

On Emergent Social World Models — Evidence for Functional Integration of Theory of Mind and Pragmatic Reasoning in Language Models

Polina Tsvilodub* and Jan-Felix Klumpp and Amir Mohammadpour

Department of Linguistics
University of Tübingen

*polina.tsvilodub@uni-tuebingen.de

Jennifer Hu

Department of Cognitive Science
Johns Hopkins University

Michael Franke

Department of Linguistics
University of Tübingen

Abstract

This paper investigates whether LMs recruit shared computational mechanisms for general Theory of Mind (ToM) and language-specific pragmatic reasoning in order to contribute to the general question of whether LMs may be said to have emergent “social world models,” i.e., representations of mental states that are repurposed across tasks (the *functional integration hypothesis*). Using behavioral evaluations and causal-mechanistic experiments via functional localization methods inspired by cognitive neuroscience, we analyze LMs’ performance across seven subcategories of ToM abilities (Beaudoin et al., 2020) on a substantially larger localizer dataset than used in prior like-minded work. Results from stringent hypothesis-driven statistical testing offer suggestive evidence for the functional integration hypothesis, indicating that LMs may develop interconnected “social world models” rather than isolated competencies. This work contributes novel ToM localizer data, methodological refinements to functional localization techniques, and empirical insights into the emergence of social cognition in artificial systems.

1 Introduction

Large language models (LMs) possess astonishing abilities and prove useful for a plethora of downstream tasks, but controversy persists regarding how to conceptualize their capacities. Some view LMs as mere “stochastic parrots,” fundamentally incapable of genuine understanding (Bender and Koller, 2020; Bender et al., 2021). Others argue that highly proficient language ability presupposes the emergence of “world models,” i.e., representations of (relevant aspects of) the language-generating processes (Yudkowsky, 2023; Bereska and Gavves, 2024). Arguments for the latter view typically come from models optimized for solving finite-state problems like Chess (Toshniwal et al., 2022), Othello (Li et al., 2023), or taxi navigation

(Vafa et al., 2024), or from narrow domains like text-based adventure games (Li et al., 2021). But to what extent might large pretrained LMs recruit representations and computational mechanisms suggestive of “world models” for tasks crucial for *general-purpose language understanding*?

Language understanding comprises *formal* and *functional* abilities (Mahowald et al., 2024). Formal abilities (e.g., syntax) may be acquired relatively straightforwardly from local statistical patterns in training data. Functional abilities relate to language use in communication, e.g., *pragmatic reasoning*, such as comprehension of contextual meaning and social implications of utterances (Levinson, 1983). Functional ability seems more difficult to extract from text alone, since the surface language signal provides less direct information about relevant latent variables, such as interlocutors’ mental states. This raises the issue: to what extent do LMs develop rich “social world models”—entertaining concepts of self and other, social commitment, and nested beliefs—in order to achieve their observed level of performance?

We here address this big-picture question with a concrete case study on the relationship between LMs’ capacities for (i) Theory of Mind reasoning and (ii) pragmatic reasoning. Theory of Mind (ToM) refers to the capability to attribute mental states—unobservable, latent properties such as beliefs or desires—to oneself and other agents, and plays a fundamental role in social cognition (Beaudoin et al., 2020) and pragmatic language use (Grice, 1975; Goodman and Frank, 2016). Evidence from cognitive neuroscience indicates that humans indeed have, at least to some extent, shared functional mechanisms, namely, overlapping neural substrates supporting both Theory of Mind and pragmatic reasoning (Saxe and Kanwisher, 2003; Bosco et al., 2017; Hauptman et al., 2023).

What holds for human cognition need not, of course, generalize to LMs. Nothing in their archi-

ecture explicitly enforces structural similarity to human solutions for functional language abilities. Nevertheless, if shared computational mechanisms govern linguistic and non-linguistic social reasoning in humans, efficient compression principles might favor cross-purposing internal mechanisms across these domains in stochastic learners as well. We therefore formulate two contrasting hypotheses. The **functional specialization** hypothesis holds that LMs use different computational mechanisms for pragmatic and ToM-reasoning. In contrast, the **functional integration** hypothesis maintains that computational mechanisms for pragmatic and ToM-reasoning are shared, at least in substantial part.

We explore whether LMs exhibit functional specialization or functional integration through evidence from behavioral evaluations and causal-mechanistic experiments. The latter builds on recent work by AlKhamissi et al. (2025a) on *functional localization* inspired from cognitive neuroscience which identifies subsets of neurons in LMs supporting performance on target tasks (cf., Duan et al., 2025; Hanna et al., 2025). While their work did not find strong evidence for functional ToM localization based on a very small data set, we do find causal evidence for it based on a much larger data set which is part of this paper’s contribution.

The two hypotheses structure our investigation, even if this dichotomy is a simplification in several respects. The repurposing of computational mechanisms likely exists along a continuum rather than as a binary distinction, and the constructed domains of pragmatic and ToM-reasoning may themselves lack internal homogeneity. To acknowledge this general question of the level of granularity at which function sharing should be investigated, we investigate LMs’ ToM capabilities while taking into account the “Aspects in Theory Of Mind Space” (ATOMS) framework. ATOMS identifies seven fine-grained subcategories in ToM abilities (Beaudoin et al., 2020; beliefs, intentions, desires, emotions, knowledge, percepts, mentalistic understanding of non-literal communication; described in detail in Appendix A.1). We conduct evaluations of LMs on datasets tapping into different ATOMS, building on like-minded work (Ma et al., 2023).

Related Work & Contributions Our paper builds on ideas from numerous studies in psychology and cognitive science, showing that Theory of Mind enables humans to interpret others’ actions as causally driven by these mental states (e.g.,

Beaudoin et al., 2020; Goodman et al., 2006; Wimmer and Perner, 1983). A longstanding tradition in philosophy, linguistics, and cognitive science conceptualizes communication as a specialized form of goal-directed action (Austin, 1975; Grice, 1975), such that understanding language is, in principle, structurally similar to interpreting non-linguistic actions, and has also been conceptualized as interlocutors engaging in probabilistic recursive reasoning about each other’s mental states (Grice, 1975; Goodman and Frank, 2016).

Additionally, we contribute to related work evaluating LMs, often separately focusing on ToM and linguistic abilities, in the domain of mechanistic interpretability (Wang et al., 2022), in particular LMs’ language capabilities (Oh et al., 2025) and mental state representations (Bortoletto et al., 2024; Zhu et al., 2024), and to the growing number of behavioral evaluations of LMs’ ToM and pragmatic abilities (which, in part, provided mixed results; e.g., Hu et al., 2023; Ullman, 2023; Sap et al., 2022; Ma et al., 2025). While the functional connection of these two abilities has been addressed more in human research (Baldwin and Saylor, 2005; Hauptman et al., 2023), work on the functional organization of LMs has mostly focused on aligning only LMs’ language representations to human brains (Tuckute et al., 2024; AlKhamissi et al., 2025b). We follow recent arguments for more integrated investigations of linguistic abilities of LMs grounded in cognitive science approaches (Sumers et al., 2023; McCoy et al., 2024).

In sum, this paper contributes behavioral and causal-mechanistic experiments (Sections 2 & 3 respectively), paired with stringent hypothesis-driven statistical testing, that provide suggestive evidence in favor of the functional integration view, thus contributing to the question of emergent social world models. In doing so, the paper contributes novel, empirically grounded ToM localizer data and methodological refinements of the functional localization method.¹

2 Behavioral Evaluation

According to the functional integration hypothesis, if ToM and pragmatic reasoning share computational resources, we expect a systematic relation between performance on these tasks across differ-

¹We release all code and data at <https://github.com/polina-tsvilodub/lm-emergent-social-world-models>.

ent LMs, which we operationalize in three predictions to be tested by statistical analyses:

P1 (simple correlation): Across LMs, accuracies on ToM and pragmatic reasoning are positively correlated.

P2 (predictive redundancy): Under appropriate controls, when predicting accuracy for a task, information about whether a task is a pragmatic or a ToM reasoning task is not relevant for predicting an LMs’ accuracy for it.

P3 (predictive gains): Under appropriate controls, information on an LM’s ToM accuracy is better for predicting its pragmatic accuracy than information on general language ability.

While P1–P2 address the behavioral correlation between ToM and pragmatic capabilities, P3 helps to disentangle this connection from potential confounds: both ToM and pragmatic capabilities likely correlate with general language ability (e.g., for instruction comprehension), as well as on general cognitive abilities and world knowledge. While the following analyses focus on controlling for general linguistic ability through performance on two linguistic benchmarks, we provide additional exploratory mechanistic analyses of the models also on general intelligence tasks like analogical reasoning and entity tracking in Section 3.5 and in Appendix D.1.

2.1 Datasets

We select 16 datasets testing ten different pragmatic phenomena from extant pragmatics benchmarks (Ma et al., 2025), and 22 datasets from existing ToM benchmarks. We use several ToM benchmarks with ATOMS annotations introduced by Ma et al. (2023), and select additional ToM datasets, annotating them with ATOMS we consider the benchmarks to likely test. All datasets have a multiple-choice format: for each item x , given a context c that contains the target vignette, a question, and optionally instructions or the answer options, there are $2 \leq n \leq 6$ answer options, one of which is the correct answer that is consistent with ToM-based or pragmatic reasoning. For evaluating general language ability of each model, the BLiMP dataset (Warstadt et al., 2020) and the SNLI dataset (Bowman et al., 2015) are used. The detailed list of datasets and annotations can be found in Appendix A, full prompts for each dataset are reported in Appendix A.3.

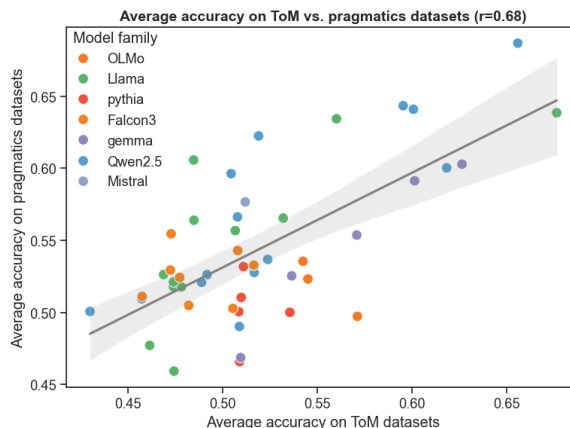


Figure 1: Average accuracy of 48 models on ToM (x-axis) and pragmatics (y-axis) tasks.

2.2 Models & Procedure

We test 48 open-weights and/or open-source LMs from seven families (Llama, Qwen, Falcon, Mistral, OLMo, Pythia, Gemma), ranging from 0.5B to 72B parameters (listed in Appendix A.4). Where possible, we test base and fine-tuned versions, the type of which (instruction-tuning, RLHF, both) depends on the model.

LMs’ predictions are retrieved via conditional log-probability scoring. For each item, we retrieve the log-probability of the tokens of each answer option, conditioned on the input prompt, and average across the token log-probabilities to correct for different lengths of different answer options. The model’s prediction is the option with the highest normalized log-probability. Accuracy is the proportion of correct answers.

2.3 Results

Figure 1 shows the correlation between models’ accuracy on pragmatic and ToM benchmarks. As per prediction P1, we find a moderate but significant positive correlation ($r = 0.68$, $p = 1.24 \cdot 10^{-7}$).

To address prediction P2, we fit a Bayesian beta regression model in R using the brms package (Bürkner, 2017), regressing the average accuracy of each model on each dataset against the main effects of model family (seven levels), model size (where models up to 8B parameters were labeled as small, up to 32B as medium, and up to 72B as large), model type (base vs. fine-tuned), dataset type (instruction / answer options in context vs. sentence completion) and domain (ToM vs. pragmatics), an interaction between model size and type, and ran-

dom by-dataset intercepts.² Here and below, we report posterior means and 95% credible intervals for the posterior of contrasts of interest. All independent variables were sum-coded. In line with most evaluations in the literature, we found that, across domains, fine-tuned models performed credibly better than base models ($\beta = 0.11[0.05, 0.16]$), large models performed better than medium models ($\beta = 0.20[0.11, 0.29]$) and medium models performed better than small models ($\beta = 0.26[0.20, 0.32]$). We found no differences between the dataset types ($\beta = -0.06[-0.88, 0.73]$). Crucially, in line with prediction P2, we also found no difference between the domains ToM and pragmatics ($\beta = -0.03[-0.74, 0.67]$), suggesting that the information about the domain of a given task is not crucial for predicting a given model’s accuracy if sufficient information about the model is given. Further accuracy results on the single datasets are in the Appendix (Figure 5–Figure 8).

Finally, we address P3 with model comparison for two Bayesian beta regression models predicting the average accuracy on pragmatics datasets for each model against the baseline predictors of model size and model type.³ One model (M0) contains the additional predictor of the model’s average accuracy on the linguistic benchmarks. The other model (M1) contains information of the model’s average accuracy on ToM tasks. Comparing M0 and M1 via leave-one-out cross-validation (LOO-CV), we found that model M1 was significantly better than M0, measured through expected log predictive density (ELPD) ($ELPD = -16.1$, p -value = 0), thus supporting prediction P3. Additional details on our statistical analyses are in Appendix A.5.

To explore potential diversity in ToM reasoning further, we also ask whether there are more specific *aspects* of the ToM capabilities that LMs perform particularly well on. We operationalize this question through our annotations of the evaluation datasets with the ATOMS framework (Beaudoin et al., 2020). We compare the baseline beta regression model (MB) where a model’s average accuracy on ToM evaluation datasets is regressed against the model family, size, type and dataset type, to models which additionally include all possible combina-

tions of the seven ATOMS main effects, resulting in 128 different regression models. Comparison was also conducted via LOO-CV. We found that the baseline MB was numerically worse than any model with ATOMS predictors. The best model additionally contained the aspects intentions, desires, emotions, percepts and non-literal communication. The difference between MB and the best model was significant ($ELPD = -58.80$ p -value = 0; full results are in the Appendix, Figure 9). Model comparison further suggests that several models are on par with the best model, and several models next to MB are clearly worse than others. The least performing models do not contain the predictor percepts, while containing up to three additional other ATOMS predictors. This suggests that the factor percepts added the most predictive power to the statistical models of the LMs’ accuracy, specifically distinguishing datasets which tapped into this ToM aspect (false-belief tasks about the location of an object and tasks about agent properties) from other evaluation datasets. Conjointly, these behavioral results provide first evidence for the functional integration hypothesis. Next, we investigate whether the hypothesis is supported at the level of LMs’ internal mechanisms.

3 Functional Localization

Functional subnetwork localization is a method that has been widely used in cognitive neuroscience to identify regions (or, subnetworks) in the human brain that are implied in the execution of a particular task or ability, including Theory of Mind (e.g., Saxe et al., 2006; Fedorenko et al., 2010), and indeed Hauptman et al. (2023) suggest that pragmatic language use in humans relies on the same ToM subnetwork. The procedure uses so-called *localizer suites* that contrast a *target* condition which contains stimuli that require the capacity of interest for processing, against a closely matched *control* condition which does not require it. For example, the ToM network has been localized by contrasting target stories in which an actor holds a false belief (Baron-Cohen et al., 1985) against control stories about pictures containing false information (Zaitchik, 1990). While reading these stories, the brain activity of human participants is recorded and localization proceeds via by-individual identification of regions that are selectively active under the target, but not under the control stimuli. However, given that ToM may itself not be a single monolithic ca-

²The regression model in R syntax: $\text{accuracy} \sim \text{model_family} + \text{size} * \text{model_type} + \text{ds_type} + \text{domain} + (1 | \text{dataset})$. A model with full interaction structure did not converge.

³Including the model family fixed effect led to convergence issues.

pability (see Section 1), cognitive neuroscience has also used other localizers than false belief tasks, e.g., economic games or stories involving moral intent, which have been found to activate the same functional brain region in humans. For our localizations, we build on recent work extending this procedure to language models (AlKhamissi et al., 2025a), described next.

3.1 Localizer Suites

A *localizer suite* L comprises a set of sets of target stimuli $\{S_T^{(j)}\}_{1 \leq j \leq n_j}$ and a set of sets of control stimuli $\{S_C^{(k)}\}_{1 \leq k \leq n_k}$. Localizer suites with $n_j = n_k = 1$ are called *simple*. We synthetically generated four localizer suites based on extant experimental material which was successful in previous fMRI studies for singling out the ToM network in human brains. These prior studies were selected for variety and to cover all of the seven top-level categories of the ATOMS framework, while still sticking with empirically validated material:⁴ The **LatentBeliefs** suite covers the commonly used false-belief items (Saxe and Kanwisher, 2003), **CommunicativeIntent** looks at mental state ascriptions in communication, and might be particularly relevant for comparison to pragmatic reasoning (Bosco et al., 2017), **GameBeliefs** considers ToM reasoning in the context of economic games (trust & ultimatum; Chang et al., 2023), and **MoralIntent** considers issues of responsibility in decision making (Young et al., 2007). Our localizer suites contain different numbers of target and control sets (each called condition): the two simple suites (GameBeliefs and MoralIntent) have target and control sets in which items are exactly paired, two suites (LatentBeliefs and CommunicativeIntent) have several unpaired sets (see Table 1 for examples).

To mitigate data contamination and to control for irrelevant surface-level variation in the localizer materials that may affect LMs more than humans (e.g., differences in semantic content of target vs. control stimuli), we used GPT-5 (not evaluated in this study) with a 5-shot in-context prompt to generate 100 novel synthetic stimuli for each condition in each suite (see Appendix B for all prompts). To critically assess the quality of the novel synthetic stimuli against the established origi-

⁴For example, belief and percepts are relevant for the LatentBelief suite, intentions and knowledge are relevant for the MoralIntent suite, desires and emotions are relevant for the GameBelief suite, and the CommunicativeIntent suite addresses non-literal communication.

nal stimuli, we ran a Principal Component Analysis with two components comparing sentence embeddings from SentenceTransformer’s all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019) of the original and synthetic stimulus stories. For all conditions except HumanDescr from the LatentBeliefs suite, we confirmed that original and synthetic stimuli did not form clearly separate clusters, indicating that the synthetic stimuli are overall comparable to the original ones. For the more divergent HumanDescr condition, the most likely reason is that the original stimuli only described the physical appearance of humans, while the synthetic stimuli covered other aspects as well. This difference, however, does not impact their role as a non-ToM control condition. More details and visualizations of the PCA results are in Appendix C.3.

All localizer items consist of context sentences and end in a task question, which would require a human participant to engage in Theory of Mind reasoning in the target conditions, and multiple choice answer options. More details about all localizer suites are given in Appendix C.

3.2 Localization Procedure in LMs

For an autoregressive transformer model with a hidden dimension d , let $a_{l,t} \in \mathbb{R}^d$ be the output vector of each transformer block of the l -th layer at token t of the input sequence (target or control stimulus). We record the outputs of each unit $i \in \{1, \dots, d\}$, which we call *activations*: $a_{l,t,i}$. To represent activations for the entire stimulus, we use activation at the last token before the model is required to produce the answer. Given a *localizer dataset* with a set of *target stimuli* S_T and a set of *control stimuli* S_C , we generate two sets of activations for every unit (l, i) (i.e., target and control activations). We identify functional subnetworks based on relative differences in activation for S_T in contrast to S_C , as measured by an unpaired or paired Welch’s t -test (the latter used for the simple suites GameBelief and MoralIntent). We write $\mathcal{T}(i, S_T, S_C)$ for the t -test statistic when comparing activations at unit i for target stimuli in S_T and control stimuli in S_C .

We use two statistical approaches for localization. The **simple** approach defines a metric for unit i under localizer L as $m_i^L = \mathcal{T}\left(i, \bigcup_j S_T^{(j)}, \bigcup_k S_C^{(k)}\right)$, i.e., the relevant t -test statistic for the union of all target and control items. We add a stricter **conjunctive** approach, which is common in cognitive neuroscience and uses a so-

LatentBeliefs (Example of GPT-5 generated stimuli based on data from Saxe and Kanwisher, 2003)	
FalseBelief (trgt)	Tom dropped his keys in the bowl by the door before going for a run. While he was out, Mia put the keys into the desk drawer to tidy up. Tom came back a bit later. S / Q: Tom will first look for his keys in the (• bowl • drawer)
Desire (trgt)	Coach Ramirez booked the soccer field for Saturday practice. He told the team they’d be outside unless lightning forced them into the gym. A storm never came, and they kept the plan. S / Q: The coach wanted to practice (• indoors • outdoors)
FalsePhoto (ctrl)	A bus schedule was printed last month. Yesterday the transit agency changed Route 9 to depart an hour later. All the old paper schedules are still circulating. S / Q: On the printed schedule, Route 9 departs (• earlier • later)

GameBeliefs (Example of GPT-5 generated stimuli based on data from Chang et al., 2023)	
GameBelief (trgt)	Nora and Ben play an Ultimatum Game. Nora is the proposer and receives 20 Euros to divide between her and another player. She intends to offer half to Ben, but a glitch records her offer as 1 Euro. Ben sees 1 Euro on his screen, frowns, and rejects immediately. S / Q: Ben evidently believes that (• Nora is trying to cheat him • a glitch reduced the offer)
GameOutcome (ctrl)	Nora and Ben play an Ultimatum Game. Nora is the proposer and receives 20 Euros to divide between her and another player. She intends to offer half to Ben, but a glitch records her offer as 1 Euro. Ben sees 1 Euro on his screen, frowns, and rejects immediately. S / Q: Ben will now have (• 1 Euro • 0 Euro)

Table 1: Example synthetic stimuli for five conditions from two localizer suites. Dots indicate answer options. “S” / “Q:” stand for “Statement” / “question:” used in the prompts. “trgt” indicates a target stimulus, “ctrl” a control stimulus. Examples for all suites are in the Appendix, Table 3.

called *minimum statistic* approach ([Nichols et al., 2005](#)) to detect significant differences between any pair of relevant target and control items. This metric for unit i and suite L selects the lowest t -statistic for all pairwise comparisons: $m_i^L = \min_{j,k} \mathcal{T}(i, S_T^{(j)}, S_C^{(k)})$. For simple suites, the conjunctive approach reduces to the simple.

Based on localizer L , we select the *subnetwork* of units i , such that m_i is statistically significant at the significance level $\alpha = 0.05$ (modulo appropriate controlling for multiple comparisons). If that set comprises more than 1% of all units, we restrict the selection to the top 1% of units with the highest unique absolute measure $|m_i^L|$. For critical ToM localization with each suite L , we also construct a control localization condition based on the *least active* subnetwork by selecting the same number of units with the smallest absolute $|m_i^L|$ -value from the subset of units whose m_i^L is not significant. To investigate the causal role of the localized subnetworks through ablation experiments (Section 3.5), units in the subnetwork are set to zero.

Altogether, we consider eight *localizers* for identifying relevant subnetworks for subsequent analy-

ses. We have five *simple localizers*, four of which are obtained from each of the the four localizer suites, and another one from the union of all target and control sets from all localizer suites (the **all** localizer). We consider three *conjunctive localizers*, two of which come from the LatentBeliefs and the CommunicativeIntent suites. A third conjunctive localizer was constructed from the union of these two suites (the **LB + CI** localizer), as they have non-paired stimuli and were considered to be most relevant for pragmatic reasoning.

3.3 Models and Evaluation Datasets

We run localization experiments on language models from four families: Qwen-2.5, Llama, Falcon-3 and Gemma-2, where both base and fine-tuned models are available (20 models in total). The models were selected among the models evaluated behaviorally (see Section 2.2) to cover a wide range of sizes, while generally performing above chance on ToM and pragmatics. To assess the role of the localized subnetworks for the performance on these domains, for each of the 20 models we selected 19 to 24 datasets on which the model performed above chance from ToM and pragmatics among the 36

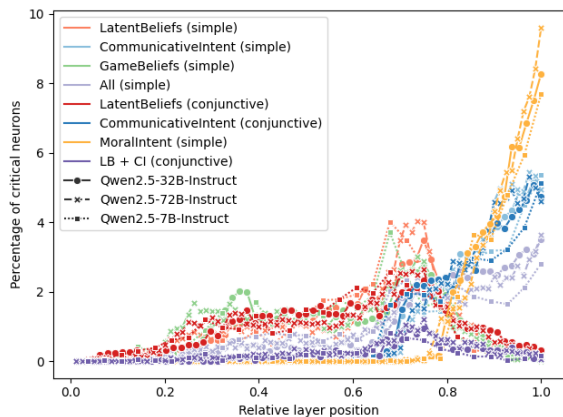


Figure 2: Percentage of localized units across model layers (x-axis) identified by different target subnetwork localizers (colors) for Qwen-2.5 Instruct models (shapes).

datasets used for behavioral evaluations.

3.4 Localization Results

The distributions of units selected by the eight localizers are shown in Figure 2 for three models. The plot suggests differences in the distributions among localizers; e.g., for CommunicativeIntent highly active units are concentrated in the last layers, while for LatentBeliefs mid to late layers are active. Also, the LB + CI localizer seems to be more conservative: often less than 1% of the models’ units were significantly active. The patterns appear to be qualitatively similar across model sizes (see Figures 12–13 in the Appendix for full results).

To assess the general robustness of the localizations, we used 10-fold cross-validation, splitting the localizer stimuli into 90:10 localization and testing splits. We identified active units on the localization split, recorded the activations of these units for the target and control condition of the testing split, and conducted a two-sided t -test ($\alpha = 0.05$) on these test activations for each fold, determining whether the units’ activity is significantly different. Across models, we find the highest generalization rate (≥ 9 significant folds) for the localizers all, MoralIntent and GameBeliefs, high rates (≥ 8 significant folds) for simple CommunicativeIntent and LatentBeliefs, and lower generalization rates (5–7 significant folds) for conjunctive suites.⁵

These results suggest localization is systematic and reproducible, non-random across layers, and that different localizer suites may yield different

⁵Due to a coding error, results for the LB + CI suite could not be included at the time of writing.

localized subnetworks, and localizations with the conjunctive suites may be subject to more noise.

3.5 Ablating Localized Subnetworks

To assess the causal role of the localized networks, we first ablated, for all 20 models and each localizer L , the ToM network identified by L and the “least active” control network for L , and then evaluated the performance on selected datasets as described in Section 2.2 for each ablation.

We test predictions about three causal effects (on ToM, on pragmatics and on baseline general linguistic abilities (BLiMP and SNLI)), which, if observed, conjunctively provide evidence for functional integration (see the upper panel of Figure 3):

- P1.1** ToM-network ablation decreases performance on ToM-related tasks.
- P1.2** CE (causal effect) on ToM-task performance is bigger for ToM ablation than for control.
- P2.1** ToM-network ablation decreases performance on pragmatic tasks.
- P2.2** CE on pragmatic abilities is bigger for ToM ablation than for control ablation.
- P3.1** ToM-network ablation does not decrease performance on baseline linguistic benchmarks.
- P3.2** CE is bigger on pragmatics than on baseline.

These predictions can be tested at different levels of aggregation. For example, results for some localizers may conform to our predictions better than results from others (see the two right-most plots in the bottom row of Figure 3). In light of such variation, the most insightful conclusions are those from a “global analysis,” which marginalizes over all LMs and all localization methods and whose results are shown in the left-most facet in the bottom row of Figure 3. In order to test predictions from the “global analysis” we inspect the relevant posterior probabilities from a Bayesian beta regression model.⁶ Marginalizing across localizers and models, both P1.1 ($\beta = 0.25[0.14, 0.35]$) and P1.2 ($\beta = 0.06[0.02, 0.11]$) were credibly borne out in our results, providing evidence for a general causal role of the localized ToM networks for LMs’ ToM performance. Similarly, both P2.1 ($\beta = 0.30[0.20, 0.39]$) and P2.2 ($\beta = 0.07[0.02, 0.12]$) were credibly borne out, in-

⁶Model in R syntax: $\text{accuracy} \sim \text{model_type} + \text{model_size} + \text{ablation_localizer} * \text{domain}$, where $\text{ablation_localizer}$ is a variable with 17 levels combining the two ablations and eight localizations procedures and the baseline intact condition (results from Section 2.3).

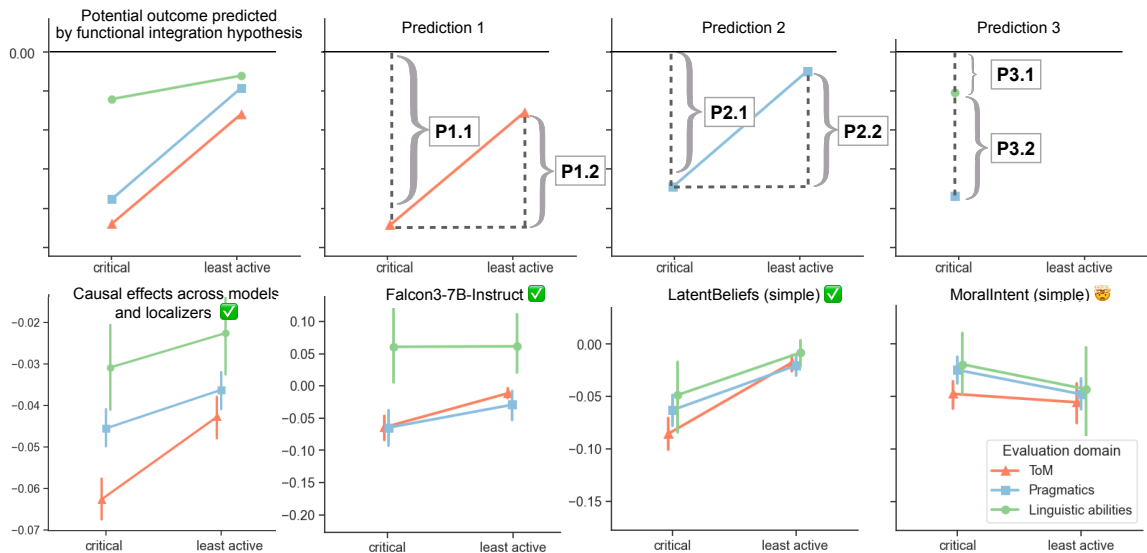


Figure 3: **All:** Plots show the mean causal effects (y-axis), i.e., difference in accuracy relative to the non-ablated baseline when ablating the critical (ToM) subnetworks or the control (least active) sub-networks (x-axis), for different test sets (colors and shapes). 95% CIs indicate change in by-dataset accuracy across localizer suites. **Top row:** Fictitious, stylized result that could come from our experiments (left-most plot), and a visualization of the three main predictions made by the functional integration hypothesis. **Bottom row:** Four observed outcomes at different levels of analysis. Results from the “global analysis” (marginalizing over all language models and localizers) are shown on the left; followed by example results for data from only one language model (averaging over localizers) and two localizer sets (averaging over language models). Predictions are supported for the “global analysis,” but not for all subunits at finer levels of analysis (e.g., the right-most example in the bottom row).

dicating that the ToM subnetworks were generally causally involved in LMs’ pragmatic performance. Finally, the data supported P3.1 (no credible decrease of linguistic benchmark performance under ToM ablation: $\beta = 0.13[-0.10, 0.35]$), but not P3.2 ($\beta = 0.17[-0.07, 0.41]$), offering mild evidence that the causal role of the ToM subnetworks is stronger for pragmatic than general linguistic capabilities, from the “global” point of view. Together, our results support to the idea that ToM and pragmatic reasoning *do* share causally relevant functional units in LMs to a non-trivial degree.

Towards further exploration, we tested predictions separately for each of the eight localizers. Across models, predictions 1.1, 2.1, and 3.1 were borne out for all localizers (except for CommunicativeIntent (conjunctive) (1.1) and GameBeliefs (3.1)). Predictions 1.2 and 2.2 were borne out only for LatentBeliefs (simple & marginally conjunctive) and GameBeliefs, suggesting that these three localizers mainly drive the results above (shown in Figure 14). This suggests that subnetworks tapping into the ToM aspects beliefs, percepts, desires and emotions (captured by these three localizers) might functionally support both ToM and pragmatic performance of LMs. We note that these three local-

izers also showed similar distributional patterns in the models (Figure 2), different from other localizers. Results for all localizers and models are shown in the Appendix in Figures 16–17, full statistical results are presented in Table 5.

Moreover, we investigated to which extent the identified subnetworks functionally support mental state related reasoning as opposed to other cognitive skills that may be recruited in ToM and pragmatic tasks. In addition to the general linguistic ability benchmarks, we evaluated selected models on the entity tracking (Kim and Schuster, 2023) and analogical reasoning tasks (Webb et al., 2023). When ablating the critical and control subnetworks, e.g., localized by LatentBeliefs and GameBeliefs, we found that the critical ablations had a similar effect on entity tracking performance as on ToM and pragmatic performance across models, but did not as consistently affect other datasets, suggesting that ToM and pragmatic capabilities might indeed also rely on shared entity tracking mechanisms. Details of the analysis are reported in Appendix D.1.

For better understanding the impact of the statistical methods for localization, we compared the causal effects of ablations of subnetworks localized with the simple vs. the conjunctive

method for LatentBeliefs (LB) and CommunicativeIntent (CI). Aggregating causal effects on ToM and pragmatics relative to the baseline (P1.1 and P2.1), we found no credible differences for LB ($\beta = 0.05[-0.03, 0.15]$), nor CI ($\beta = 0.06, [-0.03, 0.15]$). Relative to the control ablation (P1.2 and P2.2), we also found no credible differences (LB: $\beta = 0.10[-0.03, 0.23]$, CI: $\beta = 0.07, [-0.06, 0.20]$). The proportion of the units localized by both methods is moderately high (LB = 0.50, CI = 0.66). This suggests that (i) simple and conjunctive strategies may be interchangeable, and (ii) the observed causal effects may be due to only a subset of the localized subnetworks.

Finally, we compared the causal effects of the ToM ablation between different model sizes using a beta regression model with a three-way interaction.⁷ Averaged across localizers, we found no credible differences in the causal effect on ToM and pragmatic performance of models of different sizes (L vs. M models: ToM: $\beta = 0.15[-0.12, 0.42]$, pragmatics: $\beta = 0.13[-0.12, 0.38]$; M vs. S models: ToM: $\beta = -0.04[-0.26, 0.18]$, pragmatics: $\beta = -0.11[-0.41, 0.20]$), suggesting that varying capacity of the models may not directly affect functional integration.

4 Discussion

This paper offered five contributions towards addressing the overarching research question of whether LMs contain “social world models” — latent representations of information related to social reasoning about other minds. (1) We outlined the functional integration hypothesis, according to which internal representations and computational mechanisms are shared between ToM and pragmatic reasoning tasks, and we derived concrete, testable predictions from this view. (2) We report stringent theory-driven statistical analyses based on data from both behavioral evaluations and causal ablation studies which provide suggestive evidence in favor of the functional integration view. (3) Improving on previous work (AlKhamissi et al., 2025a), this paper contributes four novel synthetically generated localization suites with a total of 1400 stimuli based on localizers carefully assembled from prior work in cognitive neuroscience that spans a substantial swath of various aspects of ToM reasoning. (4) We ported established meth-

ods from neuroscience (the conjunctive analysis) into the realm of language models, and presented first results to assess their value. (5) Finally, we explored a finer-grained analysis of ToM based on the ATOMS framework and found that its subcategories are indeed predictive of benchmark accuracies. They also help to interpret the results from the causal ablation studies, where the ablation of localized subnetworks addressing those ATOMS categories that were part of the best explanations of the behavioral data also had the strongest causal effect of interest on LMs’ ToM and pragmatic capabilities, although it remains puzzling why ablating the subnetwork identified by the CommunicativeIntent localizer did *not* have the predicted causal effect.

Even though we found support for various causal effects predicted by a functional integration view, our results hinge on whether the datasets used for localizations and behavioral analysis indeed operationalize the right kinds of concepts (general ToM, specific aspects in the ATOMS framework, pragmatic reasoning etc.). As is the case with human experiments, the validity of methods and operations require converging evidence from many like-minded investigations and replications. What is more, while it is commonly assumed that ToM and pragmatics tasks recruit reasoning about mental causes, we should be open to the possibility that, also in humans, explicit social reasoning trades off with cheaper heuristics (Rubio-Fernandez, 2024). Additionally, both tasks may share not only social reasoning, but likely also other general intelligence capabilities. While the exact nature of such general mechanisms involved in ToM and pragmatics remains elusive (for both humans and machines), our novel large localizer datasets and fine-grained statistical methodology for establishing functional integration between such tasks show one potential path towards contributing to this deep issue.

Together, these considerations underline the importance of an interdisciplinary perspective on language abilities, closely comparing human and machine solutions to superficially similar tasks. In this spirit, the present paper contributes by investigating LMs’ functional mechanisms of different aspects of social reasoning at a high resolution, offering exciting avenues for future investigations of the exact relation between different capabilities and of the architectural pressures shaping LMs’ “social world models.”

⁷Model in R syntax: `accuracy ~ model_type + model_size * ablation_localizer * domain`.

Limitations

The presented functional localizations and evaluations make a number of design choices in the procedure which are known to potentially impact the results (Bommasani et al., 2023), and future work should evaluate the robustness of the results to such design choices.

For the baseline and ablation behavioral evaluations, we chose to use means over the conditional log probabilities of the answer tokens. For the subnetwork localization, we used one general instruction and one specific prompt format across stimuli. All localizer stimuli contained a forced-choice task, which was formulated as a sentence continuation task with two answer options for all localizer suites, except for CommunicativeIntent, for which the task was a meta-linguistic judgment task to select whether an utterance was literal, ironic, deceitful or meaningless.

Furthermore, to ablate the localized subnetwork, the identified units were set to zero; however, alternative approaches also exist (Heimersheim and Nanda, 2024). The size of the ablated subnetworks is also set to the minimum of the top-most active 1% of the units and all significantly active units; however, smaller or larger ablation rates could also be explored (cf. AlKhamissi et al., 2025a).

The ATOMS annotations used for more fine-grained analyses of ToM were, in part, provided by the authors, resulting in a low number of annotations and potential noise due to ambiguity of commonly used ToM and pragmatics evaluation benchmarks. The annotations could be improved upon in future work through, e.g., crowdsourcing.

Furthermore, future work could expand the selection of datasets used to evaluate the general linguistic abilities of models beyond the BLiMP and SNLI datasets. Additionally, all datasets focus exclusively on English, a resource-rich language; in particular, pragmatics datasets in other languages may differ with respect to the cultural and social norms that must be reasoned about. The precision and diversity of the pragmatics datasets may also vary between the selected benchmarks, e.g., due to semi-automated construction of some of the materials. The ToM datasets contain a large proportion of false-belief tasks, whereas real-world language use may provide more evidence for true-belief-based reasoning, which may therefore be strongly represented in LMs’ training data. Finally, since some of our datasets have been publicly available for sev-

eral years, there is a risk that the training data of recent models are contaminated with these datasets.

Although the synthetic generation of novel localizer materials was based on validated materials from neuroscience research and allowed studying language models at a larger scale than related previous work, these new materials have not been validated in human fMRI studies. Therefore, while drawing inspiration from cognitive neuroscience research, our conclusions cannot be directly translated to claims about human-likeness of the functional organization of LMs.

In general, localized subnetwork ablation experiments offer only limited, coarse-grained insights about LMs’ internal organization; achieving a more precise understanding of the causal role of single units for ToM and pragmatic performance will require alternative techniques like causal tracing or mediation methods, such as activation and attribution patching and circuit discovery (Wang et al., 2022).

Acknowledgments

JFK, AM and MF were supported by the Volkswagen Foundation through a Momentum grant. MF is a member of the Machine Learning Cluster of Excellence at University of Tübingen, EXC number 2064/2 – Project number 39072764. PT and MF gratefully acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG, and the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1803-1 FUGG and INST 35/1804-1 LAGG.

References

- Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025a. [The LLM language network: A neuroscientific approach for identifying causally task-relevant units](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.
- Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025b. [From language to cognition:](#)

- [How LLMs outgrow the human language network](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24321–24339, Suzhou, China. Association for Computational Linguistics.
- John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.
- Dare A. Baldwin and Megan M. Saylor. 2005. [Language Promotes Structural Alignment in the Acquisition of Mentalistic Concepts](#). In Janet Wilde Astington and Jodie A. Baird, editors, *Why Language Matters for Theory of Mind*, page 0. Oxford University Press.
- Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a “theory of mind”? *Cognition*, 21(1):37–46.
- Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics.
- Leonard Bereska and Stratis Gavves. 2024. [Mechanistic interpretability for AI safety - a review](#). *Transactions on Machine Learning Research*.
- Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.
- Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024. Brittle minds, fixable activations: Understanding belief representations in language models. *arXiv preprint arXiv:2406.17513*.
- Francesca M. Bosco, Alberto Parola, Maria C. Valentini, and Rosalba Morese. 2017. [Neural correlates underlying the comprehension of deceitful and ironic communicative intentions](#). *Cortex*, 94:73–86.
- Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.
- Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.
- Li-Ang Chang, Konstantinos Armaos, Lotte Warns, Ava Q Ma de Sousa, Femke Paauwe, Christin Scholz, and Jan B Engelmann. 2023. Mentalizing in an economic games context is associated with enhanced activation and connectivity in the left temporoparietal junction. *Social cognitive and affective neuroscience*, 18(1):nsad023.
- Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang Cai. 2025. [Unveiling language competence neurons: A psycholinguistic approach to model interpretability](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.
- Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194.
- 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.
- Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.
- Noah D Goodman, Chris L Baker, Elizabeth Baraff Bonawitz, Vikash K Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua B Tenenbaum. 2006. Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the twenty-eighth annual conference of the cognitive science society*, volume 6. Cognitive Science Society Vancouver.
- Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.
- Paul Herbert Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York.
- 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*.
- Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2025. Are formal and functional linguistic mechanisms dissociated in language models? *Computational Linguistics*, pages 1–40.

- Miriam Hauptman, Idan Blank, and Evelina Fedorenko. 2023. Non-literal language processing is jointly supported by the language and theory of mind networks: evidence from a novel meta-analytic fmri approach. *Cortex*, 162:96–114.
- Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2025. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270.
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. [Are natural language inference models IMPPRESSive? Learning IMPLICature and PRESupposition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.
- 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.
- Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. [Revisiting the evaluation of theory of mind through question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.
- Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.
- Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of ACL*.
- Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Emergent world representations: Exploring a sequence model trained on a synthetic task](#). In *Proceedings of ICLR 11*.
- Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. [Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.
- Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. [Towards a holistic landscape of situated theory of mind in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. [Dissociating language and thought in large language models](#). *Trends in Cognitive Sciences*, 28(6):517–540.
- Nicole Maslan, Melissa Roemmele, and Andrew S Gordon. 2015. One hundred challenge problems for logical formalizations of commonsense psychology. In *AAAI Spring Symposia*.
- R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.
- Thomas Nichols, Matthew Brett, Jesper Andersson, Tor Wager, and Jean-Baptiste Poline. 2005. Valid conjunction inference with the minimum statistic. *Neuroimage*, 25(3):653–660.
- Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2025. Tug-of-war between idiom’s figurative and literal meanings in llms. *arXiv preprint arXiv:2506.01723*.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. [Creating and characterizing a diverse corpus of sarcasm in dialogue](#). In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

- Paula Rubio-Fernandez. 2024. [Cultural evolutionary pragmatics: Investigating the codevelopment and co-evolution of language and social cognition](#). *Psychological Review*, 131(1):18–35.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. [EmoBench: Evaluating the emotional intelligence of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. 2025. A systematic review on the evaluation of large language models in theory of mind tasks. *arXiv preprint arXiv:2502.08796*.
- Rebecca Saxe, Matthew Brett, and Nancy Kanwisher. 2006. [Divide and conquer: A defense of functional localizers](#). *NeuroImage*, 30(4):1088–1096.
- Rebecca Saxe and Nancy Kanwisher. 2003. [People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”](#). *NeuroImage*, 19(4):1835–1842.
- Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. [How well do large language models perform on faux pas tests?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.
- Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. [PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.
- Theodore Summers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. [Cognitive architectures for language agents](#). *Transactions on Machine Learning Research*.
- Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2022. [Chess as a testbed for language model state tracking](#). *Preprint, arXiv:2102.13249*.
- Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024. [Language in brains, minds, and machines](#). *Annual Review of Neuroscience*, 47(2024):277–301.
- Tomer Ullman. 2023. [Large language models fail on trivial alterations to theory-of-mind tasks](#). *arXiv preprint arXiv:2302.08399*.
- Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. 2024. [Evaluating the world model implicit in a generative model](#). In *Proceedings of NeurIPS 38*.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. [Interpretability in the wild: A circuit for indirect object identification in GPT-2 small](#). *arXiv preprint arXiv:2211.00593*.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. [Emergent analogical reasoning in large language models](#). *Nature Human Behaviour*, 7(9):1526–1541.
- 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.
- 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.
- Liane Young, Fiery Cushman, Marc Hauser, and Rebecca Saxe. 2007. [The neural basis of the interaction between theory of mind and moral judgment](#). *Proceedings of the National Academy of Sciences*, 104(20):8235–8240.
- Eliezer Yudkowsky. 2023. [Gpts are predictors, not imitators](#). *lesswrong.com*.
- Deborah Zaitchik. 1990. [When representations conflict with reality: The preschooler’s problem with false beliefs and “false” photographs](#). *Cognition*, 35(1):41–68.
- Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. [Language models represent beliefs of self and others](#). *arXiv preprint arXiv:2402.18496*.

A Behavioral Evaluation

A.1 ATOMS Categories

The ATOMS framework (Beaudoin et al., 2020) is an approach to classify ToM experiments based on the specific capabilities they investigate. While Beaudoin et al. (2020) use the framework to mostly classify ToM measures into disjoint categories (such that each task is placed in the best-fitting ATOMS category only), in practice, many experiments address more than one aspect of ToM. We follow Ma et al. (2023) in that we characterize ToM tasks according to all ATOMS aspects for which we deem them relevant.

The ATOMS framework encompasses seven broad categories, each with subdivisions: 1) The **beliefs** category targets the capability to ascribe beliefs to actors that may be different from one’s own beliefs, such as required in the Sally-Anne task (Baron-Cohen et al., 1985). This also includes second-order beliefs, i.e., beliefs about other people’s beliefs. 2) The **intentions** category targets the capability to ascribe intentions to people and to distinguish these intentions from their actions. 3) The **desires** category encompasses tasks that require ascribing desires to actors. 4) **Emotions** refers to the ability to predict other people’s emotions from events, and to understand that other people may hide or regulate their emotions such that they are not directly accessible. 5) The **knowledge** category concerns being aware of different people having different knowledge which may result in different actions. 6) **Percepts** include perspective taking (visual and otherwise), and being aware that actors’ actions may differ depending on what they have perceived. 7) The **mentalistic understanding of non-literal communication** category includes tasks involving irony, humor, or lies. For detailed introductions of the categories, see Beaudoin et al. (2020); Ma et al. (2023).

The ATOMS categories (on a coarse-grained level) addressed by the test datasets used in this study are listed in Table 2.

A.2 Datasets

A.2.1 Theory-of-Mind Datasets

BigToM (Gandhi et al., 2023) is a benchmark on social reasoning / ToM capabilities. The individual items in the dataset are paired false-belief and true-belief situations generated by an LLM in accordance to a predefined framework. The questions in BigToM involve either forward ToM, i.e., in-

ferring a person’s beliefs based on what they did or did not observe, and backward ToM, i.e., inferring a person’s beliefs based on their subsequent actions. The items contain two answer options per item. We consider an LM’s prediction for a single item correct only if the correct answer options are selected both for the false-belief and the true-belief situations.

EmoBench (Sabour et al., 2024) is a benchmark on emotional intelligence assessing both emotional understanding and its application. The EmoBench tasks focus on aspects such as complex emotions, perspective taking, emotional cues, personal beliefs and experiences, and subcategories of these aspects. The associated questions are phrased as multiple choice questions, where we compare the correct answer with all incorrect answers and consider a model’s choice as correct if the conditional probability of the correct answer, given then context, is higher than all conditional probabilities of the incorrect answers.

EPITOME (Jones et al., 2024) is a battery of six ToM-related experiments, of which we use two (in addition to the scalar implicature task described in the following section): recursive mindreading and false belief. The recursive mindreading task tests the representation of recursively embedded mental states. The false belief task uses a setting similar to the Sally-Anne task (Baron-Cohen et al., 1985). All questions are phrased as multiple-choice questions.

EWoK (Ivanova et al., 2025) is a framework for evaluating world modeling abilities. It consists of eleven tasks, out of which we use three: agent properties, social properties, and social interactions. The agent properties split requires to estimate which one of two propositions about the properties of an agent (such as mental states or preferences) is true, given some context. In the social properties split, the two alternatives are socially relevant agent properties (e.g., character traits). In the social interactions split, there are two agents, and the answer alternatives are possible social relations between them (e.g., cooperating or competing with each other). Each item contains two contexts and two answer options, each being the correct answer for one of the contexts. We apply the same procedure for retrieving predictions for each item as for BigToM.

FauxPasEAI (Shapira et al., 2023) is a dataset intended for evaluating the ability to recognize situationally inappropriate utterances. The original

Dataset	beliefs	intentions	desires	emotions	knowledge	percepts	NLC
BigToM (Gandhi et al., 2023)							
• forward	✓	×	×	×	×	✓	×
• backward	✓	×	×	×	×	✓	×
EmoBench (Sabour et al., 2024)	✓	×	×	✓	×	×	×
EPITOME (Jones et al., 2024)							
• false belief	✓	×	×	×	×	×	×
• recursive mindreading	✓	×	×	×	×	×	×
EWoK (Ivanova et al., 2025)							
• agent properties	✓	×	×	×	✓	✓	×
• social properties	×	×	×	✓	×	×	×
• social interactions	×	✓	×	✓	×	×	×
FauxPasEAI (Shapira et al., 2023)							
• Q1 & Q4	✓	✓	×	×	×	×	✓
OpenToM (Xu et al., 2024)							
• coarse-grained first-order	✓	×	×	×	×	✓	×
• coarse-grained second-order	✓	×	×	×	×	✓	×
• fine-grained first-order	✓	×	×	×	×	✓	×
• fine-grained second-order	✓	×	×	×	×	✓	×
• multi-hop first-order	✓	✓	✓	✓	×	✓	×
• multi-hop second-order	✓	✓	✓	✓	×	✓	×
• attitude	✓	×	✓	✓	×	✓	×
SimpleToM (Gu et al., 2024)	×	✓	×	×	×	×	×
Social IQa (Sap et al., 2019)	×	✓	×	✓	×	×	×
ToMi (Le et al., 2019)							
• first-order false-belief	✓	×	×	×	×	×	×
Triangle-COPA (Maslan et al., 2015)	×	✓	×	✓	×	×	×

Table 2: Classification of ToM datasets according to the ATOMS framework (Beaudoin et al., 2020). NLC stands for “non-literal communication”. This table takes inspiration and, in part, annotations from those in Ma et al. (2023) and Saritaş et al. (2025), and includes our own annotations of additional datasets and partial revisions of the annotations.

dataset has four questions Q1-Q4, of which we exclude Q2 because it is an open-ended question and Q3 because it does not directly concern ToM abilities. The remaining questions Q1 and Q4 are both “yes/no”-questions, where Q1 targets recognizing whether there is a situationally inappropriate utterance, while Q4 concerns the knowledge available to the person uttering it.

OpenToM (Xu et al., 2024) is a ToM benchmark for LLM evaluation consisting of several ToM-related subtasks where an object is moved in the presence or absence of observers (following the schema of the Sally-Anne test, Baron-Cohen et al., 1985). The dataset encompasses 1) first-order false belief tasks with both coarse-grained questions (which ask about somebody’s belief whether the location has changed) and fine-grained questions (which ask about somebody’s belief about the exact location), 2) second-order false belief tasks, 3) multi-hop tasks, which are modifications of the false-belief tasks aiming to address the understanding of social norms, and 4) an attitude task that concerns the attitude of an observer towards the mover’s action.

SimpleToM (Gu et al., 2024) is a dataset designed to test LLMs’ abilities to infer mental states implicitly. This is tested using three types of ques-

tions, namely, mental state questions, behavior prediction questions, and judgement questions. We only use the behavior prediction split, where the task is to predict which of two actions a person is more likely to perform next in some context (based on implicit reasoning about their mental state).

The **Social IQa** dataset (Sap et al., 2019) is designed to test language models’ competence in the areas of social and emotional intelligence. It consists of descriptions of situations and multiple-choice questions about aspects such as motivations, reactions, and social consequences.

ToMi (Le et al., 2019) is a synthetically constructed dataset of situations resembling the Sally-Anne test (Baron-Cohen et al., 1985), i.e., where a false-belief is created by an object being moved in the absence of a person. The dataset encompasses questions about first- and second-order true and false beliefs as well as “reality” and “memory” questions. We only use first-order false-beliefs from (Sap et al., 2022), as is also done by (AlKhamissi et al., 2025a). For this dataset, we additionally use the following instruction and format following AlKhamissi et al. (2025a): “The following multiple choice questions is based on the following story. The question is related to Theory-of-Mind. Read the story and then answer the ques-

tions. Choose the best answer from the options provided by printing it as is without any modifications. \n Story: {story} \n Question: {question} \n Options: {answer options} \n Answer:”.

The **Triangle-COPA** dataset (Maslan et al., 2015) is a commonsense psychology dataset measuring the ability to infer mental or emotional causes and social relations from actions. The agents are represented as geometric shapes (e.g., triangles), which perform some action. A following binary choice question then offers two motivations for the observed behavior, and the task is to identify the more plausible of them.

A.2.2 Pragmatics Datasets

From **EPITOME** (Jones et al., 2024) (see also the section above), we include the scalar implicature task among the pragmatics experiments. This task tests whether *some* is more likely to be interpreted as meaning *not all* when the speaker has complete knowledge (because then they could have said *all* as well).

The dataset from Hu et al. (2023), here called **HuFloyd**, consists of short stories involving non-literal utterances, and questions about the actual intention of the person uttering them. For each story and utterance, the answer options include the target intention, the literal interpretation, and non-literal distractors (except for the coherence task, where there are only two answer options). The dataset targets a range of pragmatic phenomena, namely, deceptions, indirect speech, irony, maxims, metaphor, humour, and coherence.

IMPRES (Jeretic et al., 2020) consists of an implicature and a presupposition task, both of which are also part of the PUB benchmark, through which we access this dataset (Srivanthi et al., 2024, see below). In the implicature task, a premise and a hypothesis are given which involve a scalar implicature trigger. The task is then to judge whether the hypothesis is true given the premise or not. The presupposition task is structured analogously, but involves presupposition triggers. We modify the questions and answer options relative to PUB such that instead of asking whether the hypothesis is true or not true, they ask whether it follows from the premise (such that the answer options are: “The hypothesis contradicts the premise.”, “The hypothesis follows from the premise.”, “The hypothesis is in no clear relationship with the premise.”).

The dataset called **LUDWIG** (George and Mamidi, 2020) assesses the ability to interpret con-

versational implicature correctly. Each stimulus consists of a short dialogue, where a question is followed by a response that does not directly entail “yes” or “no”, but from which the answer can be inferred via conversational implicature. The task then is to judge whether the response means “yes” or “no”.

PUB (Srivanthi et al., 2024) is a large benchmark of 14 pragmatics-related tasks for LLM evaluation. Of these tasks, apart from the two ImpPres tasks described above, we use the following four: 1) indirectness classification (i.e., distinguishing direct from indirect responses), 2) agreement detection (i.e., assessing whether two speakers agree), 3) understanding sarcasm, 4) deictic question answering (i.e., correctly inferring whether an object is in some location based on deictic expressions in a dialogue). All these tasks are phrased such that there are exactly two answer options. For agreement detection, we modify the answer prefix to be “Speaker 2”, and the answer options to be “agrees” and “disagrees”. For indirectness classification, we modify the answer prefix to be “The response is”, and the answer options to be “direct” and “indirect”.

SarcV2 (Oraby et al., 2016) is a corpus of utterances from social media data, which are to be classified into sarcastic and not-sarcastic utterances. The corpus is divided into three subcorpora (generic, hyperbole, and rhetorical questions). We use all three subcorpora without making more fine-grained distinctions.

A.2.3 Other Datasets

For comparison, we use items from two datasets testing LMs’ general linguistic performance. First, across behavioral evaluations and functional localization, we use **BLiMP** (Warstadt et al., 2020), a widely known benchmark on grammatical phenomena in English. **BLiMP** consists of minimal pairs, where in each pair, one sentence is grammatical and one is ungrammatical. LMs are evaluated by comparing the probabilities they assign to the grammatical and ungrammatical sentences. We use a subset of 10 sample items per category. Second, for behavioral evaluations we also use **SNLI** (Bowman et al., 2015), a natural language inference benchmark in English. It consists of pairs of sentences, a premise and a hypothesis. The task is to identify whether the relation between these two sentences is entailment, contradiction or in neutral relation. The models are evaluated by comparing the con-

ditional probabilities the LM assigns to the labels (e.g., “The hypothesis follows from the premise.”). The label to which the model assigns the highest probability given the sentence pair is the model’s prediction. We use a subset of 670 items, to match the size of the BLiMP dataset.

A.3 Prompt Formatting

The following datasets contain and answer options which can be considered simple continuations of the statement or partial sentence posing the target task: EPITOME, EWoK, SimpleToM, BigToM, HuFloyd, and LUDWIG. When models that have chat templates were evaluated on these datasets, the scored answer options were formatted as part of the user turn content.

For the remaining datasets, the optional task, the context and the question or trigger statement were part of the user content, while the answer options were scored as the assistant turn content.

No additional task instructions or prompt formatting were used for any datasets beyond the instructions or formats provided in the original sources of the datasets or described above. Only for the following datasets the answer prefix “The answer is: ” was used: FauxPasEAI, SocialIQa, EmoBench.

A.4 Models

The following models from seven families were evaluated in the behavioral analysis. All models’ weights were accessed through the HuggingFace platform. The HuggingFace model names are listed below.

OLMo: allenai/OLMo-1B-0724-hf, allenai/OLMo-7B-0724-hf, allenai/OLMo-7B-0724-Instruct-hf

Pythia: EleutherAI/pythia-1b-deduped, EleutherAI/pythia-1.4b-deduped, EleutherAI/pythia-2.8b-deduped, EleutherAI/pythia-6.9b-deduped, EleutherAI/pythia-12b-deduped

Mistral: mistralai/Mistral-7B-Instruct-v0.3

Llama-2: meta-llama/Llama-2-7b-hf, meta-llama/Llama-2-7b-chat-hf, meta-llama/Llama-2-13b-hf, meta-llama/Llama-2-13b-chat-hf, meta-llama/Llama-2-70b-hf, meta-llama/Llama-2-70b-chat-hf

Llama-3: meta-llama/Llama-3.2-1B, meta-llama/Llama-3.1-8B, meta-llama/Llama-3.1-8B-Instruct, meta-llama/Llama-3.1-70B, meta-llama/Llama-3.3-70B-Instruct

Falcon: tiuae/Falcon3-1B-Base, tiuae/Falcon3-1B-Instruct, tiuae/Falcon3-3B-Base,

tiuae/Falcon3-3B-Instruct, tiuae/Falcon3-7B-Base, tiuae/Falcon3-7B-Instruct, tiuae/Falcon3-10B-Base, tiuae/Falcon3-10B-Instruct

Gemma: google/gemma-2-2b, google/gemma-2-9b, google/gemma-2-9b-it, google/gemma-2-27b, google/gemma-2-27b-it

Qwen: Qwen/Qwen2.5-0.5B, Qwen/Qwen2.5-0.5B-Instruct, Qwen/Qwen2.5-1.5B, Qwen/Qwen2.5-1.5B-Instruct, Qwen/Qwen2.5-3B, Qwen/Qwen2.5-3B-Instruct, Qwen/Qwen2.5-7B, Qwen/Qwen2.5-7B-Instruct, Qwen/Qwen2.5-14B, Qwen/Qwen2.5-14B-Instruct, Qwen/Qwen2.5-32B, Qwen/Qwen2.5-32B-Instruct, Qwen/Qwen2.5-72B, Qwen/Qwen2.5-72B-Instruct

A.5 Statistical Analyses

Below, we provide details on how the predictions P1–P3 in Section 2 are operationalized and tested statistically. For P2, we used a Bayesian Beta regression model, which predicts the average accuracy of an LM on a given dataset (dependent variable, bound between 0 and 1) according to a Beta distribution, where the predictor for the mean of the distribution is parameterized by several independent variables (i.e., intercept, model family, model size, model training type, dataset type, dataset domain) respective regression coefficients are estimated. For P2, we hypothesized the following: if computational mechanisms are shared between pragmatics and ToM, there will be no credible differences in the accuracy on ToM vs. pragmatics tasks, when controlling for other factors which may lead to differences independently of these mechanisms (e.g., model size), which will be borne out statistically as a non-credible regression coefficient for the predictor “domain” (ToM or pragmatics), which we report in Section 2.3.

For P3, we used two versions of a Bayesian Beta regression model predicting the independent variable “accuracy on pragmatics” (for each dataset for each model) with the dependent variables model size, model type, and either the average accuracy on linguistic benchmarks BLiMP and SNLI (model M0), or on ToM (model M1). As standard in leave-on-out cross-validation, we calculated the expected log predictive density (ELPD) of each of the two models, by iteratively fitting each model on the empirical data but a left out data point (done through the R package brms, Bürkner, 2017). We then compared the two models by evaluating the log

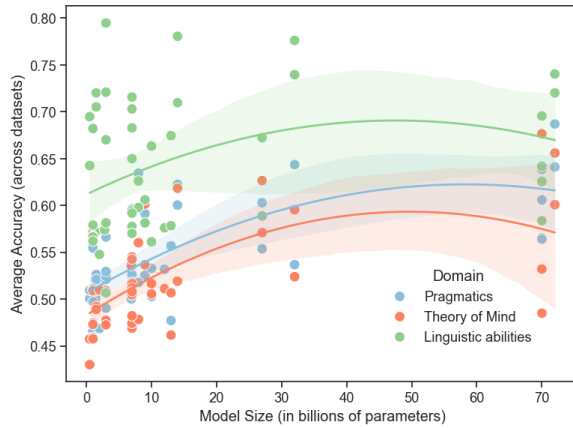


Figure 4: Average accuracy (y-axis) for the domains pragmatics, theory of mind and general linguistic abilities (color) by model size (x-axis).

predictive density of the held-out observation under each model, computing a p -value for this using Lambert’s z -score method.

A.6 Additional Results

We show the models’ average accuracies on the linguistic benchmarks BLiMP and SNLI, next to the accuracies on ToM and pragmatics in Figure 4. Consistent with the literature, accuracy on the complex domains ToM and pragmatics increases over model size. However, just model size or model type (base or fine-tuned) are not sufficient for explaining away the connection between pragmatic and ToM performance of LMs, as shown via model comparison in Section 2.3.

B Prompts for Stimulus Generation

For the generation of synthetic stimuli for the conditions FalseBelief, Desire, FalsePhotograph, HumanDescr, NonhumanDescr, MechInf, GameBelief, and MoralIntent, we used the following prompt:

Please generate twenty more stories and questions following the structure of the given examples:

—
{examples}

The examples consisted of five randomly selected items from the original or manually adjusted dataset, separated by "—" and each structured as follows:

Story:
{story}

Question:
{question}

For the conditions Deceptive, Ironic, Literal, and Meaningless, where the question is the same for all items, the following prompt was used:

Please generate twenty more stories following the structure of the given examples:

—
{examples}

The examples again were five randomly selected story items from the original dataset, separated by "—" and each prefixed by "Story:"

For GameOutcome, which is paired with GameBelief by using the same stories but different questions, we used the following prompt:

For the given story, generate a question analogous to the examples.

Examples:

{examples}

—

Story:

{original story}

Here, the examples were five randomly drawn items (each consisting of the story and one question) from the original GameOutcome data, and the original story was a single story item from the synthetic GameBelief data.

For DecisionOutcome, which is paired with moral intent by sharing the same overall setting (but not the whole narrative text), we used the following prompt:

Convert the given story into a similar story (accompanied by a corresponding question) where no moral decision is involved, in analogy to the given examples.

Examples:

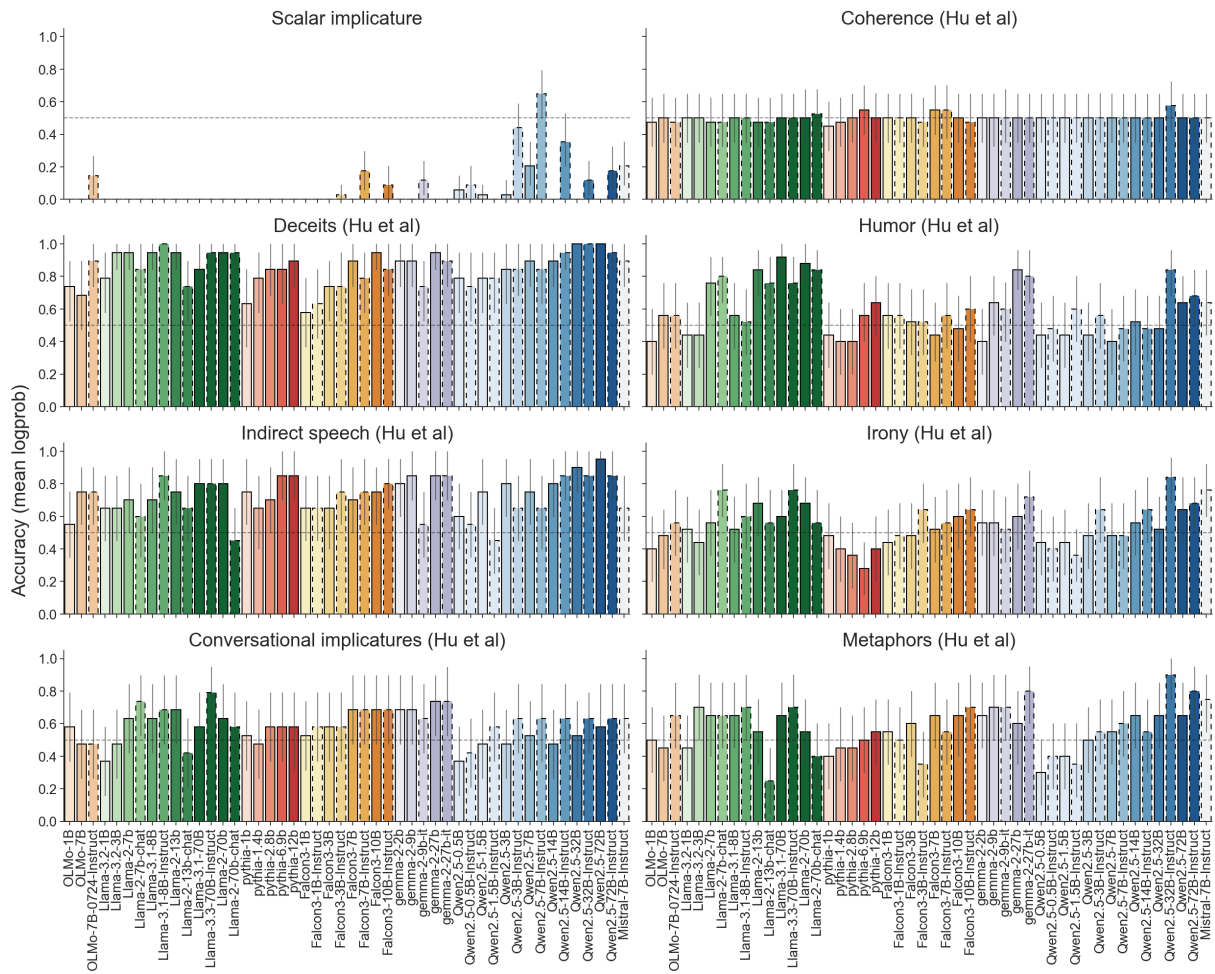


Figure 5: Detailed overview of the accuracy (y-axis) results of all models (x-axis) on the first eight pragmatic datasets (facets).

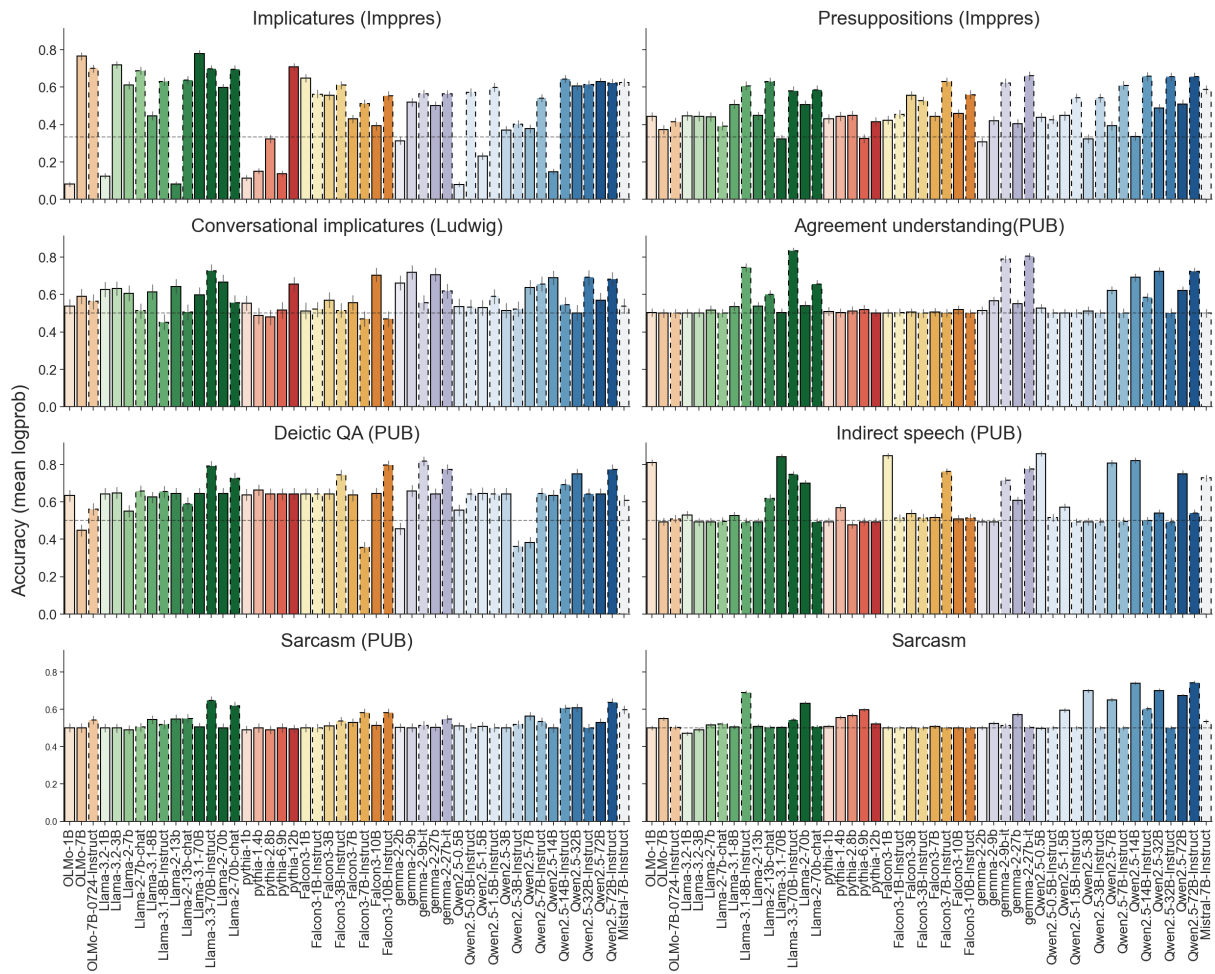


Figure 6: Detailed overview of the accuracy (y-axis) results of all models (x-axis) on the remaining eight pragmatic datasets (facets).

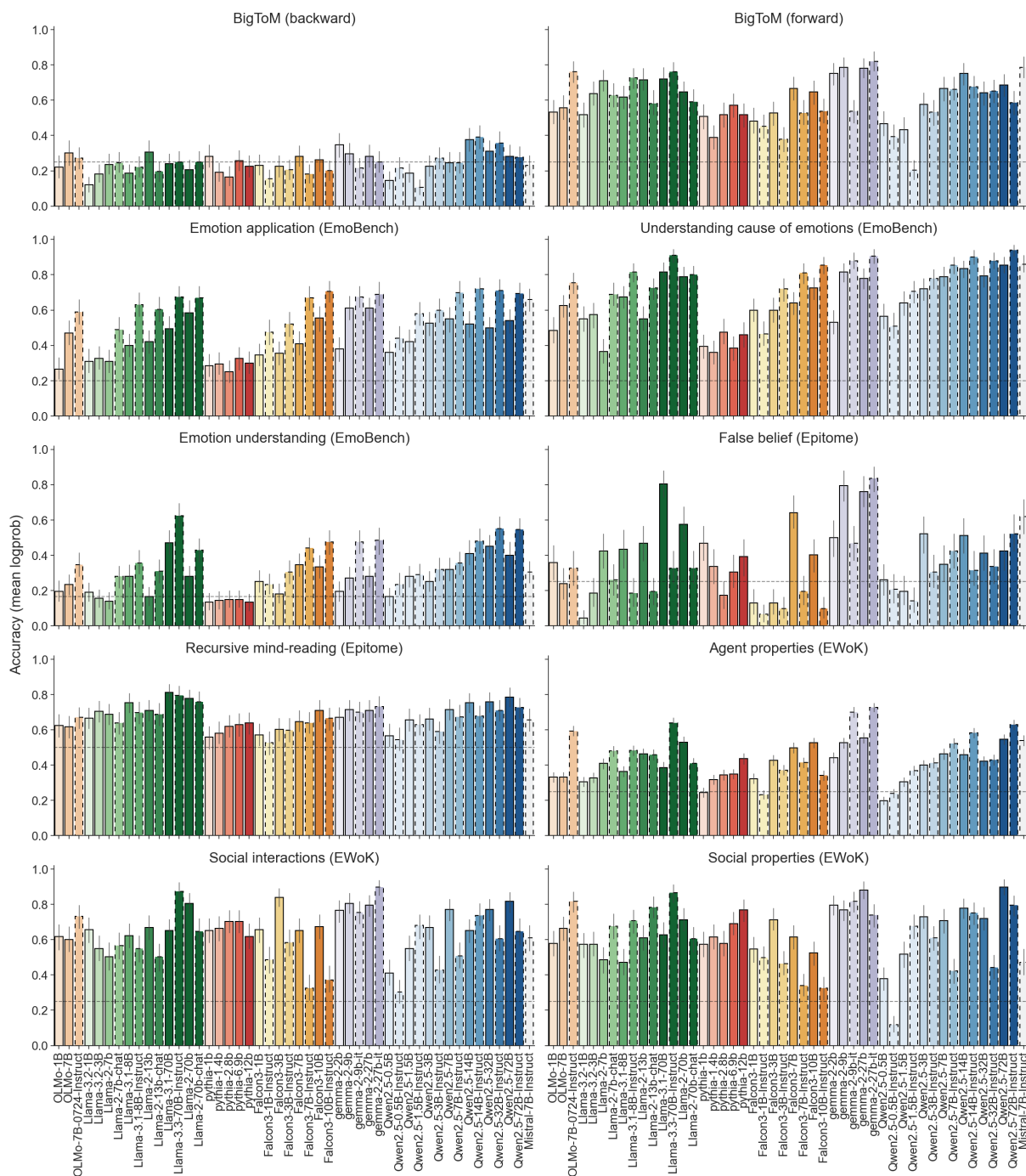


Figure 7: Detailed overview of the accuracy (y-axis) results of all models (x-axis) on ten of the ToM datasets (facets).

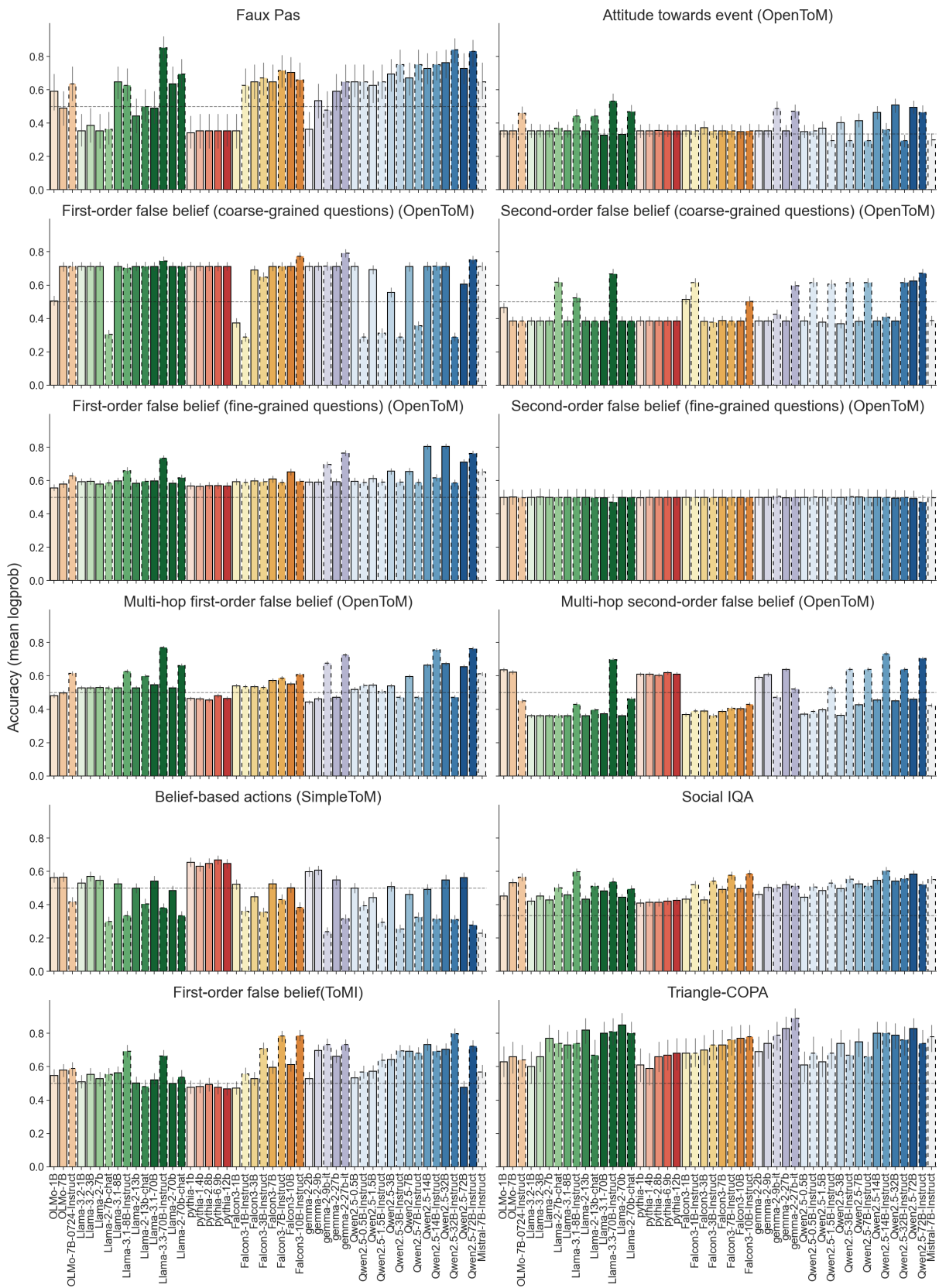


Figure 8: Detailed overview of the accuracy (y-axis) results of all models (x-axis) on the remaining twelve ToM datasets (facets).

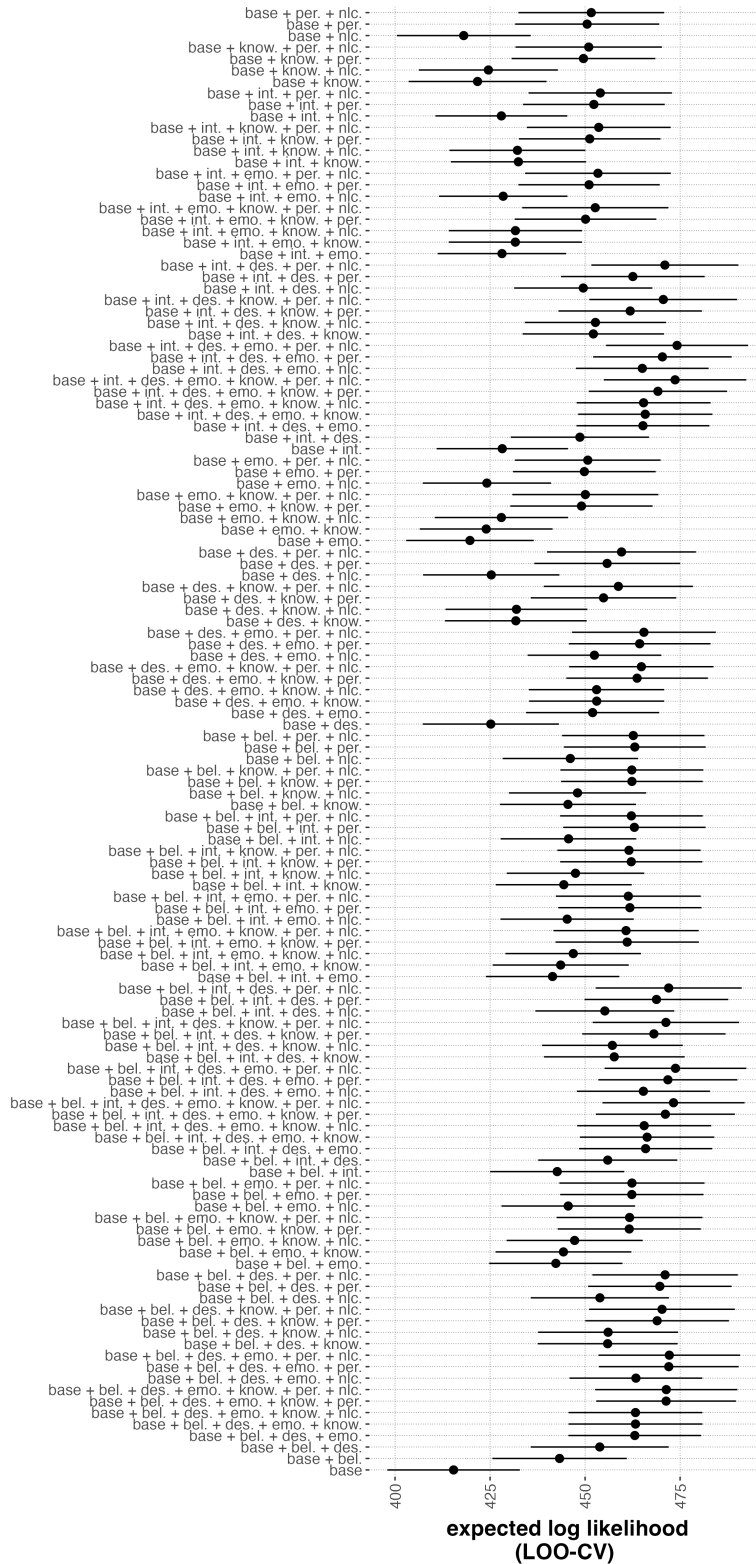


Figure 9: Statistical comparison of models with different combinations of up to seven ATOMS predictors with LOO cross-validation, showing the estimated expected log-likelihoods for LOO-CV (x-axis) for different regression models (y-axis). Higher values indicate a model’s better predictive fit to the behavioral accuracy on ToM datasets. All models contain baseline predictors model family, size, type and dataset type (summarized as “base”). The ATOMS predictors are abbreviated (“bel.” = belief, “des.” = desire, “emo.” = emotions, “int.” = intentions, “know.” = knowledge, “per.” = percepts, “nlc.” = non-literal communication). The best model contains the predictors “base + int. + des. + emo. + per. + nlc”, the worst model is the “base” only model.

{examples}

—

Original story:

{original story}

Here, the examples consisted of five triplets of a MoralIntent story and a DecisionOutcome story and question from the original or manually constructed data (introduced by "Original:", "New story:", and "Question:" respectively), and the original story was a single story item from the synthetic MoralIntent data.

C Localizer Suites

C.1 Details on each Localizer Suite

1. **LatentBeliefs:** to localize the ability to recognize latent mental action causes, we generate synthetic stimuli based on 6 items from [Saxe and Kanwisher \(2003\)](#) where a person acts out of a false belief (henceforth called the FalseBelief condition) and on 6 items from the same paper where a person's desires do not match with the actual course of events (the Desire condition). The corresponding control stimuli are based on four other conditions from [Saxe and Kanwisher \(2003\)](#) where no reasoning about mental causes is required: the FalsePhotograph condition involves photographs or other kinds of depictions that describe a state that does not exist anymore (i.e., false content), the HumanDescr condition involves physical descriptions of people (to contrast mental causes from mere mentioning of human attributes), the NonhumanDescr condition involves descriptions of non-human objects, and the MechInf conditions requires inferring a latent mechanical (i.e., non-mental) cause of an event. Since there are no questions given for the MechInf stories in [Saxe and Kanwisher \(2003\)](#), we manually add corresponding questions to align the format to that of the other conditions.⁸
2. **CommunicativeIntent:** to localize the ability to recognize latent mental states in communication, we use synthetic stimuli based on data

⁸We only use a subset of 9 of the original mechanical inference stories where the actual hidden cause is unambiguous.

from [Bosco et al. \(2017\)](#). Following [Bosco et al. \(2017\)](#), we contrast a Deceptive condition, where an utterance is used in a deliberately misleading way, and an Ironic condition, where an utterance is used ironically, with control conditions where an utterance is meant literally (Literal) or where it is not related to the context at all (Meaningless).

3. **GameBeliefs:** to localize the ability to reason about the latent causes for agent behavior in the specific setting of strategic games, we use synthetic stimuli based on data from [Chang et al. \(2023\)](#). We generate synthetic stimuli for a GameBelief condition, where reasoning about the motivation for a specific decision in a trust game or ultimatum game is necessary, as well as paired GameOutcome stimuli, which concern the same game settings but merely involve the outcome of the game rather than the player's motivations. This suite is paired.
4. **MoralIntent:** to localize the ability to ascribe moral intent to actors, we generate synthetic stimuli based on 12 stimuli from [Young et al. \(2007\)](#), where a person chooses to perform an action despite being aware of its harmful consequences for another actor (MoralIntent). We base our synthetic control stimuli on manually created paired control items, where a decision accidentally leads to a harmful outcome for the acting person themselves (DecisionOutcome). This suite is paired.

Table 3 includes one example stimulus for each condition, and Table 4 contains the ATOMS categories (on a coarse-grained level) addressed by the localization target conditions.

C.2 Formatting in Localization Procedures

For all localizer suites, the following instruction was prepended to each item before they were passed through the model and the activations were retrieved:

In this experiment, you will read a story, a question or a statement and select the best answer among the provided options.

Additionally, for the GameBeliefs localizer, the following instructions were appended to the general instruction (selected to match the game mentioned in each particular item):

Ultimatum game: *The story is about the Ultimatum Game. There are two players who divide a sum of money in this game. The first player, the proposer, proposes a division of the sum with the second player, the responder. The responder can either accept the proposed division or reject it. If the responder accepts, the money is split according to the proposal; if the responder rejects, neither player receives anything.*

Trust game: *The story is about the Trust Game. There are two players who play an investment game. Both players are given some quantity of money. The first player is told that they must send some amount of their money to an anonymous second player, though the amount sent may be zero. The first player is also informed that whatever they send will be tripled by the experimenter and given to the second player. The second player is then told to also give some amount of the now-tripled money back to the first player, even if that amount is zero.*

The localization stimuli were formatted into the following template, resulting in the final language model input: *{instruction}Story: {story}\n Statement / question: {question} \n Options:\n{options}\n*, where the answer options were formatted as a bullet point list with dashes.

For all language models that expect a special template for formatting inputs, i.e., when the corresponding tokenizer provides a chat template, the inputs were put into the chat template as the user turn of the conversation. Then, an assistant turn was appended with the following content: *“Answer: {answer prefix}”*. The answer prefix contained the sentence that was to be completed with the answer options of the questions, but without any answers. The parameter to continue the final message was set to True. The last token at which the activations were retrieved was therefore set up to, ideally, concentrate representations of the target latent information required to answer the question.

C.3 Principal Component Analyses of Synthetic Localizers

To assess the GPT-5 generated localizer stimuli in comparison to original localizer materials from the literature in terms of surface-level and statistical properties to which LMs might be particularly sensitive, a Principal Component Analysis (PCA)

with two components was performed. For each condition of each suite, sentence embeddings for the story part (i.e., the context narrative without the question or answer options) of all original stimuli and an equal number of randomly selected synthetic stimuli were calculated using the Sentence-Transformer’s all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019). The results of the PCA analysis of the original and synthetic data for each condition are showing in Figures 10–11. Visual inspection confirms that original and synthetic stimuli do not cluster separately, indicating that there is no systematic mismatch between them.

Only for one of the conditions (HumanDescr from the LatentBeliefs suite), the two groups can clearly be separated; manual inspection suggests that the difference lies in the fact that the original stimuli describe the physical appearance of a person only, while the synthetic stimuli include other characteristics such as habits. Since the task at hand (i.e., answering simple questions about the described person) does not require Theory-of-Mind reasoning in either case, this does not affect using the synthetic stimuli as a non-ToM control condition. For two other conditions (FalsePhotograph from the LatentBeliefs suite and DecOutcome from the MoralIntent suite), some weak clustering is visible, but original and synthetic stimuli are not clearly separated, and the amount of clustering differs strongly between different samples of synthetic stimuli. Finally, for GameBelief and GameOutcome from the GameBeliefs suite, the sentence embeddings form two clusters, both including a similar number of original and synthetic stimuli; this can be explained by the presence of two types of strategic games in the dataset (trust games and ultimatum games).

D Subnetwork Localization Procedure Details

We select the following 20 models for our localization experiments.

Llama: Llama-3.1-8B, Llama-2-13b-hf, Llama-2-13b-chat-hf, Llama-3.1-8B-Instruct, Llama-3.1-70B, Llama-3.3-70B-Instruct

Falcon: Falcon3-7B, Falcon3-7B-Instruct, Falcon3-10B, Falcon3-10B-Instruct

Gemma: gemma-2-9b, gemma-2-9b-it, gemma-2-27b, gemma-2-27b-it

Qwen: Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-32B, Qwen2.5-32B-Instruct, Qwen2.5-

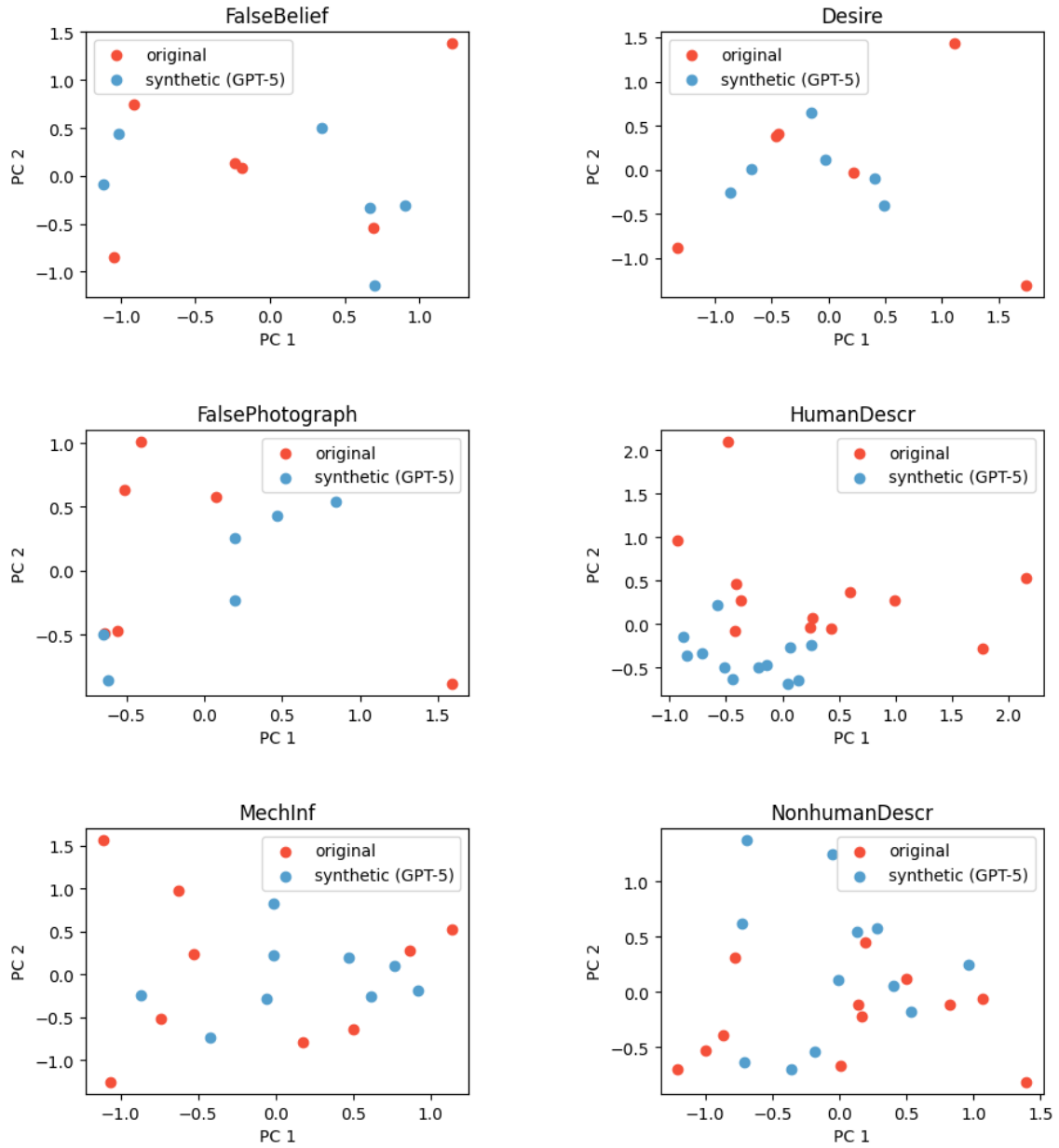


Figure 10: PCA with two components of sentence embeddings (from SentenceTransformer’s all-MiniLM-L6-v2 model) of the original stimuli and an equal number of randomly selected synthetic stimuli, for the conditions FalseBelief, Desire, FalsePhotograph, HumanDescr, MechInf, and NonhumanDescr (all from the LatentBeliefs suite).

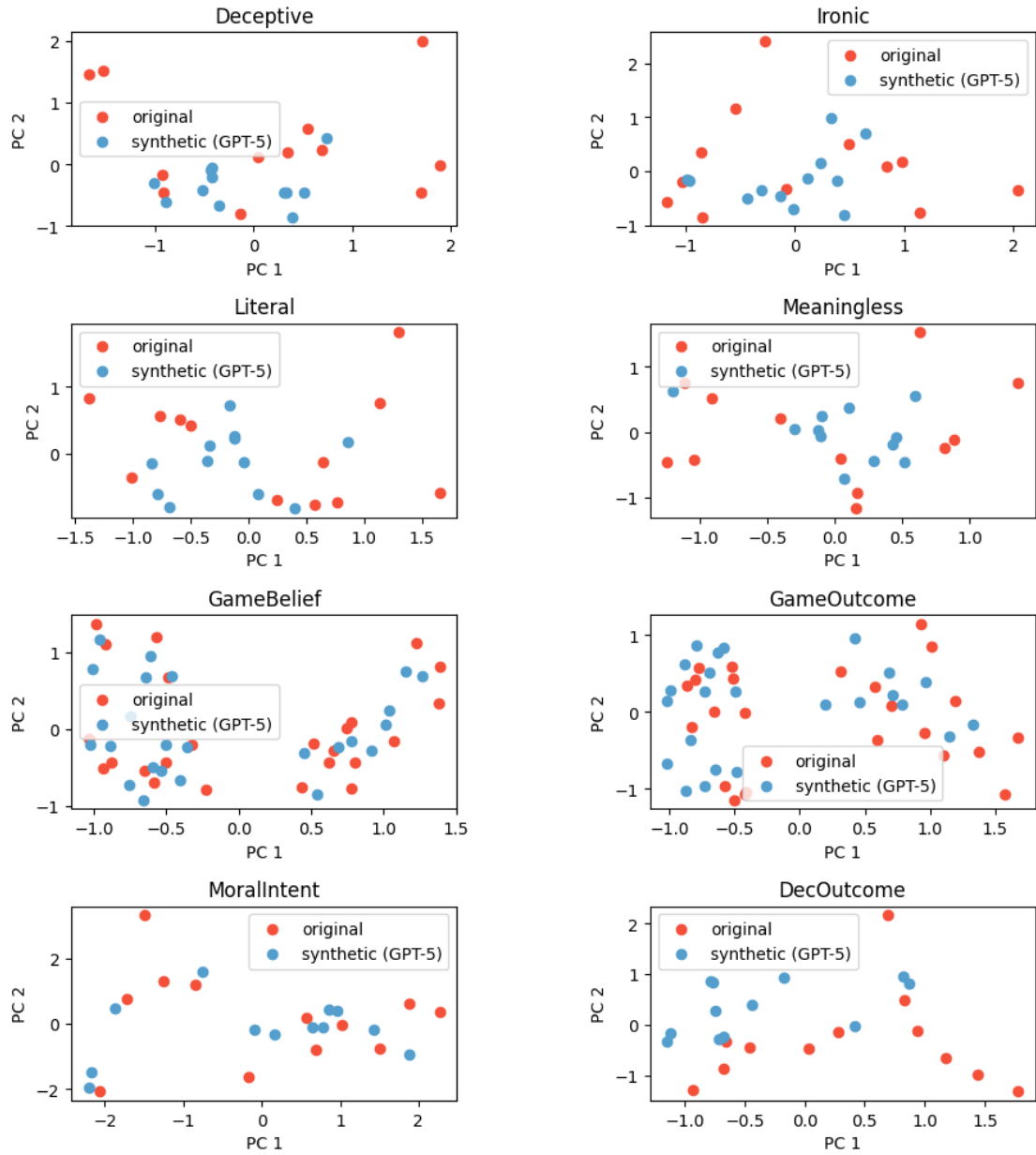


Figure 11: PCA with two components of sentence embeddings (from SentenceTransformer’s all-MiniLM-L6-v2 model) of the original stimuli and an equal number of randomly selected synthetic stimuli, for the conditions Deceptive, Ironic, Literal, Meaningless (from the CommunicativeIntent suite), GameBelief, GameOutcome (from GameBeliefs), MoralIntent, DecOutcome (from MoralIntent, respectively).

72B, Qwen2.5-72B-Instruct

In the ablation experiments (Section 3.5), each model was evaluated on BLiMP (Warstadt et al., 2020) and SNLI (Bowman et al., 2015), and on the following selected ToM and pragmatics datasets on which the intact models showed an above-chance performance (described in detail above in Section A.2):

- Llama-3.1-8B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Llama-3.1-8B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (understanding sarcasm), SocialIQA, ToMi (first-order false-belief), Triangle-COPA
- Llama-2-13b-hf: BigToM (forward), EmoBench (emotion application), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (understanding sarcasm), SocialIQA, ToMi (first-order false-belief), Triangle-COPA
- Llama-2-13b-chat-hf: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Llama-3.1-70B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), LUDWIG, PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Llama-3.3-70B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Falcon3-7B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Falcon3-7B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA

- (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), OpenToM (attitude), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Falcon3-10B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
 - Falcon3-10B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), OpenToM (attitude), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
 - all gemma models: ToMi (first-order false-belief), PUB (agreement detection), PUB (indirectness classification), FauxPasEAI (Q1 & Q4), LUDWIG, OpenToM (attitude), HuFloyd (irony), IMPPRES (presupposition), PUB (understanding sarcasm), PUB (deictic QA), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), EmoBench (understanding emotion), Triangle-COPA, EWoK (agent properties), HuFloyd (indirect speech), HuFloyd (deceits), BigToM (forward), EWoK (social interactions), SocialIQA, EPITOME (recursive mindreading), EmoBench (emotion application)
 - Qwen2.5-7B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
 - Qwen2.5-7B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, PUB (deictic QA), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
 - Qwen2.5-32B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
 - Qwen2.5-32B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, PUB (deictic QA), SocialIQA, Triangle-COPA

- Qwen2.5-72B: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA
- Qwen2.5-72B-Instruct: BigToM (forward), EmoBench (emotion application), EmoBench (understanding emotion), EPITOME (recursive mindreading), EWoK (agent properties), EWoK (social interactions), FauxPasEAI (Q1 & Q4), HuFloyd (deceits), HuFloyd (indirect speech), HuFloyd (irony), HuFloyd (Gricean maxims), HuFloyd (metaphor), IMPPRES (implicature), IMPPRES (presupposition), LUDWIG, OpenToM (attitude), PUB (agreement detection), PUB (deictic QA), PUB (indirectness classification), PUB (understanding sarcasm), SocialIQA, Triangle-COPA

D.1 Subnetwork Ablation Results on Additional Datasets

Next to assessing the effect of the subnetwork ablation on ToM and pragmatics performance in comparison to general linguistic ability benchmarks, we evaluated 13 models on additional cognitive skills that may be recruited in ToM and pragmatic tasks, namely on the entity tracking and analogical reasoning benchmarks.

We cast both benchmarks into multiple choice format. For the entity tracking benchmark from [Kim and Schuster \(2023\)](#), we use the items from the test dataset from their experiment 1, and convert the masked item format into the following sentence completion prompt:

Given the description after “Description:”, write a true statement about all boxes and their contents according to the description after “Statement:”. Format the statement as Box 0 contains the A, Box 1 contains the B, Box 2 contains the C, Box 3 contains the D, Box 4 contains the E, Box 5

contains the F, Box 6 contains the G. A, B, C, D, E, F, G are placeholders for the contents of each box.

Description: Box 0 contains {A}, Box 1 contains {B}, ...

Statement: Box {X} contains

The performance is assessed by retrieving the conditional probability of each of the m different answer options under the LM, where m is the number of boxes described in the item, one of which described the correct content of the target box X . We match the size of the linguistic benchmarks and sample 670 items from the original dataset.

For analogical reasoning, the multiple choice formatted verbal analogy and story analogy tasks from [Webb et al. \(2023\)](#) were selected. For verbal analogy, the task is to identify a correct completion which is analogous to an example pair (e.g., given “clothing : jacket :: bird : ”, select among the options pigeon, dog). LMs are evaluated by scoring each of the possible completions, given the example and first member of the target pair. For story analogy, the task is to decide, given a story, which of two additional stories is a better analogy to the first story, or if they are equally analogous. We largely build on the prompts from [Webb et al. \(2023\)](#).

We evaluate the following models on these three datasets: Falcon3-7B, Falcon3-7B-Instruct, Falcon3-10B, Falcon3-10B-Instruct, Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-2-13b-hf, Llama-2-13b-chat-hf, Llama-3.1-70B, Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-32B, Qwen2.5-32B-Instruct. Each model is evaluated without interventions as well as under ablations of the critical and least active networks localized by each of our localizers. Results for each of the localizers, aggregated across models are shown in Figure 18. For most localizer suites, including the LatentBeliefs (simple & conjunctive) and GameBeliefs suites which mainly supported our overall results reported in Section 3.5, the causal effect of the critical subnetwork ablation was largest for the performance on entity tracking. For analogical reasoning, the effect was smaller (e.g. for GameBeliefs) or qualitatively different (e.g., for CommunicativeIntent conjunctive) than for the critical evaluation domains pragmatics and ToM. This suggests that ToM and pragmatic capabilities might indeed also rely on shared entity tracking mechanisms, in addition to social

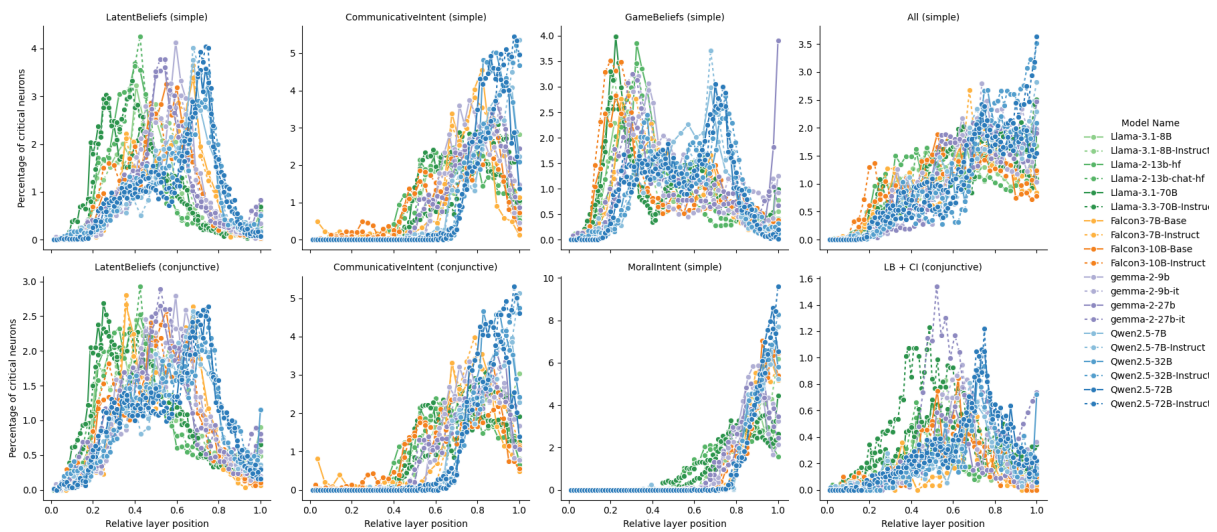


Figure 12: Distribution of top 1% of most active units across the model layers, identified with respect to the synthetically generated ToM localizer suite through last-token pooling. The mask ‘theory-of-mind’ identified most significant units under simple selection, the mask ‘theory-of-mind-conjunctive’ identified most significant units under conjunctive selection, and the random masks identified the least significant units with respect to the same selection approaches, respectively.

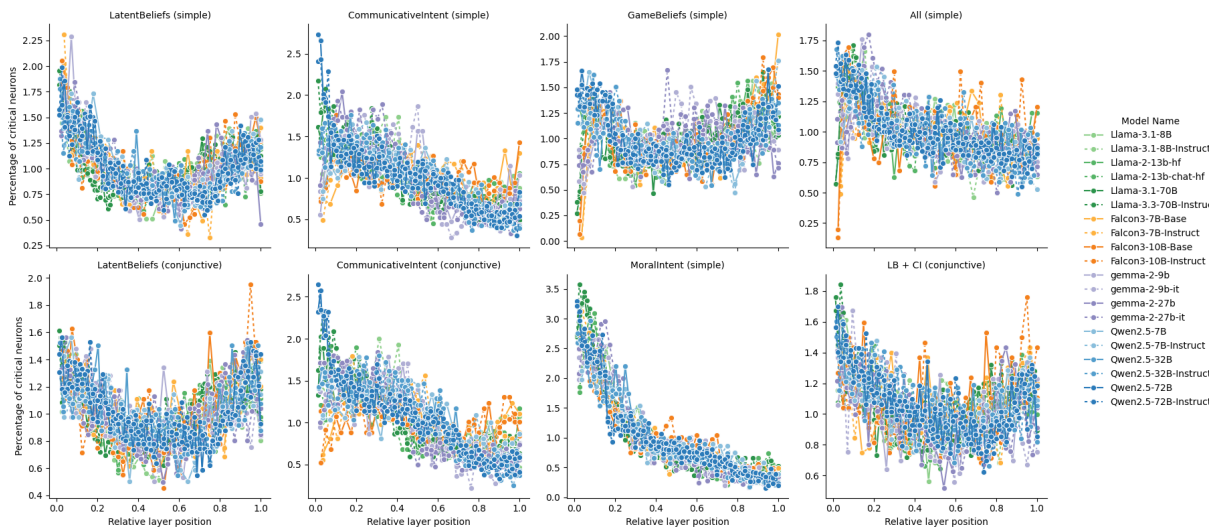


Figure 13: Distribution of top 1% of least active units across the model layers, identified with respect to the synthetically generated ToM localizer suite through last-token pooling. The mask ‘theory-of-mind’ identified most significant units under simple selection, the mask ‘theory-of-mind-conjunctive’ identified most significant units under conjunctive selection, and the random masks identified the least significant units with respect to the same selection approaches, respectively.

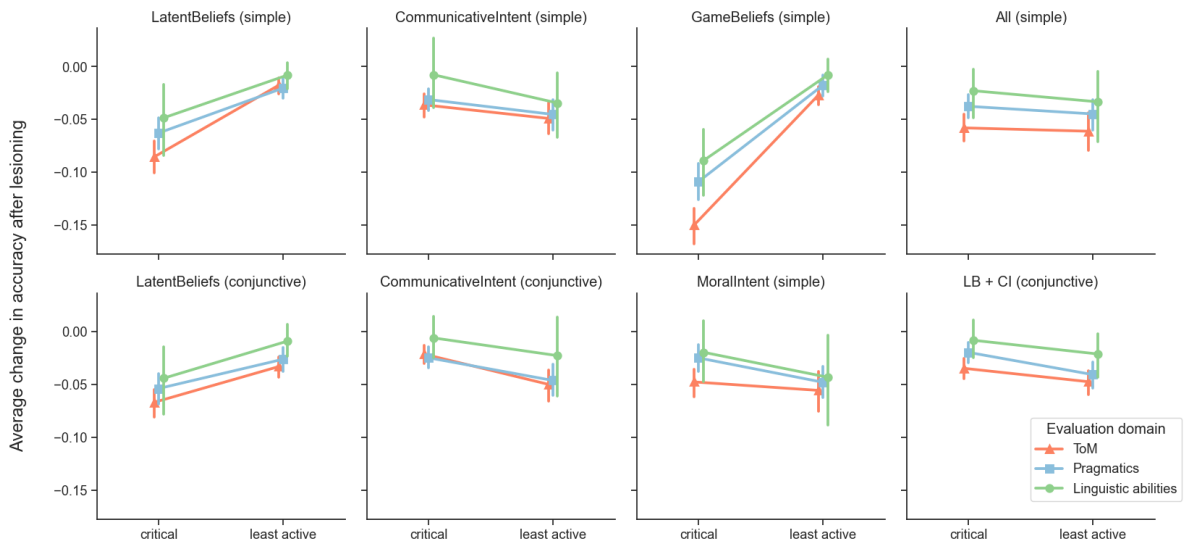


Figure 14: Average difference in accuracy, by localizer, across models.

reasoning mechanisms, and that these may share representations. In contrast, analogical reasoning seems to be less likely to be part of ToM and pragmatics tasks. Results for all 13 models for each localizer are shown in Figures 19–20 (aggregating the three reasoning datasets into the “General reasoning” domain).

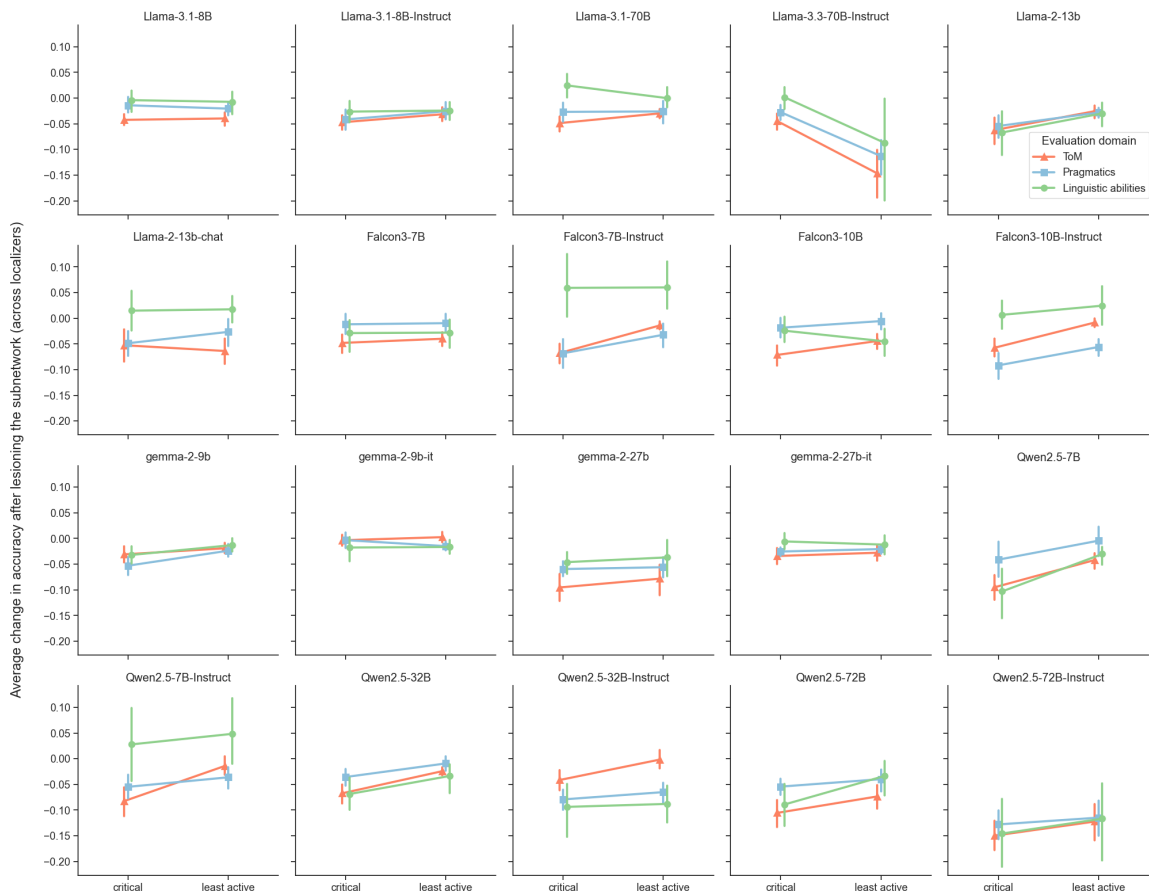


Figure 15: Average change in accuracy (y-axis) relative to the intact model when applying ablations of the target or least active networks (x-axis) across evaluation datasets and localizer suites for different evaluated models (facets). A stronger decrease in accuracy for the critical ablation on target domains (ToM and pragmatics, color) is better. 95% CIs indicate change in by-dataset accuracy across localizer suites.

LatentBeliefs (Saxe and Kanwisher, 2003)	
FalseBelief (trgt)	Story: Tom dropped his keys in the bowl by the door before going for a run. While he was out, Mia put the keys into the desk drawer to tidy up. Tom came back a bit later. Statement / question: Tom will first look for his keys in the (● bowl ● drawer)
Desire (trgt)	Story: Coach Ramirez booked the soccer field for Saturday practice. He told the team they'd be outside unless lightning forced them into the gym. A storm never came, and they kept the plan. Statement / question: The coach wanted to practice (● indoors ● outdoors)
FalsePhoto (ctrl)	Story: A bus schedule was printed last month. Yesterday the transit agency changed Route 9 to depart an hour later. All the old paper schedules are still circulating. Statement / question: On the printed schedule, Route 9 departs (● earlier ● later)
HumanDescr (ctrl)	Story: Nora never goes out in the rain without her favorite outfit. She wears a bright yellow raincoat and matte black rubber boots. Even her umbrella is speckled with tiny white dots. Statement / question: Nora's raincoat is (● yellow ● black)
MechInf (ctrl)	Story: She forgot a full glass bottle of soda in the freezer overnight. By morning, the freezer door was stuck and there were shards of glass around the bottle. The cap was still on, but the liquid had turned to ice and expanded. Statement / question: The bottle shattered because (● freezing water expanded ● a thief broke in)
NonhumDescr (ctrl)	Story: Penguins are birds that cannot fly but swim swiftly. They use their stiff wings like flippers to push through the water. A layer of fat keeps them warm in icy seas. Statement / question: In the water, penguins propel themselves mainly with their (● wings ● tails)
CommunicativeIntent (Bosco et al., 2017)	
Deceptive (trgt)	Story: Oliver has been invited to his manager's poetry reading. He finds the poems boring and overwrought, but he wants to make a good impression at work. After the reading, his manager asks what he thought. Oliver replies: "It was deeply moving, I loved every minute" Statement / question: Was this (● literal ● deceitful ● ironic ● meaningless?)
Ironic (trgt)	Story: Luca and Giulia are driving to a wedding. They hit a standstill on the ring road, horns blaring and heat shimmering off the tarmac. The GPS suddenly adds another hour to their arrival time even though the venue is close by. Giulia looks at the bouquet wilting on the dashboard. Luca sighs and says: "Great, we'll be early" Statement / question: Was this (● literal ● deceitful ● ironic ● meaningless?)
Literal (ctrl)	Story: Elena is making toast for breakfast but accidentally burns it black. Her partner Diego walks in and raises an eyebrow. Elena answers: "I'll make another one." Statement / question: Was this (● literal ● deceitful ● ironic ● meaningless?)
Meaningless (ctrl)	Story: Liam and Nora are doing the weekly shop at the supermarket. They've already picked up milk, bread, and fruit. Liam reaches for a carton and asks Nora if they still have eggs at home. Nora answers: "The station is two blocks away" Statement / question: Was this (● literal ● deceitful ● ironic ● meaningless?)
GameBeliefs* (Chang et al., 2023)	
GameBelief (trgt)	Story: Nora and Ben play an Ultimatum Game. Nora is the proposer and receives 20 Euros to divide between her and another player. She intends to offer half to Ben, but a glitch records her offer as 1 Euro. Ben sees 1 Euro on his screen, frowns, and rejects immediately. Statement / question: Ben evidently believes that (● Nora is trying to cheat him ● a glitch reduced the offer)
GameOutcome (ctrl)	Story: Nora and Ben play an Ultimatum Game. Nora is the proposer and receives 20 Euros to divide between her and another player. She intends to offer half to Ben, but a glitch records her offer as 1 Euro. Ben sees 1 Euro on his screen, frowns, and rejects immediately. Statement / question: Ben will now have (● 1 Euro ● 0 Euro)
MoralIntent* (Young et al., 2007)	
MoralIntent (trgt)	Story: Anna and her friend are hiking along a cliffside trail when they reach a narrow rock ledge overlooking a canyon. The ledge is unstable and will crumble if someone steps onto it, causing a deadly fall. Because Anna sees deep cracks running through the rock and read a posted warning about recent collapses, she believes the ledge is unsafe and stepping on it could be fatal. Anna urges her friend to step onto the ledge for a better view; the ledge collapses and her friend falls to her death. Statement / question: Urging her friend onto the ledge was (● permissible ● forbidden)
DecOutcome (ctrl)	Story: Anna and her friend are hiking along a cliffside trail when they reach a narrow rock ledge overlooking a canyon. The ledge is unstable and will crumble if someone steps onto it, causing a deadly fall. While trying to take a photo of the view, Anna steps onto the ledge; it collapses, and she falls to her death. Statement / question: Stepping onto the ledge was (● a good idea ● a bad idea)

Table 3: Examples of synthetic GPT-5 generated stimuli for all localizer suites. Dots indicate answer options. The references indicate the source of the original stimuli that were used to prompt GPT-5.

Dataset	beliefs	intentions	desires	emotions	knowledge	percepts	NLC
LatentBelief (Saxe and Kanwisher, 2003)							
• FalseBelief	✓	×	×	×	×	✓	×
• Desire	×	✓	✓	×	×	×	×
CommunicativeIntent (Bosco et al., 2017)							
• Deceptive	×	✓	×	×	×	×	✓
• Irony	×	✓	×	×	×	×	✓
MoralIntent (Chang et al., 2023)	×	✓	×	×	✓	×	×
GameBelief (Young et al., 2007)	✓	×	✓	✓	×	✓	×

Table 4: Classification of localizer suites according to the ATOMS framework (Beaudoin et al., 2020). NLC stands for “non-literal communication”.

Localizer	P1.1	P1.2	P2.1	P2.2	P3.1	P3.2
GB	0.59	0.49	0.58	0.50	0.36	0.22
(simple)	[0.46, 0.73]	[0.36, 0.61]	[0.46, 0.7]	[0.39, 0.63]	[0.06, 0.64]	[-0.09, 0.55]
MI	0.19	-0.11	0.16	-0.08	0.08	0.08
(simple)	[0.05, 0.32]	[-0.24, 0.02]	[0.04, 0.30]	[-0.20, 0.05]	[-0.22, 0.38]	[-0.24, 0.41]
LB + CI	0.14	-0.04	0.21	-0.10	0.04	0.17
(conj.)	[0.01, 0.28]	[-0.17, 0.09]	[0.08, 0.34]	[-0.22, 0.03]	[-0.26, 0.33]	[-0.14, 0.50]
All	0.22	-0.05	0.29	0.05	0.10	0.19
(simple)	[0.09, 0.36]	[-0.19, 0.08]	[0.17, 0.42]	[-0.07, 0.17]	[-0.19, 0.40]	[-0.14, 0.50]
CI	0.08	-0.12	0.16	-0.13	0.03	0.13
(conj.)	[-0.05, 0.22]	[-0.24, 0.02]	[0.04, 0.29]	[-0.26, -0.01]	[-0.27, 0.33]	[-0.19, 0.45]
CI	0.14	-0.05	0.23	-0.05	0.03	0.20
(simple)	[0.01, 0.28]	[-0.19, 0.08]	[0.10, 0.36]	[-0.18, 0.07]	[-0.26, 0.32]	[-0.12, 0.52]
LB	0.26	0.14	0.36	0.15	0.18	0.18
(conj.)	[0.13, 0.40]	[0.00, 0.27]	[0.23, 0.48]	[0.02, 0.28]	[-0.12, 0.47]	[-0.14, 0.50]
LB	0.34	0.27	0.39	0.22	0.20	0.19
(simple)	[0.20, 0.47]	[0.14, 0.40]	[0.26, 0.52]	[0.09, 0.34]	[-0.10, 0.50]	[-0.13, 0.51]

Table 5: Posterior mean estimates and [95% credible intervals] for assessing the main predictions separately by-localizer. When the credible interval excludes zero, the prediction is borne out. “conj.” stands for conjunctive localizers.



Figure 16: Average change in accuracy across evaluation datasets (y-axis) when different localizer datasets are used (columns) in Llama and Falcon models (rows) when ablating the critical vs. the least active subnetworks (x-axis).

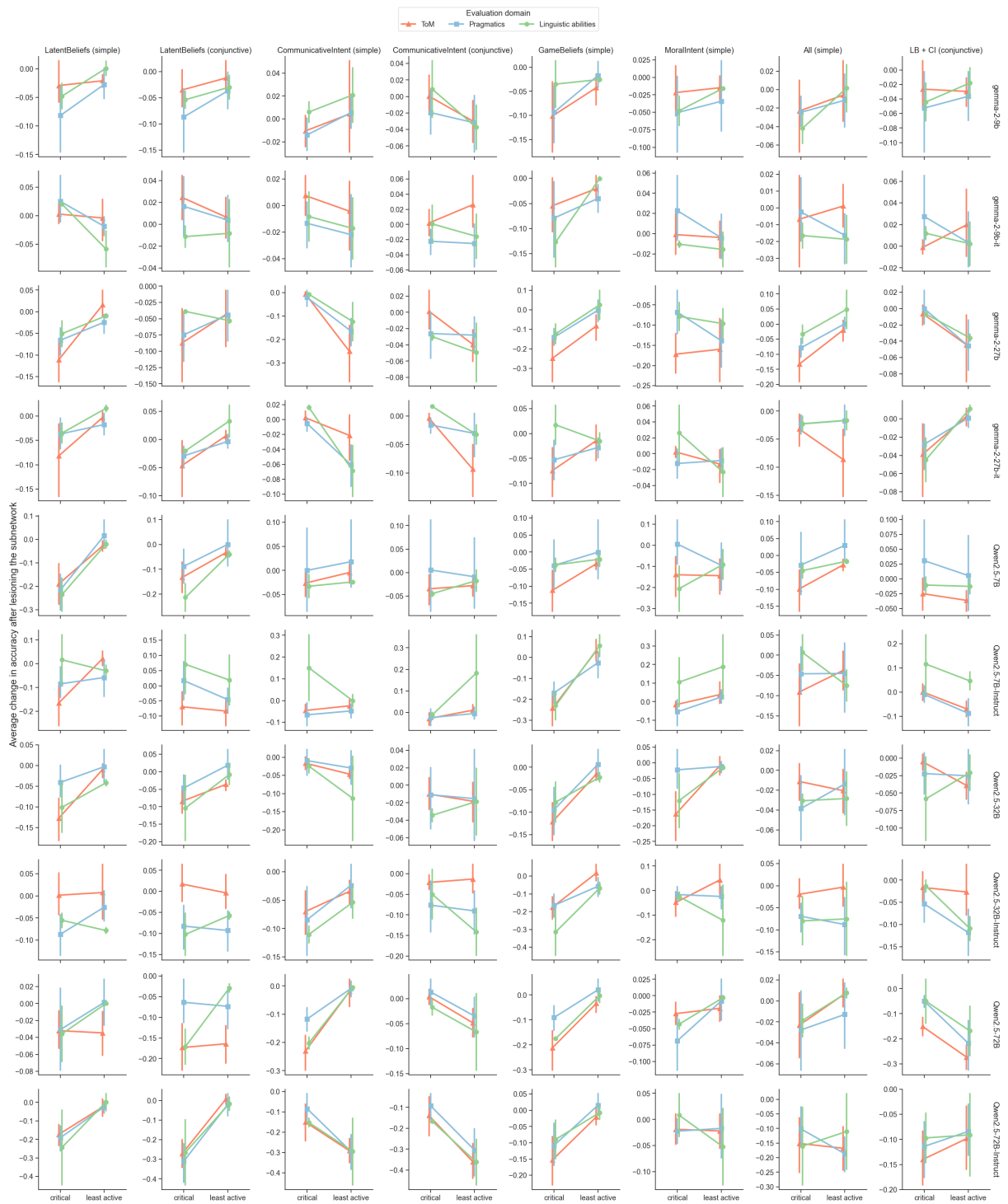


Figure 17: Average change in accuracy across evaluation datasets (y-axis) when different localizer datasets are used (columns) in Gemma and Qwen models (rows) when ablating the critical vs. the least active subnetworks (x-axis).

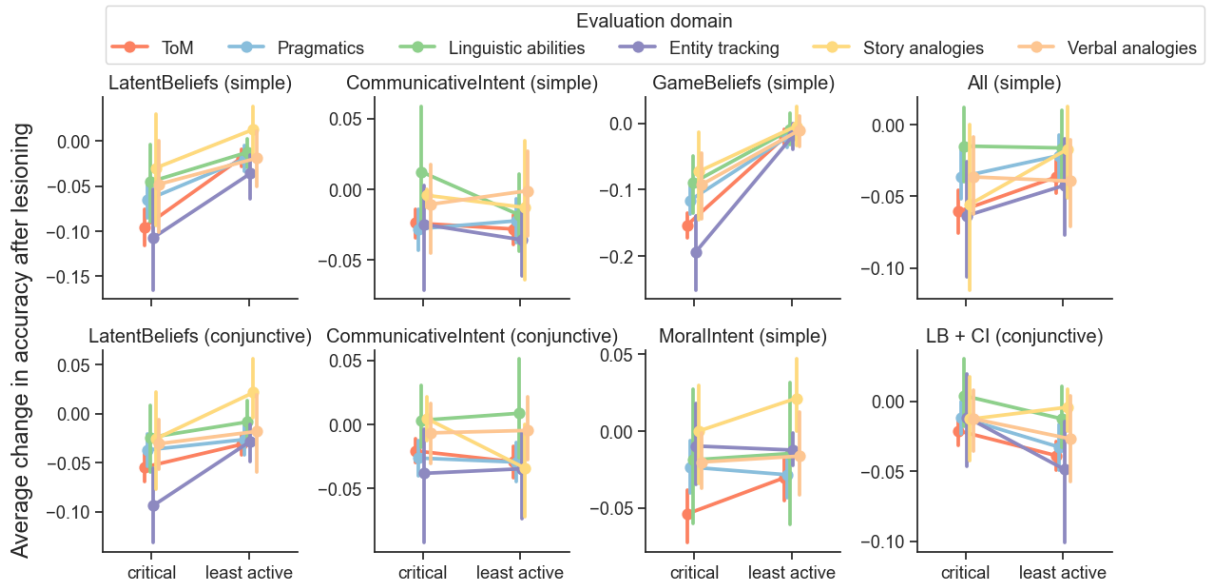


Figure 18: Average difference in accuracy under ablations of the localized subnetworks, by localizer, across 13 models. Additionally to the results reported in the main text, change in performance on general reasoning tasks (entity tracking Kim and Schuster, 2023; verbal and story analogies Webb et al., 2023) is shown.

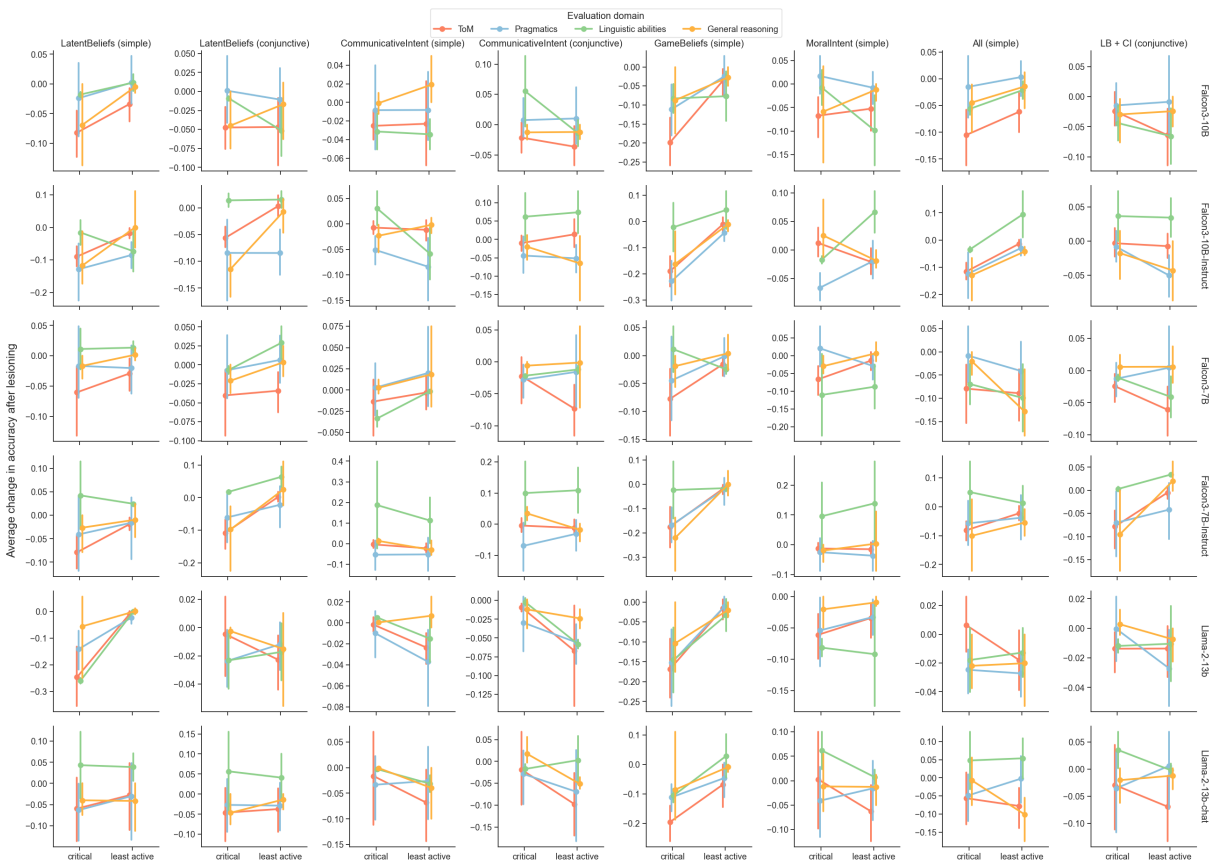


Figure 19: Average difference in accuracy under ablations of the localized subnetworks, for each localizer and selected six models. Additionally to the results reported in the main text, average change in performance on control general reasoning tasks is shown.

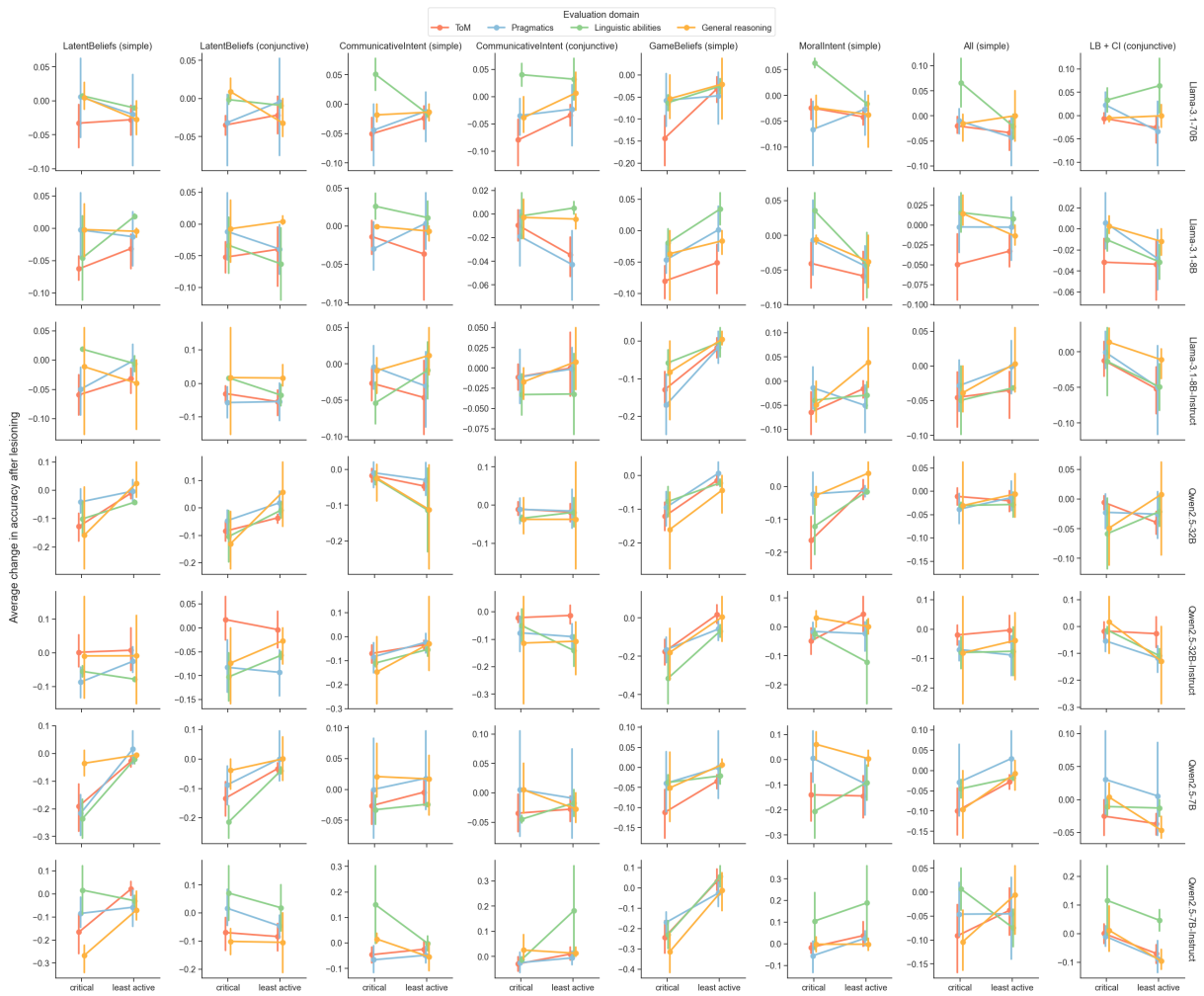


Figure 20: Average difference in accuracy under ablations of the localized subnetworks, for each localizer and additional selected seven models. Additionally to the results reported in the main text, average change in performance on control general reasoning tasks is shown.