 model *families* (Gemma Team et al., 2024; Grattafiori et al., 2024; OLMo Team et al., 2025; Biderman et al., 2023; Qwen Team et al., 2025), where "family" refers to a collection of model instances released by the same model provider; these model instances typically share a number of features (such as pre-training methodology), but also vary in key respects (such as number of parameters, or whether a model is instruction-tuned).[1]

[1] See Appendix A for an analysis of additional models and model families accessed through the HuggingFace Inference Endpoints API.

Family	Huggingface ID	Versions
Gemma 2	google/gemma-2-2b*	Base, Instruct
Llama 3	meta-llama/Meta-Llama-3-8B*	Base, Instruct
OLMo-2	allenai/OLMo-2-*-1124	Base, Instruct
Pythia	EleutherAI/pythia-*	Base only
Qwen 2.5	Qwen/Qwen2.5-*	Base, Instruct

Table 2: Model families tested in our experiments. See Appendix A for an additional results with models run through the HuggingFace Inference Endpoints API.

3.3 Procedure

Model instances were run locally on an NVIDIA DGX-H200 using the HuggingFace *transformers* package (Wolf et al., 2020); see Appendix A for a supplementary analysis using larger models accessed through the HuggingFace Inference Endpoints API.

Each model instance was assessed in the following way. A stimulus was first *tokenized* using that model's tokenizer. Following Trott et al. (2023), we then presented the tokens to the model and extracted the log probability corresponding to the Start location token (e.g., "box") and the End location token (e.g., "basket").[2] We then calculated the Log Odds by taking the difference between these log probabilities, such that a positive Log Odds indicated a higher probability was assigned to the Start location token. This was repeated for all 192 passages, for each model instance.

For instruction-tuned models, we carried out an additional analysis in which the models were prompted to provide an answer to a question (as opposed to simply completing a sentence), which was intended to mirror the instruction-tuning process (see Appendix B).

3.4 Human Data

We used publicly available data from Trott et al. (2023). After applying exclusion criteria, a total of 613 participants completed the original single-trial experiment, which involved reading the target passage and completing the final sentence, e.g., "John thinks the book is in the ...'. These were coded in terms of whether the participant responded with the Start or End location.

4 Results

All analyses were conducted in R (R Core Team, 2025). Mixed effects models were constructed using the *lme4* package (Bates et al., 2015), and visualizations were produced using *ggplot2* (Wickham, 2016).

4.1 RQ 1: Are any open-weight LMs sensitive to Knowledge State?

We first asked whether any of the open-weight models we tested were *sensitive* to implied belief states. Specifically, as in past work (Trott et al., 2023), we asked whether the distribution of Log Odds produced by a given model instance varied systematically according to the manipulation of Knowledge State. For each model instance, we constructed a linear mixed effects model with Log Odds as a dependent variable; fixed effects of Knowledge State, Knowledge Cue, First Mention, and Recent Mention; and by-item random slopes for the effect of Knowledge State. We then compared this model to a model omitting only the effect of Knowledge State using a likelihood ratio test (LRT).

After applying Holm-Bonferroni corrections, 14 of the 41 models tested (approximately 34%) exhibited sensitivity to Knowledge State (see Figure 1 for a subset of these models). In each case, the effect was in the right direction: a more negative Log Odds was observed for the True Belief condition, corresponding to a (relatively) higher probability assigned to the End (correct) location. Likelihood-ratios for this subset of models ranged from 11.9 to 21.4. We also conducted a follow-up analysis in which all LMs were analyzed together in a single mixed effects model, with by-LM random slopes for the effect of Knowledge State. Again, this full model explained significantly more variance than a model omitting only the effect of Knowledge State $[\chi^2(1) = 14.78, p < .001]$, and more negative Log Odds were again observed in the True Belief condition $[\beta = -1.5, SE = 0.37, p < .001]$.

Viewed from the lens of *accuracy*—positive Log Odds for False Belief passages, and negative Log Odds for True Belief passages—the top-performing LMs achieved 74.5% accuracy, roughly analogous to the performance of GPT-3 in past work. The best performance was achieved by instruction-tuned variants of the OLMo 2 13B model, as well as the base model. Average accuracy across the entire sample of LMs was 56.4%.

4.2 RQ 2: Do any open-weight LMs "explain away" the effect on humans?

Past work found that LMs did not fully "explain away" the effect of Knowledge State on human responses. We replicated this "distributional base-

[2] Multi-token words were addressed by summing the log probabilities across all constituent tokens.

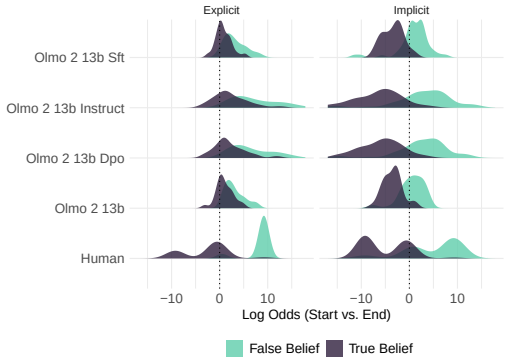


Figure 1: Distribution of Log Odds by condition (True Belief vs. False Belief) for the best-performing models, along with the by-item human baseline. Human Log Odds were calculated by comparing the proportion of Start vs. End responses for each item across participants.

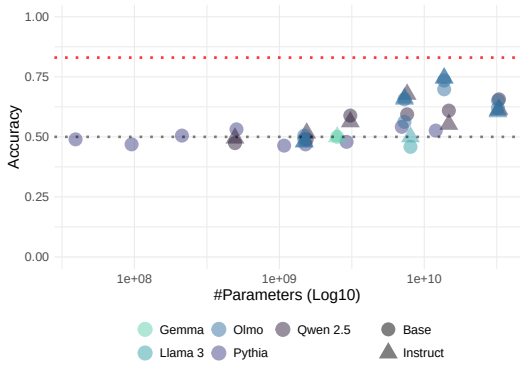


Figure 2: Accuracy on the false belief task by number of parameters. Overall, larger models performed better, though no model attained or surpassed the human baseline. Dashed red line corresponds to average human performance.

lines" analysis for each LM by constructing a generalized linear mixed effects model with Human Response as a dependent variable (Start vs. End); fixed effects of Knowledge State, Log Odds, Knowledge Cue, Recent Mention, and First Mention; and random intercepts for each item. We compared this full model to a model omitting only the effect of Knowledge State to determine whether the manipulation explained variance in behavior beyond Log Odds. For all LMs, the full model explained significantly more variance (likelihood ratios ranged from 123.6 to 195.4). That is, Log Odds were *insufficient* to fully account for the effect of Knowledge State on human behavior.

More straightforwardly, human accuracy (approximately 82.7%) surpassed even the top-performing LMs (74.5%) (see Figure 2). We also calculated an approximate effect size by comparing the proportion of Start responses across False Belief and True Belief conditions—a difference that should be positive, since characters are more likely to look in the Start (incorrect) location in the False Belief condition. As depicted in Figure 4b, this effect was considerably larger for humans than for any LM tested. Thus, the distributional statistics of language—as operationalized by this sample of LMs—account for *some sensitivity* to manipulations of Knowledge State, but do not currently account for human-level sensitivity.

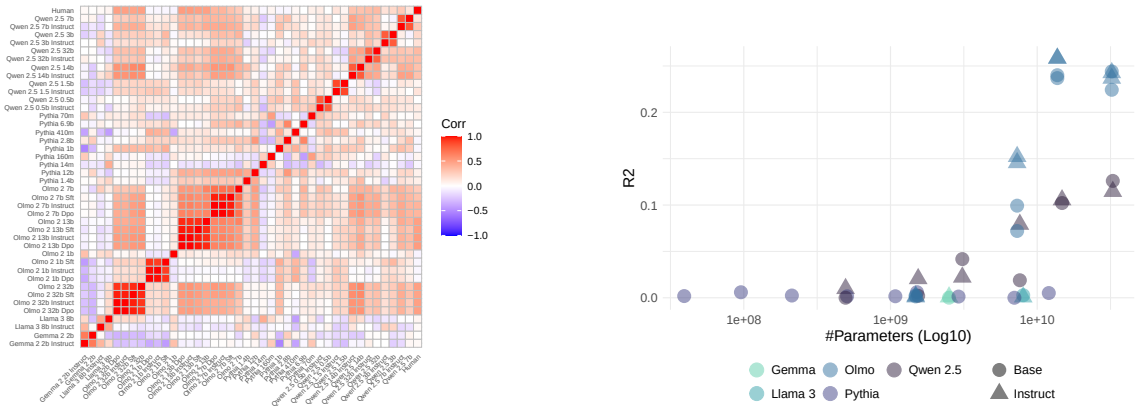
4.3 RQ 3: Which factors predict LM performance?

A key advantage of testing a large sample of LMs, as opposed to a single instance, is that it allows researchers to investigate which properties of individual model instances predict variance in performance. We focused on three factors found to be relevant to performance on various tasks in past work (Kaplan et al., 2020; Ouyang et al., 2022): the number of parameters in a model, the number of tokens a model was trained on, and whether a model was subject to instruction-tuning.

These factors were entered as predictors in a linear regression model with mean accuracy as a dependent variable (parameter count and token count were both log-transformed). Altogether, this model achieved an $R^2 = 0.44$, though only Log Number of Parameters had a significant (positive) coefficient $[\beta = 0.08, SE = 0.02, p < .001]$: that is, each order of magnitude increase in parameter count was associated with an expected increase of about 8% in accuracy (see also Figure 2). Of course, this relationship was not perfectly monotonic: the best-performing model (OLMo 2 13B) was about half the size of the largest models tested (e.g., OLMo 2 32B). Note that we also carried out additional analyses of instruction-tuned variants to investigate the role of specific prompt formatting; these results are described in Appendix B.

4.4 RQ 4: Which factors contribute to psychometric predictive power?

Another metric of performance is *psychometric predictive power* (PPP), i.e., the extent to which a given LM's behavior *predicts* human behavior (Wilcox et al., 2023). Although no LM "explained away" the effect of Knowledge State on human responses (see RQ 2), LMs did vary considerably in their performance, leading us to ask whether some



(a) Correlation matrix between model outputs and human responses. Human responses were aggregated to create a Log Odds measure for each item.

(b) Psychometric predictive power (PPP) for each LM instance, as determined by the proportion of variance explained (R^2) in by-item human Log Odds by that LM's Log Odds. Larger models achieved higher PPP.

Figure 3: Although no LM fully "explained away" the effect of Knowledge State in humans, some LMs did produce behavior that was more correlated with human behavior overall.

LMs were more correlated with human behavior than others.

We calculated PPP in two ways. First, we asked how much variance in human behavior could be accounted for as a function of LM Log Odds—setting aside other potential explanatory variables such as Knowledge State. For a direct comparison, we aggregated across human responses for each item to compute a Log Odds measure (i.e., the log ratio of the proportion of participants choosing the Start location vs. the proportion of participants choosing the End location). We then calculated the Pearson's correlation between each LM's Log Odds and the human Log Odds, as well as each pair of LMs (see Figure 3a). PPP for each LM was approximated by calculating R^2, i.e., the proportion of variance in by-item Log Odds explained by that LM's Log Odds. The best-performing LM (an instruction-tuned variant of OLMo 2 13B) achieved an $R^2 = 0.26$. To ask which factors contributed to variance in PPP, we fit a multiple regression model predicting PPP from Log Parameter count, Log Training Tokens, and whether an LM was instruction-tuned: only Log Parameter Count was a significant predictor $[\beta = 0.09, SE = 0.02, p < .001]$ (see also Figure 3b).

In a second approach (PPP_2), we asked which LMs produced behavior that *explained additional variance* above and beyond Knowledge State. We fit a generalized linear mixed effects model predicting human response (Start vs. End location) with fixed effects of LM Log Odds, First Mention, Recent Mention, Knowledge State, and Knowledge Cue, as well as random intercepts for each item. We compared the Akaike Information Criterion (AIC) of this model to the AIC of a model omitting only LM Log Odds. Using this approach, a lower (more negative) ΔAIC corresponded to higher PPP relative to the other predictors, including Knowledge State. ΔAIC values ranged from 2 (indicating no improvement in model fit, along with a penalty for the additional Log Odds parameter) to approximately -8. Lower ΔAIC was generally achieved by the models with the highest accuracy, though the two were not perfectly correlated (e.g., OLMo 2 7B obtained the lowest ΔAIC). PPP was significantly predicted by Log Parameter Count in the expected direction, i.e., larger models achieved a relatively larger reduction in AIC $[\beta = -1.57, SE = 0.68, p = 0.03]$.

Together with RQ2, these results suggest that larger models were *better at the FB Task* (RQ2) and better at accounting for human behavior on the FB task, as well as human behavior *independent* of the manipulation of Knowledge State.

4.5 RQ 5: Can LM behavior be used to generate novel hypotheses?

One proposed benefit of using LMs in psycholinguistic research is the possibility of generating and testing *novel* hypotheses (Misra and Kim, 2024; Lakretz et al., 2021). Past work (Trott et al., 2023) observed an effect of Knowledge Cue on GPT-3 Log Odds: namely, Log Odds were systematically higher in the Explicit condition ("John thinks the

book is in the...") than the Implicit condition ("John looks for the book in the..."). In other words, GPT-3 displayed a bias towards the *incorrect* (Start) location when a *non-factive* verb ("thinks") was used.

Although Trott et al. (2023) did not expand on this finding, there is a large body of work exploring the different presuppositions and entailments derived from factive and non-factive verbs (Hooper, 1975; Kiparsky and Kiparsky, 1970). Indeed, some (Chemla, 2008) have suggested that non-factive verbs (e.g., "think" or "believe") carry an *anti-presupposition*: that is, "John *thinks* A" may weakly imply *NOT A*. This would be broadly consistent with the finding described above (i.e., imputing an incorrect belief to an agent when a non-factive verb is used), even though the original study was not designed with this hypothesis in mind. Moreover, there is also some evidence that LMs are sensitive to epistemic cues, such as expressions of uncertainty (Zhou et al., 2023).

We extended the original finding by asking whether any LMs in the current sample also exhibited sensitivity to Knowledge Cue, whether humans showed analogous sensitivity, and whether the human effect was larger or smaller than the LM effect. For more direct comparison of the coefficients, we converted each LM's Log Odds into a categorical measure reflecting whether the ratio was positive (Start-biased) or negative (End-biased). Then, for each LM, we constructed a generalized linear mixed effects model predicting Response (Start vs. End), with fixed effects of Knowledge State, Knowledge Cue, First Mention, and Recent Mention (and by-item slopes of Knowledge State). We extracted the coefficient for Knowledge Cue for each LM: the mean across all LMs tested was negative $\bar{\beta}_{\text{KC}} = -1.6$ (median $= -1.08$, $SE = 0.43$[3]), and values ranged from -11.67 to 1.5. A multiple regression model with β_{KC} as a dependent variable, and predictor terms for Log Parameter Count, Log Token Count, and Base/Instruct status, indicated that larger models showed more negative (stronger) effects on average $[\beta = -2.04, SE = 0.63, p = 0.003]$.

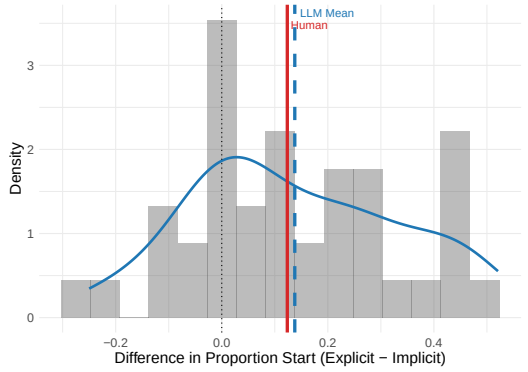
We then fit an analogous model for the human data, and the resulting coefficient for Knowledge Cue was also negative $[\beta = -1.02, SE = 0.27, p < .001]$: that is, humans were more likely to respond with the Start (incorrect) location in the Explicit condition, i.e., when the verb "thinks" was used (and more likely to respond with the End location in the Implicit condition). Together, these results suggest that both LMs and humans were biased towards reporting *incorrect* belief states when non-factive verbs like "think" were used, perhaps because of an implied *anti-presupposition* (Chemla, 2008). The human parameter estimate fell in approximately the middle percentile (49%) of LM parameter estimates.

As with the primary Knowledge State effect, we visualized these effect sizes by comparing the probability of a Start response across Explicit vs. Implicit conditions (see Figure 4a). Here, a positive score reflects a bias towards the Start location in the Explicit condition (i.e., when a non-factive verb is used). In contrast to the effect of Knowledge State (where the human effect was much bigger than the distribution of LM effects), the effect of Knowledge Cue in humans fell squarely in the middle of the distribution of LM effect sizes. That is, distributional statistics are in principle *sufficient* to account for the average effect size of Knowledge Cue on human responses, but not for the average effect size of Knowledge State.

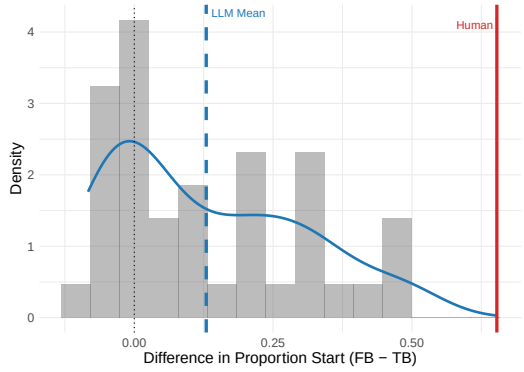
5 General Discussion

We tested a large sample of open-weight LMs (including a number of fully open-source LMs) on the *false belief task* and used their behavior to predict individual human behavior on the same task. Our results point to several theoretical conclusions. First, LMs trained on the distributional statistics of language can develop *sensitivity* to implied belief states, consistent with (though not entailing) the hypothesis that language exposure can partially account for mental state reasoning capacity in humans (Figure 1). Moreover, larger LMs are more sensitive to mental states (Figure 2) and better predict human behavior on the same task (Figure 3b). However, no LM fully "explained away" the human effect, suggesting that mental state reasoning in humans might well depend on other factors or processes beyond language statistics (Figure 4b). This is particularly striking given that the LMs tested are trained on orders of magnitude more language data than most humans encounter in a lifetime (Warstadt et al., 2023). That is, humans are more *sample-efficient* in terms of the ratio of linguistic input to FB task performance (though not necessarily in terms of other input sources). The question of what

[3] Standard error values were calculated by taking the standard deviation of the coefficient estimates across the entire sample and dividing by the square root of the sample size.



(a) Distribution of differences in proportion of Start location responses across Explicit and Implicit conditions.



(b) Distribution of differences in proportion of Start location responses across False Belief and True Belief conditions.

Figure 4: Both LMs and humans exhibited a bias towards the Start (incorrect) location when a non-factive verb ("thinks") was used in the probe sentence. Both populations were also more likely to respond with the Start location in the False Belief condition, but this effect was considerably larger in humans.

accounts for this gap—grounding, social interaction, or even innate biological constraints—remains unknown, but could potentially be explored by systematically operationalizing each proposed mechanism in the context of LM training regimes or architectures.

The insufficiency of distributional statistics in accounting for manipulations of knowledge state also stands in sharp contrast to a novel finding reported in the current work: namely, LMs were (approximately) equally sensitive as humans to the manipulation of *knowledge cue* (Figure 4a). One potential explanation for this asymmetry is that the "anti-presupposition" implicit in non-factive verbs like "thinks" is plausibly a lexical or conventionalized phenomenon: if "X thinks P" appears in many cases where P is uncertain or even false, then the association between "thinks" and false beliefs could be learned through the distributional statistics of language—provided the LM has also learned to encode the likelihood of "P" being true. The FB task, on the other hand, may require constructing a *situation model* (Zwaan et al., 1995) and representing the likely belief states of various entities in the discourse. This computation might be more challenging to learn through language statistics.

These results build directly on published work relying on (now-deprecated) closed-source LMs (Trott et al., 2023). The findings are also broadly consistent with a growing body of research suggesting that although LMs are surprisingly capable on mental state reasoning tasks (Kosinski, 2023; Jones et al., 2024), their performance remains relatively brittle in the face of small perturbations (Shapira et al., 2024) and does not match human behavior (Trott et al., 2023). As argued in Section 1, a key contribution of the current work is the use of a larger sample of open-weight (and open-data) LMs, which has several benefits. First, it makes the work more reproducible. Second, it allows us to investigate the *generalizability* of claims about LM capabilities, as well as the factors contributing to variation across model instances, which is crucial for building a systematic science of LM behavior and capabilities; such findings can also help "taxonomize" LMs in terms of their behavioral similarity (see also Appendix C). And third, it provides a *distribution* of "baseline" effect sizes attributable to language statistics, as opposed to relying on a point estimate from a single LM—which in turn facilitates more robust statistical comparison to human effect sizes. For instance, we observed that a large subset of LMs showed human-level effects for the manipulation of Knowledge Cue but not Knowledge State (Figure 4a and Figure 4b). This latter approach may prove useful for future work evaluating the cognitive capacities of LMs or using LMs as "model organisms" to test (or generate) hypotheses about human cognition.

6 Limitations

One limitation of the current work concerns the false belief task itself. As others have argued (Bloom and German, 2000; Gernsbacher and Yergeau, 2019), the task likely involves capacities other than mental state reasoning and may not accurately reflect "Theory of Mind" even in human samples. Inferences about mental state rea-

soning capacities are further complicated when applying the test to LMs, which may apply heuristics or "short-cuts" to solve the task based on patterns observed in the training data (Shapira et al., 2024)—raising the question about whether the task (or any measure originally developed for humans subjects) has *differential construct validity* for humans and LMs (Trott et al., 2023; Ivanova, 2025; Saxon et al., 2024). Our aim here was not to resolve these important theoretical questions, but rather to reproduce and extend existing work using open-weight models and to ask which behaviors can emerge *in principle* from exposure to language statistics. We refer interested readers to more detailed discussions of construct validity elsewhere (Ivanova, 2025; Trott et al., 2023; Shapira et al., 2024; Hu et al., 2025; Bean et al., 2025), and note that addressing these challenges will likely require open-weight (and ideally open-data) models, for which relevant properties—such as training data composition—can be more accurately characterized. That said, future work would likely benefit from greater *coverage* of possible scenarios (i.e., stimuli) requiring mental state reasoning to minimize the likelihood that good performance by LMs (or humans) is driven by "shortcuts" overfit to a particular stimulus structure. Future work could also adopt a mechanistic approach (Wu et al., 2025) to ask whether consistent "circuits" are consistently recruited for a variety of scenarios involving mental state reasoning.

A second limitation is *external validity*: we assessed a (relatively) large sample of open-weight LMs from multiple model families, but it is unclear whether this sample is *representative* of distributional learners more generally, or what such a representative sample might look like (Constantinescu et al., 2025). The question of *generalizability* (i.e., across architectures, training regimes, random seeds, and scale) has emerged as a key concern in recent work studying LM behavior and internal mechanisms (Olah et al., 2020; Olah, 2023; Tigges et al., 2024; van der Wal et al., 2025; Trott, 2025; Michaelov et al., 2025; Zhao et al., 2025). While there has been some work attempting to "taxonomize" LMs (Yax et al., 2024), it remains unclear which empirical findings we should expect to generalize from a given LM instance to another. We aimed to sample a large range of model *sizes* and also considered both "base" and "instruct" versions of different models, but the space of *possible* models is vast and largely unexplored; therefore, we limit our claims to the observed sample of LMs. One exciting direction would be to ask which properties allow researchers to predict whether two LMs will exhibit convergent or divergent behavior on the same task, perhaps allowing for a task-driven taxonomy of model instances (see also Appendix C).

Notably, the current sample was limited in that it excluded closed-source, state-of-the-art LMs. It is plausible that such LMs would attain better performance on the FB task—at the same time, the lack of transparency around their training process or composition (e.g., system prompts, use of external "tools", etc.) makes them less appropriate candidates for operationalizing theories about the sufficiency of distributional statistics specifically. Future work could benefit from exploring which kinds of modifications to the base LM architecture or training objective augment performance on the FB task; such research would also inform theories of the mechanisms and processes undergirding human mental state reasoning.

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.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, and 1 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR. ISSN: 2640-3498.

Paul Bloom and Tim P German. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):B25–B31.

Elisabeth EF Bradford, Victoria EA Brunsdon, and Heather J Ferguson. 2020. The neural basis of belief-attribution across the lifespan: False-belief reasoning and the n400 effect. *cortex*, 126:265–280.

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.

Josep Call and Michael Tomasello. 2008. Does the chimpanzee have a theory of mind? 30 years later. *Trends in cognitive sciences*, 12(5):187–192.

Emmanuel Chemla. 2008. An epistemic step for anti-presuppositions. *Journal of Semantics*, 25(2):141–173.

Javier Conde, María Grandury, Tairan Fu, Carlos Arriaga, Gonzalo Martínez, Thomas Clark, Sean Trott, Clarence Gerald Green, Pedro Reviriego, and Marc Brysbaert. 2025. Adding llms to the psycholinguistic norming toolbox: A practical guide to getting the most out of human ratings. *arXiv preprint arXiv:2509.14405*.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Jill De Villiers. 2007. The interface of language and theory of mind. *Lingua*, 117(11):1858–1878.

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.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Guy Dove. 2020. More than a scaffold: Language is a neuroenhancement. *Cognitive neuropsychology*, 37(5-6):288–311.

Sarah Fairchild and Anna Papafragou. 2021. The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, 45(2):e12938.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 178 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*. ArXiv:2408.00118 [cs].

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.

Alice Ann Howard Gola. 2012. Mental verb input for promoting children's theory of mind: A training study. *Cognitive Development*, 27(1):64–76.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs] version: 3.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

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.

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

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.

Joan B Hooper. 1975. On assertive predicates. In *Syntax and Semantics volume 4*, pages 91–124. Brill.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499.

Ann Huang, Satpreet H Singh, Flavio Martinelli, and Kanaka Rajan. 2025. Measuring and controlling solution degeneracy across task-trained recurrent neural networks. *ArXiv*, pages arXiv–2410.

Zak Hussain, Marcel Binz, Rui Mata, and Dirk U Wulff. 2024. A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, 56(8):8214–8237.

Anna A Ivanova. 2025. How to evaluate the cognitive abilities of llms. *Nature Human Behaviour*, 9(2):230–233.

Cameron R Jones, Sean Trott, and Benjamin Bergen. 2024. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (epitome). *Transactions of the Association for Computational Linguistics*, 12:803–819.

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

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In Manfred Bierwisch and Karl Erich Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, The Hague.

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

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

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*, 354(6308):110–114.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699.

Martha Lewis and Melanie Mitchell. 2025. Evaluating the robustness of analogical reasoning in large language models. *Transactions on Machine Learning Research*.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1774–1787.

James A Michaelov, Roger P Levy, and Benjamin K Bergen. 2025. Language model behavioral phases are consistent across architecture, training data, and scale. *arXiv preprint arXiv:2510.24963*.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Chris Olah. 2023. Interpretability dreams. *Transformer Circuits*. Https://transformer-circuits.pub/2023/interpretability-dreams/index.html.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 OLMo 2 Furious. *arXiv preprint*. ArXiv:2501.00656 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Damian KF Pang, Severin KY Pang, Marianne D Broeker, and Alix Hibble. 2025. Do large language models have a theory of mind? *Proceedings of the National Academy of Sciences*, 122(28):e2507080122.

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

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. In *First Conference on Language Modeling*.

Dana Schneider, Virginia P Slaughter, Stefanie I Becker, and Paul E Dux. 2014. Implicit false-belief processing in the human brain. *NeuroImage*, 101:268–275.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. Llm circuit analyses are consistent across training and scale. *Advances in Neural Information Processing Systems*, 37:40699–40731.

Sean Trott. 2025. Toward a theory of generalizability in LLM mechanistic interpretability research. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

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.

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

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. 2025. Polypythias: Stability and outliers across fifty language model pretraining runs. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, pages 1–25.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Yuheng Wu, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, and Denghui Zhang. 2025. How large language models encode theory-of-mind: a study on sparse parameter patterns. *npj Artificial Intelligence*, 1(1):20.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. 2024. Phylolm: Inferring the phylogeny of large language models and predicting their performances in benchmarks. *arXiv preprint arXiv:2404.04671*.

Rosie Zhao, Tian Qin, David Alvarez-Melis, Sham Kakade, and Naomi Saphra. 2025. Distributional scaling laws for emergent capabilities. *arXiv preprint arXiv:2502.17356*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297.

A Models run through Endpoints API

Select larger models were accessed using the HuggingFace Inference Endpoints API, which allows users to access model outputs through a server hosted by HuggingFace with the same *transformers* (Wolf et al., 2020) library used for the models

that were run locally. Specifically, we ran Llama 3 70B (Base and Instruct), Qwen 2.5 72B (Instruct), Llama 3.1 70B (Base and Instruct), and Mixtral 8x7B (Base and Instruct). The same procedure was applied as for the models run locally (i.e., tokenizing the passage and Log Odds of completing the passage with the Start vs. End token). We then replicated the primary analyses above.

First, all models tested were *sensitive* to manipulations of knowledge state, even after adjusting for multiple comparisons (likelihood ratios > 6 in all cases); accuracy ranged from 57.8% (for Mixtral 8x7B Instruct) to 73.4% (for Llama 3 70B Instruct). Second, no model fully "explained away" the effect of Knowledge State in humans: that is, a mixed model containing a fixed effect of Knowledge State consistently explained variance in human responses above and beyond Log Odds from each of the LMs tested (likelihood ratios ranged from 117.6 to 174.5). Finally, all models showed a significant negative effect of Knowledge Cue, i.e., passage completions were biased towards the End (correct) location in the Implicit condition—consistent with the effects reported in the primary manuscript.

B Assessing instruction-tuned models with question prompts

Past work on instruction-tuned models has generally found a performance benefit of instruction-tuning (Ouyang et al., 2022), including on the False Belief task (Trott et al., 2023). One potential limitation of the approach adopted in the primary manuscript is that the same prompt strategy was used for "base" models and for instruction-tuned variants, even though instruction-tuning typically involves training models to follow explicit prompts and generate responses to queries, rather than computing next-token probabilities for sentence completions. Although this was an intentional feature of our experimental design (to control for the exact prompt used), this could unfairly bias the results *against* instruction-tuned models. To address this limitation, we conducted a follow-up study on the subset of instruction-tuned models using a prompt format that mirrored the approach taken in their training process.

B.1 Procedure

We selected the subset of 21 instruction-tuned models used in the original analysis. Each passage was presented to each LM, embedded in an appropriate *prompt template* based on the instruction-tuning procedure unique to each model, and rephrased into a question-answer format minimally different from its original phrasing. As in the original analysis, we then calculated the Log Odds of the Start vs. End token following the *answer* portion of the prompt template. These different formats are shown in Figure 5.

```
David and Marta go out to get some wine for
the party. When they get home, David stores
the wine in the garage and grabs a drink from
the fridge. Then, David goes out to get some
snacks. While David is gone, Marta decides the
wine would be best cooled, so she moves the
wine out of the garage and into the fridge.
David returns home and wants to put out the
wine. David thinks the wine is in the [MASK]

<start_of_turn>user\nDavid and Marta go out
to get some wine for the party. When they get
home, David stores the wine in the garage and
grabs a drink from the fridge. Then, David
goes out to get some snacks. While David is
gone, Marta decides the wine would be best
cooled, so she moves the wine out of the
garage and into the fridge. David returns
home and wants to put out the wine. David
thinks the wine is in the:
<end_of_turn>\n<start_of_turn>model\n[MASK]

<start_of_turn>user\nDavid and Marta go out
to get some wine for the party. When they get
home, David stores the wine in the garage and
grabs a drink from the fridge. Then, David
goes out to get some snacks. While David is
gone, Marta decides the wine would be best
cooled, so she moves the wine out of the
garage and into the fridge. David returns
home and wants to put out the wine.
Where does David think the wine is?
<end_of_turn>\n<start_of_turn>model\nDavid
thinks the wine is in the [MASK]
```

Figure 5: An example stimulus, formatted for continuation mode (top), instruction mode (middle), and instruction mode in a question-answer format (bottom). The text in red remains the same, but changes position slightly in the question-answer format. The blue [MASK] is where the probe of the Start and End token occurs. **Text in bold** are tokens introduced through the chat template for instruction-tuned models. In this example, the Gemma chat template is shown. These tokens will look different depending on the specific chat template unique to each model. The underlined text is the question probe used only in the question-answer format.

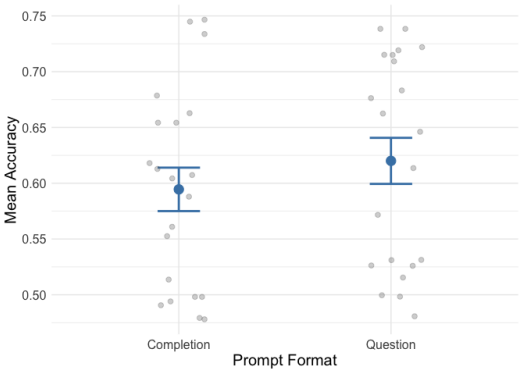


Figure 6: Mean accuracy was slightly higher when instruction-tuned models were using a question/answer (Question) format than when asked to complete the passage (Completion).

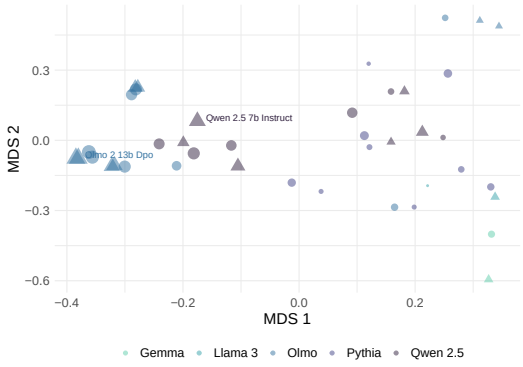


Figure 7: Results of multi-dimensional scaling (MDS) applied to the correlation matrix presented in the primary manuscript. Size of each dot indicates mean accuracy of that model instance; shapes represent Base/Instruct (circle = Base, triangle = Instruct).

B.2 Results

On average, the instruction-tuned LMs achieved slightly higher performance when prompted using the Question strategy (62%) vs. the default Completion strategy (59.4%). A linear mixed effects model predicting the mean accuracy of each LM with Prompt Strategy (Question vs. Completion) and a random intercept for each LM revealed a significant effect of Prompt Strategy, with higher accuracy predicted in the Question condition $[\beta = 0.03, SE = 0.01, p = .03]$. Notably, this effect size was relatively small (an expected difference of about 3%); moreover, the top-performing LMs did not *improve* when prompted using the Question strategy. Indeed, the maximum accuracy achieved was actually in the Completion condition (74.4%, as reported in the primary manuscript). Together, these results suggest that the theoretical conclusions reached in the primary manuscript (e.g., that no LMs attain human-level performance on the FB Task) are not particularly dependent on the prompting strategy used for these instruction-tuned models.

C MDS analysis of models

In the primary manuscript, we presented a correlation matrix illustration the Pearson's correlation between the behavior of each pair of models. Following other recent work in the study of larger LM samples (Huang et al., 2025; Trott, 2025), we performed multi-dimensional scaling (MDS) on this correlation matrix to visualize how each individual clustered relative to other model instances. As depicted in Figure 7, model instances clustered strongly by *model family*. Additionally, the first MDS component appeared to track *accuracy*, with more accurate model instances receiving more negative scores. Together, this suggests that cross-model correlations can be explained both by overall performance (more accurate models are more similar to each other) and by aspects of the architecture and training procedure (instances from the same family cluster together).