Do Different Theory of Mind Tasks for LLMs Measure the Same Thing?

Sean Trott
Rutgers University - Newark
sean.trott@rutgers.edu

Cameron Jones
Stony Brook University
cameron.jones@stonybrook.edu

Pamela Rivière
Rutgers University - Newark
pamela.riviereruiz@rutgers.edu

Abstract

The capacity to infer and reason about the mental states of others—sometimes called Theory of Mind (ToM)—has become a topic of considerable interest in research on large language models (LLMs). Yet measures of ToM in LLMs face questions of *construct validity*: are these tasks measuring what they purport to measure? In this work, we approach this question through the lens of convergent validity, asking whether different ToM tasks produce correlated observations across a sample of 36 LLMs from three model families. We find evidence for moderate convergence across some tasks (e.g., variants of the false belief task), while other tasks exhibit weak or even negative correlations. Moreover, we find that pairwise correlations between ToM tasks are consistently weaker than correlations between those tasks and other measures, such as model size or performance on a control task. We discuss several interpretations of these findings, which vary in their stance on whether ToM is a unified construct in LLMs (and we simply need to identify which tasks are best at measuring it) or whether ToM is a collection of only partially related abilities (in which case we should not expect convergence). We argue that the field needs systematic investigation of ToM measures, drawing on insights from psychometrics and the history and philosophy of measurement before strong inferences about LLM ToM can be drawn from any individual task.

1 Introduction

The ability to infer and reason about the mental states of others plays a key role in human social cognition, cooperation, and pragmatic language understanding (Tomasello et al., 2005; Apperly, 2012). This capacity—sometimes called *mentalizing* or *Theory of Mind* (ToM)—is also of increasing interest to large language model (LLM) researchers for both theoretical and practical reasons, which has motivated the creation and deployment of various tasks or "benchmarks" to measure the putative ToM capacity of LLMs (Strachan et al., 2024; Trott et al., 2023; Kosinski, 2024; Jones et al., 2024; Gandhi et al., 2023; Xu et al., 2024; Trott et al., 2026; Kouwenhoven et al., 2026). Yet this work has also led to considerable debate about how to interpret these results (Ullman, 2023; Shapira et al., 2024; Hu et al., 2025), in large part centering around the question of *construct validity*. Are measures of ToM in LLMs valid—and how would we know? In this paper, we focus on one necessary (though not sufficient) condition for construct validity: *convergent validity*, i.e., whether different tasks designed to measure ToM produce correlated observations.

Construct validity is a topic that has long occupied both philosophers and practitioners of science (Cronbach and Meehl, 1955; Chang, 2004), and there continues to be debate about the most appropriate procedures for validating a measure (Borsboom et al., 2004; Alexandrova and Haybron, 2016; Larroulet Philippi, 2021).[1] Some researchers emphasize the need for a causal relation between the construct of interest and the measure of that construct (Borsboom et al., 2004); others, while endorsing the causal criterion, have argued that the validity of a measure also depends on the research context, e.g., the sample or explanatory goal (Larroulet Philippi, 2021). Typically, researchers tend to rely on convergence between measures, i.e., under the assumption that different measures of "the same thing" should produce correlated observations (Chang, 2004).

Yet as Chang (2004) argues, a central challenge in validating a novel measure of a construct is a kind of epistemic circularity: without some external "ground truth" of the construct (e.g., temperature), how can the reliability of an instrument (e.g.,

[1] Indeed, measures of ToM in humans have repeatedly faced questions about their validity, i.e., the extent to which they reliably assess variation in mental state reasoning capacity (Bloom and German, 2000; Dörrenberg et al., 2018).

a mercury thermometer) be determined? Chang (2004) suggests that historically, this challenge was overcome in part through rigorous empirical investigation and an assumption—as per above—that good measures should be somehow comparable. While validating a measure depends on more than empirical convergence alone (e.g., it also depends on the articulation of theoretical commitments and goals (Alexandrova and Haybron, 2016; Larroulet Philippi, 2021)), this step arguably represents a crucial precondition for validation: namely, determining the extent to which different measures of the "same" construct in fact produce similar observations. (See Section 4 for more discussion of how to interpret convergence or divergence between tasks.)

In the current work, we draw inspiration from this conceptual and methodological paradigm—and from work in psychometrics more broadly (Flake and Fried, 2020)—to assess the convergent validity between select measures of ToM in LLMs. Specifically, we focus on EPITOME, a collection of tasks adapted from psycholinguistic and developmental research in humans to measure ToM in LLMs (Jones et al., 2024). The inventory includes tasks that emphasize reasoning about false beliefs, as well as others emphasizing pragmatic inference dependent on a speaker's belief states (Trott and Bergen, 2020; Goodman and Stuhlmüller, 2013); each task is also associated with a human gold standard (see Section 2.1 for more details). Importantly, while (Jones et al., 2024) measured the performance of just a single, closed-source LLM (GPT-3), our goal was to measure the correlation in performance between each pair of tasks—and ideally, determine whether a single "latent factor" could be posited to account for performance across the tasks. Capturing potential correlations requires variance in measured performance, so we administered each task to a larger sample of LLMs.

We then addressed several research questions. First, we aimed to replicate the primary analyses in Jones et al. (2024), i.e., comparing the accuracy of each model to the human gold standard and to the original GPT-3 results (Section 3.1). On average, models in our sample under-performed both humans and GPT-3, though select models outperformed GPT-3 on certain tasks (Figure 1a); moreover, larger models performed better on ToM tasks across the board (Figure 1b). Second, and more importantly, we assessed convergent validity by measuring the correlation between performance on each pair of tasks (Section 3.2). Here, we found moderate positive correlations between select pairs of tasks, though correlations were very weak or even negative in other cases (Figure 3a); moreover, each pairwise correlation between tasks was weaker than each task's correlation with model size. Third, we ran a principal components analysis (PCA) to ask whether latent structure could account for the pattern of between-task correlations. We found no evidence for a single driving factor, particularly when other control tasks were included in the correlation matrix (Figure 4), ultimately suggesting that between-task correlations are difficult to disentangle from more global indicators like model size. Finally, in Section 4, we consider various interpretations of these results and discuss both philosophical and empirical directions for future work.

2 Methods

All code and data can be found on GitHub (`https://github.com/UCSD-Language-and-Cognition-Lab/llm-tom`).

2.1 Tasks

We selected four tasks from the EPITOME benchmark: the False Belief (FB) Task (Trott et al., 2023), the Recursive Mindreading (RM) Task (O'Grady et al., 2015), the Indirect Requests (IR) Task (Trott and Bergen, 2020), and the Scalar Implicature (SI) Task (Goodman and Stuhlmüller, 2013).[2] The SI Task contained two experiments, which we analyze separately. We also included a modified False Belief (FB2) Task based on the stimuli for the IR Task (Trott and Bergen, 2020).

We direct readers to the original paper (Jones et al., 2024) for a detailed discussion of each task, but we also provide a brief description here: the FB and FB2 tasks both involve explicitly reasoning about the beliefs of characters in a story, which are (in half the stimuli) false; the RM task also involves reasoning about mental states, but includes recursive beliefs up to seven levels of embedding; and

[2]We excluded the Short Story Task (Dodell-Feder et al., 2013) and the Strange Stories Task (Happé, 1994), which were both free response tasks, for two reasons: first, GPT-3 achieved or even surpassed human performance on both tasks in the original paper, suggesting they are less stringent measures of ToM in LLMs; and second, the context windows of select models in our sample were not large enough to include the entire story.

the IR and SI tasks both involving making inferences about a speaker's intended meaning on the basis of the speaker's likely knowledge states. In this sense, the FB, FB2, and RM tasks are arguably more direct measures of mental state reasoning, and the IR and SI tasks might reflect both mental state reasoning capacity and additional capacities, such as pragmatic reasoning.

Finally, we included four "control tasks" based on the attention checks for the original FB Task (Trott et al., 2023), which assess a model's ability to accurately report: the original location of an object; the final location of an object; the identity of the character who originally placed the object into the starting location; and the identity of the character who moved the object.

2.2 Models

We analyzed 36 models from three language model "families": OLMo 2 (OLMo Team et al., 2025), Pythia (Biderman et al., 2023), and Qwen 2.5 (Qwen Team et al., 2025). These sets of models were selected for several reasons. First, each model was open-weight (OW) and in some cases, open-data (OD), providing a stronger guarantee of future reproducibility. Second, and relatedly, these model suites facilitated comparison of model instances along multiple dimensions, including their size (i.e., number of parameters), amount of training data (i.e., number of tokens encountered), and training regime (i.e., "base" vs. "instruct" models).

2.3 Procedure

Following other recent work (Kouwenhoven et al., 2026; Trott et al., 2026), we ran each model instance using the *transformers* package (Wolf et al., 2020). Models were run either using an NVIDIA DGX-H200 or using the Lambda Cloud service.

Models were assessed using the same approach as in the original EPITOME paper (Jones et al., 2024). Specifically, each stimulus on each task was associated with multiple completions (e.g., correct vs. incorrect), and we compared the *surprisal* associated with each completion type. For instance, in the FB task, a model is asked to make a prediction about whether a character will look in the *original* or *final* location of an object, where the correct answer depends on whether or not the character knows the object has changed location. In this case, a correct answer reflected a relatively higher probability (lower surprisal) assigned to the original location in the False Belief condition, and a relatively higher probability (lower surprisal) assigned to the final location in the True Belief condition. (Note that the Attention Check tasks were assessed using same protocol but with a different probe sentence, e.g., "The original location of the book was").

In the FB2 and IR tasks, the model was asked to answer a yes/no question (i.e., about the character's knowledge state or intended meaning), and the surprisal associated with a "Yes" or "No" completion was compared; accuracy reflected whether a higher probability was assigned to "Yes" when the correct answer was "Yes" (and vice versa). In the RM task, we measured the probability assigned to LLMs to each continuation following the story; here, because the continuations varied in length, we used PMIDC (Holtzman et al., 2021) to control for the probability of a particular string in the absence of the story context. Accuracy corresponded to whether a higher probability was assigned to the consistent continuation. Finally, in the SI task, we calculated the probability assigned by the model to each possible response option, where accuracy was determined by whether and how a speaker's level of knowledge about a situation shifted the model's probability distribution over interpretations of the ambiguous quantifier "some". (See Jones et al. (2024) for more details on each task and analysis procedure.)

2.4 Model Meta-data

Finally, we recorded several dimensions of meta-data about each model, including: the number of parameters, the number of training tokens observed, and whether the model was a *base* or *instruct* version. Note that the number of training tokens had to be estimated in select cases based on the training corpus reported in the corresponding model paper.

3 Results

We carried out several analyses of model performance within and across tasks: first, we describe overall accuracy across models and its relation to other model properties, such as size (Section 3.1); second, we calculate the correlation between performance on various tasks, including both the ToM tasks and the control Attention Check tasks (Section 3.2); and third, we report the results of an exploratory analysis investigating whether and to what extent cross-task correlations can be accounted for by a latent "ToM factor" (Section 3.3).

3.1 RQ1: Accuracy Across Tasks

Figure 1a depicts the accuracy of each model instance tested on the original EPITOME tasks, as well as the mean of the entire sample, the mean of the human sample, and GPT-3's performance (see Jones et al. (2024)). On average, accuracy was highest on the RM task and lowest on the two SI tasks. Most models performed below both GPT-3 and the mean of the human sample, though select models surpassed GPT-3 or even (in the case of the IR task) the human mean. Average accuracy on the full set of tasks (including the modified FB task and the attention checks) is displayed in Table 1.

Table 1: Mean accuracy (and standard error) across models for each task. Attention Checks are averaged across all four control tasks.

Task	Mean	SE
Attention Checks (avg.)	0.70	0.03
RM (Mental)	0.66	0.02
False Belief	0.57	0.01
False Belief (modified)	0.56	0.01
Indirect Request	0.51	0.01
SI (E2)	0.47	0.01
SI (E1)	0.32	0.02

We then asked whether accuracy was reliably predicted by model properties such as size, training data volume, or training regime. Specifically, we regressed composite ToM performance (a model's average performance across all ToM tasks) against Log Parameter Count, Log Training Tokens, and Training Regime (Base vs. Instruct). Both Log Parameter count and Log Training Tokens were positively associated with improved ToM performance (see Figure 1b); models coded as "Instruct" showed marginally better performance, though the coefficient was on the borderline of statistical significance when other covariates were included in the model. We also carried out an identical analysis predicting average accuracy across all tasks, as well as for the control tasks; see Figure 2 for a depiction of the coefficient values and standard errors from each model.

3.2 RQ2: Convergence Between Tasks

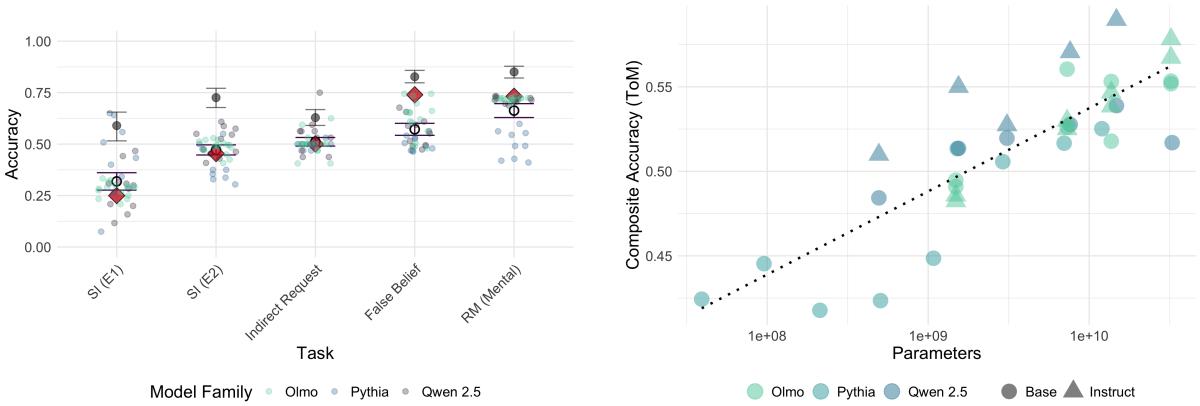
Our primary research questions concerned the relationship between tasks. To this end, we calculated the Spearman's correlation between every pair of tasks.[3] The pairwise correlations of the ToM tasks specifically are depicted in Figure 3a; see Table 3 in the Appendix for the full correlation table with all tasks. As illustrated in Figure 3a, some pairs of ToM tasks exhibited moderate to strong correlations, e.g., FB and FB2 ($\rho = 0.52$), or RM and FB2 ($\rho = 0.6$). This could be interpreted as evidence for the claim that these tasks are measuring similar constructs.

However, other tasks (such as IR or SI-E1) exhibited much weaker or even negative correlations with FB, FB2, and RM. Moreover, the correlation between select ToM tasks and control tasks (A1, A2, A3, and A4) was in some cases even stronger than between ToM tasks, e.g., between FB and A2 ($\rho = 0.72$). Similarly, both FB and FB2 demonstrated very strong correlations with *model size* ($\rho = 0.77, \rho = 0.75$), and RM's correlation with model size was of moderate strength ($\rho = 0.52$). (The correlation between composite ToM accuracy and model size was even larger: $\rho = 0.84$.)

Together, these results point to limited convergent validity across the EPITOME tasks. Even among ToM tasks that did show robust correlations, performance on those tasks was more correlated with both control tasks and with global model properties—consistent with the interpretation that the pairwise ToM task correlations may instead be tracking more general model capacities rather than a ToM-specific construct.

Finally, following recent work (Tsvilodub et al., 2026), we also asked whether average performance on ToM tasks focused on mental state reasoning (FB, FB2, and RM) predicted performance on the pragmatic reasoning tasks (SI and IR); the correlation was indistinguishable from zero, indicating that at least for these tasks, a model's ability to answer explicit questions about mental states was not predictive of its ability to infer a speaker's communicative intent on the basis of the speaker's likely mental states. This is particularly striking for the comparison between FB2 and IR: even in the exact same scenarios, a model's performance on FB2 was uncorrelated with its performance on IR.

[3] Spearman's was used instead of Pearson's *r* because it is unclear whether task results should be viewed as interval/ratio scales or ordinal scales; arguably, there is no theoretical basis for supposing that the difference between 50% performance and 55% performance is the same as that between 90% performance and 95% performance. However, Pearson's *r* did produce similar results.



(a) Accuracy across EPITOME tasks. Each dot represents an individual model. Clear circles represent the mean across model instances for that task. GPT-3 performance (red diamond) and average human performance (dark gray circle) taken from Jones et al. (2024).

(b) Model size was systematically correlated with performance across ToM tasks.

Figure 1: We replicated the primary analyses from Jones et al. (2024) and also reported a correlation between EPITOME performance and model size.

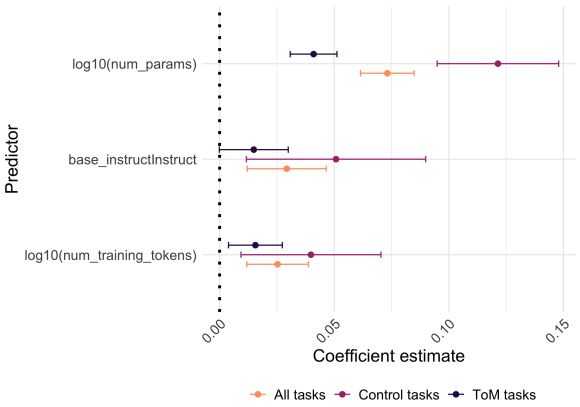


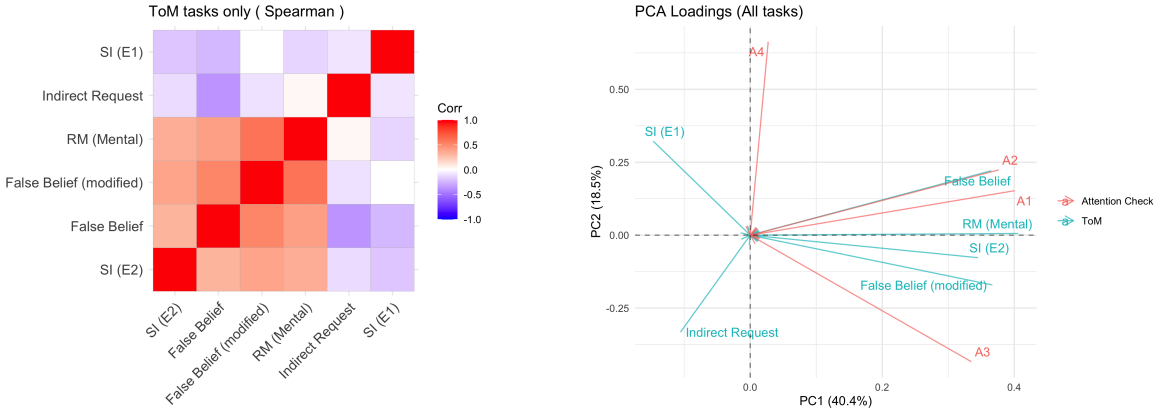
Figure 2: Coefficients from multiple regression models predicting either composite ToM accuracy, average accuracy across all tasks, or average accuracy on the control tasks.

3.3 RQ3: Exploratory Analysis of Latent Components

We then conducted an exploratory analysis, in which we asked whether a latent factor could be found across EPITOME tasks (and control tasks) that accounted specifically for ToM performance, yet was dissociable from properties like model size or performance on the control tasks. Using the correlation matrix constructed in Section 3.2, we performed principal components analysis (PCA) and inspected how each component loaded onto each task (see Figure 4). Together, the top three components accounted for approximately 80% of the variance in the original correlation matrix, yet none of these components could be plausibly interpreted as indexing ToM-specific capacity: for instance, the first PC loaded positively onto almost every task except for IR and SI, both of which were negatively correlated with all other tasks (see also Figure 3b). These results should be viewed with some caution, given the relatively small sample size ($N = 36$) for the number of tasks ($M = 10$); we emphasize that this PCA is intended as exploratory and illustrative, rather than a formal latent-variable analysis. A confirmatory analysis would require a larger model sample, as well as stronger commitments about the latent structure being tested.

4 General Discussion

Research on Theory of Mind in LLMs faces a question of construct validity: are ToM tasks actually measuring what they're designed to measure—and what criteria ought we to use to make this determination? As mentioned in the Introduction, construct validity is a complex issue involving both theoretical, empirical, and contextual considerations (Chang, 2004; Alexandrova and Haybron, 2016; Larroulet Philippi, 2021). In the current work, we focused on the specific dimension of convergent validity: namely, whether different measures of ToM produce correlated observations. The logic of convergent validity is straightforward: if two instruments measure the "same thing", they should exhibit robust "agreement". Convergent validity is by no means a sufficient criterion, but it plays an important role in the fundamentally iterative



(a) Spearman's correlation between all pairs of the selected Theory of Mind tasks.

(b) Each task is depicted in a two-dimensional space defined by the first two principal components, resulting from conducting PCA over pairwise task correlations.

Figure 3: Select EPITOME tasks exhibited some convergence, though this convergence fell short of the correlation between those tasks and control tasks.

process of construct validation (Chang, 2004).

We selected several tasks from the EPITOME benchmark (Jones et al., 2024), a collection of psycholinguistic tasks adapted for ToM assessment in LLMs, and administered them to a sample of LLMs ranging in size, training volume, and training regime. First, we established the range of accuracy on each task, and found that model performance across ToM tasks was strongly correlated with model size (Section 3.1).

We then calculated the correlation between each pair of tasks; select tasks (e.g., variants of the false belief task) exhibited moderate to strong correlations, though two tasks involving pragmatic reasoning were uncorrelated or even negatively correlated with other EPITOME tasks (Section 3.2). This result was particularly striking between the IR and FB2 tasks, which used identical scenarios. The result also diverges from other recent work (Tsvilodub et al., 2026) reporting correlated performance between pragmatic reasoning and ToM tasks. One possibility for this divergence is that the EPITOME pragmatic reasoning tasks involve both pragmatic inference and mental state reasoning, requiring LLMs to first sample mental state information about a character and later deploy this knowledge to infer their likely intent (Trott and Bergen, 2020)—thus, the EPITOME pragmatic reasoning tasks may simply be more demanding in the number of cognitive operations they require (Hu and Frank, 2024).

Further, as we note in Section 3.2, even the strongest pairwise correlations between ToM tasks were weaker than the correlation between those tasks and the control tasks (see Table 3)—or, notably, with the number of parameters in a model (Figure 1b). That is, knowing the size of a model provided more information about that model's likely performance on FB and FB2 than either task provided about the other. Does this entail that even these correlated tasks are therefore not valid measures of ToM or that LLMs do not have ToM? Our position in both cases is "no": it does not, on its own, invalidate these measures, and it also does not adjudicate the case of LLM ToM. However, it does provide evidence against the claim that the observed convergence between tasks is necessarily due to those tasks indexing the same specific construct (Theory of Mind). Instead, it opens up the possibility that the correlation between those tasks could be due to some other common cause, e.g., more general model capabilities.

Which inferences, then, are licensed from these findings? We see at least three interpretations consistent with the observed pattern of selective convergence, partial divergence, and pervasive correlations with model size. These interpretations differ both in their stance on whether some unified ToM construct might exist in (some) LLMs and whether any EPITOME tasks are valid measures of that construct. The first possibility is that some EPITOME tasks (e.g., FB and FB2) are, in fact, genuinely better measures of ToM than others (e.g., IR), as reflected by their agreement with each other; future work should thus aim to determine which tasks those are and what properties specifically make

them better instruments.

The second possibility is that there may be no unified "ToM construct" to be measured in LLMs.[4] If ToM is better understood as a collection of partially dissociable capacities (i.e., *submentalizing* (Heyes, 2014)), then the fact that distinct tasks do not converge should not be particularly surprising. Indeed, according to an *operationalist* view, each measure arguably defines a new construct.

A third possibility is that there is a coherent ToM construct, and that select EPITOME tasks do capture it to some extent, but that these tasks—particularly in this sample of models—are too confounded with model capability to measure it. We intentionally selected a wide swath of models ranging in size and training volume. One unintended consequence of this choice could be that the largest contributor to variance across models was overall model capability (as indexed, roughly, by size). Under this interpretation, a sample with more independent variation in "true" ToM capacity (e.g., among models matched on size) might yield clearer results. The challenge, of course, is determining which sample of models might plausibly vary in ToM specifically but not other potentially confounded measures.

These challenges raise the specter of a problem discussed in the Introduction: empirical validation of novel instruments almost always involves some degree of circularity, at least when the construct must be inferred through a proxy measure (Chang, 2004). Our position is that this challenge is unavoidable and, ultimately, part of what Chang (2004) calls the cycle of *epistemic iteration*. At the same time, successful epistemic iteration does require iterating—the field needs systematic, theoretically motivated comparative work on measures of ToM in LLMs across models and tasks. Indeed, the epistemology of benchmarking practices has undergone a surge of interest in recent years (Bean et al., 2025; Saxon et al., 2024; Raji et al., 2021), and the convergent validity approach demonstrated here could be productively applied to other constructs beyond ToM.

[4] Or, potentially, in humans. The issue of whether ToM exhibited convergent validity (or construct validity more broadly) in humans is a matter of considerable debate (Gernsbacher and Yergeau, 2019; Gough, 2023; Hayward and Homer, 2017). We do not take a stance on this debate here, though it is relevant to calibrating our expectations about convergence between ToM tasks in LLMs.

Limitations

One limitation of the current results is *scope*, both in terms of which tasks and which models we considered. For example, it could be that select tasks would exhibit stronger correlations if a different (or larger) sample of models was used; in the absence of such results, it is difficult to draw generalizations about how other models would likely perform and how that might impact the degree of convergence between tasks (Trott, 2025). Future work could explore a larger range of models and integrate tasks from additional benchmarks, such as OpenToM (Xu et al., 2024).

A related concern is that the pragmatic reasoning tasks—which, notably, displayed the sharpest divergences from other EPITOME tasks—may simply be less *face valid* measures of ToM, given that they involve not only mental state reasoning but also pragmatic inference. Put another way, they involve "auxiliary task demands" that may complicate our ability to directly read off ToM capacity from task performance (Hu and Frank, 2024). Indeed, performance on these tasks was consistently worse than performance on the three mental state reasoning tasks (FB, FB2, and RM). At the same time, the pragmatic reasoning tasks could be viewed as more ecologically valid applications of ToM in the sense that they require LLMs not only to respond to explicit questions about mental states but to *use* that information—in this case, to interpret a speaker's utterance.

This dilemma raises a third limitation, which is that our work focused specifically on diagnosing the convergent validity of ToM tasks. The construct validity of an instrument depends on multiple factors, including its reliability (i.e., whether it produces consistent measurements) and predictive validity (i.e., whether it predicts other outcomes of interest), neither of which were assessed here. Moreover, as others have argued (Alexandrova and Haybron, 2016), construct validity is not merely an empirical question: an instrument should ideally be motivated with respect to some *theory* articulating the phenomenon of interest. While the ToM tasks explored here were theoretically motivated, there is more philosophical work to be done giving an account of which mechanisms constitute ToM and how best to identify those mechanisms in LLMs (Hu et al., 2025).

References

Anna Alexandrova and Daniel M. Haybron. 2016. Is Construct Validation Valid? *Philosophy of Science*, 83(5):1098–1109.

Ian A. Apperly. 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65(5):825–839. Publisher: SAGE Publications.

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.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. The Concept of Validity. *Psychological Review*, 111(4):1061–1071.

Hasok Chang. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press. Google-Books-ID: yVOuV8qJkxMC.

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302. Place: US Publisher: American Psychological Association.

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.

Sebastian Dörrenberg, Hannes Rakoczy, and Ulf Liszkowski. 2018. How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, 46:12–30.

Jessica Kay Flake and Eiko I Fried. 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in methods and practices in psychological science*, 3(4):456–465.

Kanishk Gandhi, Jan-Philipp Fraenken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

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.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Joe Gough. 2023. Does the neurotypical human have a 'theory of mind'? *Journal of Autism and Developmental Disorders*, 53(2):853–857.

Francesca GE 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.

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.

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

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating Theory of Mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499.

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. Place: Cambridge, MA Publisher: MIT Press.

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

Tom Kouwenhoven, Michiel van der Meer, and Max van Duijn. 2026. Traces of social competence in large language models. *arXiv preprint arXiv:2603.04161*.

Cristian Larroulet Philippi. 2021. Valid for What? On the Very Idea of Unconditional Validity. *Philosophy of the Social Sciences*, 51(2):151–175. Publisher: SAGE Publications Inc.

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].

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.

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].

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*.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as Microscopes: A Call for Model Metrology. *arXiv preprint*. ArXiv:2407.16711 [cs].

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.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295. Publisher: Nature Publishing Group.

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.

Sean Trott. 2025. Toward a theory of generalizability in llm mechanistic interpretability research. *arXiv preprint arXiv:2509.22831*.

Sean Trott and Benjamin Bergen. 2020. When do comprehenders mentalize for pragmatic inference? *Discourse Processes*, 57(10):900–920.

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.

Sean Trott, Samuel Taylor, Cameron Jones, James A. Michaelov, and Pamela D. Rivière. 2026. Language Statistics and False Belief Reasoning: Evidence from 41 Open-Weight LMs. *arXiv preprint*. ArXiv:2602.16085 [cs].

Polina Tsvilodub, Jan-Felix Klumpp, Amir Mohammadpour, Jennifer Hu, and Michael Franke. 2026. On emergent social world models–evidence for functional integration of theory of mind and pragmatic reasoning in language models. *arXiv preprint arXiv:2602.10298*.

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint*. ArXiv:2302.08399 [cs].

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.

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.

A Full list of models

B Pairwise correlations

C Additional PCA Results

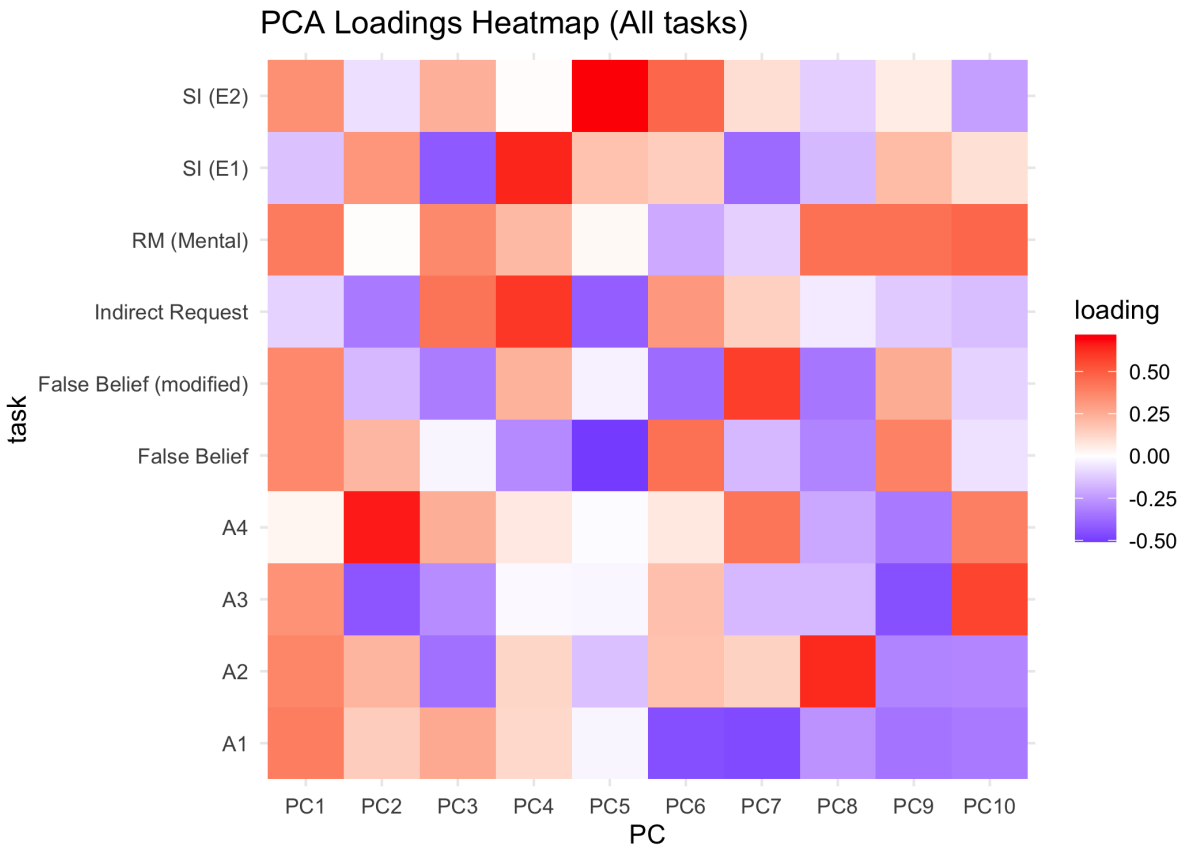


Figure 4: Loadings from each PC onto each task.

Table 2: Models tested, grouped by family.

Family	Size	Variants
OLMo 2		
Olmo	1B	Base, SFT, Instruct, DPO
Olmo	7B	Base, SFT, Instruct, DPO
Olmo	13B	Base, SFT, Instruct, DPO
Olmo	32B	Base, SFT, Instruct, DPO
Pythia		
Pythia	14M	Base
Pythia	70M	Base
Pythia	160M	Base
Pythia	5410M	Base
Pythia	1B	Base
Pythia	1.4B	Base
Pythia	2.8B	Base
Pythia	6.9B	Base
Pythia	12B	Base
Qwen 2.5		
Qwen 2.5	.5B	Base, Instruct
Qwen 2.5	1.5B	Base, Instruct
Qwen 2.5	3B	Base, Instruct
Qwen 2.5	7B	Base, Instruct
Qwen 2.5	14B	Base, Instruct
Qwen 2.5	32B	Base

Table 3: Pairwise correlation between all tasks. FB = False Belief Task, FB2 = Modified False Belief Task, IR = Indirect Request Task, RM = Recursive Mindreading Task, SI (1 and 2) = Scalar Implicature Tasks, A(1-4) = Attention Check Tasks.

	FB	FB2	IR	RM	SI1	SI2	A1	A2	A3	A4
FB	1.00	0.52	-0.37	0.41	-0.24	0.32	0.52	0.72	0.37	0.37
FB2	0.52	1.00	-0.10	0.60	0.00	0.39	0.45	0.65	0.66	-0.09
IR	-0.37	-0.10	1.00	0.03	-0.09	-0.12	-0.33	-0.46	-0.05	-0.33
RM	0.41	0.60	0.03	1.00	-0.14	0.35	0.58	0.36	0.27	0.17
SI1	-0.24	0.00	-0.09	-0.14	1.00	-0.19	-0.08	0.17	-0.01	-0.05
SI2	0.32	0.39	-0.12	0.35	-0.19	1.00	0.41	0.36	0.47	0.04
A1	0.52	0.45	-0.33	0.58	-0.08	0.41	1.00	0.60	0.41	0.31
A2	0.72	0.65	-0.46	0.36	0.17	0.36	0.60	1.00	0.49	0.37
A3	0.37	0.66	-0.05	0.27	-0.01	0.47	0.41	0.49	1.00	-0.39
A4	0.37	-0.09	-0.33	0.17	-0.05	0.04	0.31	0.37	-0.39	1.00