Traces of Social Competence in Large Language Models

Tom Kouwenhoven*, **Michiel van der Meer***, **and Max van Duijn**
Leiden Institute of Advanced Computer Science
Leiden University, Netherlands
{t.kouwenhoven, m.t.van.der.meer, m.j.van.duijn}@liacs.leidenuniv.nl

Abstract

The False Belief Test (FBT) has been the main method for assessing Theory of Mind (ToM) and related socio-cognitive competencies. For Large Language Models (LLMs), the reliability and explanatory potential of this test have remained limited due to issues like data contamination, insufficient model details, and inconsistent controls. We address these issues by testing 17 open-weight models on a balanced set of 192 FBT variants (Trott et al., 2023) using Bayesian Logistic regression to identify how model size and post-training affect socio-cognitive competence. We find that scaling model size benefits performance, but not strictly. A cross-over effect reveals that explicating propositional attitudes (X *thinks*) fundamentally alters response patterns. Instruction tuning partially mitigates this effect, but further reasoning-oriented fine-tuning amplifies it. In a case study analysing social reasoning ability throughout OLMo 2 training, we show that this cross-over effect emerges during pre-training, suggesting that models acquire stereotypical response patterns tied to mental-state vocabulary that can outweigh other scenario semantics. Finally, vector steering allows us to isolate a *think* vector as the causal driver of observed FBT behaviour.

1 Introduction

> *"Maxi and his mother store chocolate in the blue cupboard in their kitchen. When Maxi leaves for the playground, his mother takes the chocolate and eats a piece. Then she puts it back into the green cupboard and leaves. When Maxi comes home, where will he look for the chocolate?"*

A version of this scenario was presented to children aged 3-9 by Wimmer and Perner in the early 1980s, after which none of the 3-4 year old, 57% of the 4-6 year old, and 86% of the 6-9 year old children in their sample pointed "correctly" to the blue cupboard (Wimmer and Perner, 1983). This finding was interpreted as reflecting children's emerging capacity to work with the difference between one's own and somebody else's relation to the same propositional content, a form of meta-representation (Pylyshyn, 1978). Over the ensuing decades, variants of this scenario were implemented in experiments with a wide variety of human and non-human populations, including LLMs recently. Across all this work, the test subjects' ability to appreciate that a character may hold a False Belief has been taken as an indication of their broader social-cognitive competence, often described as their capacity for ToM (Apperly, 2010).

The relationship between language and the representation of false beliefs remains debated for both humans and LLMs. Studies in child development have demonstrated correlational and causal links between the mastery of certain linguistic forms (i.a., negation, clausal complement syntax, verbs of cognition) and FBT performance, suggesting that language acquisition may scaffold meta-representational ability (Milligan et al., 2007). For LLMs, the debate centres on whether distributional patterns in their linguistic training data amount to "social-world models" that generalise beyond highly specific contexts, such as the FBT, or whether they only learn to solve such tests using superficial, context-specific heuristics. Over the past years, approaches for evaluating and enhancing socio-cognitive capacities in LLMs, including FBT, have proliferated (see Chen et al., 2025, and Section 2, for an overview).

However, common problems remain that, for LLMs, test and evaluation approaches vary in robustness. Oftentimes, a small number of LLMs are used, which complicates the determination of the generalisability of empirical findings (Trott, 2025). Little information is released about the majority of models, making it difficult to control for leakage of test data into the training data and to systematically map the effects of differences in model size,

*Equal contribution

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 742–759

architecture, pre-, mid- and post-training (Hu et al., 2025). In addition, difficulties in disentangling test formulations from underlying model representations have hampered progress on fundamental questions of mechanisms and generalisability. Addressing these concerns is the focus of this paper.

Leveraging the increasing availability of open-source and open-weight model families, we provide such a systematic comparison for a total of 17 model variants based on computed probabilities (Pimentel and Meister, 2024) of completions of 192 FBT scenario variants from Trott et al. (2023). We apply vector steering (Subramani et al., 2022; Rimsky et al., 2024) to develop a new approach for distinguishing between different drivers of observed FBT performance or failure. Our findings not only concern the social-cognitive abilities of LLMs but also contribute to long-standing debates in psycholinguistics and developmental psychology about the extent to which these abilities can be learned through linguistic bootstrapping.

2 Background & Related work

The concept of ToM, also known as mindreading, is classically defined as the capacity to attribute mental states to others and oneself, to explain and anticipate behaviour (Premack and Woodruff, 1978). ToM is held to be a precondition for many aspects of social and cultural living, including communication (Sperber, 2000), forming and maintaining social networks (Sutcliffe et al., 2014), and cultural learning (Tomasello, 2014). The first tests based on fictional characters with different knowledge states were described by Flavell et al. (1968), a format that was further developed by Dennett (1978) and Wimmer and Perner (1983), before the most renowned FBT format involving two dolls named Sally and Anne was introduced by Baron-Cohen et al. (1985). Variants of this test have been implemented in experiments with infants (Barone et al., 2019), adults (Meinhardt et al., 2011), non-human animals (Call and Tomasello, 2008; van der Vaart et al., 2012), and computer models (Arslan et al., 2017; Rabinowitz et al., 2018). There is consensus that the FBT captures an essential part of ToM competence. Test-retest reliability is reportedly high (Hughes et al., 2000) and correlations with real-world social abilities (e.g. detecting lies, understanding non-literal language, *faux pas* avoidance) are well-documented (Beaudoin et al., 2020; Apperly, 2010). However, there is also debate about the actual cognitive requirements for passing the FBT; for an overview and a rebuttal see Jacob (2020). Scott and Baillargeon (2017) argue that toddlers and even infants under the age of 2 can pass an FBT once reliance on complex language is removed, whereas others have argued that only language-based tests suffice to assess actual false belief understanding (Heyes, 2014).

After training on vast amounts of language data, LLMs impress on numerous tasks beyond natural language processing (Qin et al., 2025). The discovery of emergent abilities on cognitive and behavioural tasks has sparked numerous studies in "machine psychology" (Hagendorff et al., 2024), targeting, for example, the evolution of language (Kouwenhoven et al., 2025b,a) or strategic (Gandhi et al., 2023), emotional (Sabour et al., 2024), and moral (Oh and Demberg, 2025) reasoning. Much interest has gone out to the ToM capabilities of LLMs, following roughly three sorts of approaches: building test procedures and benchmarks from text-based paradigms in experimental psychology (e.g, Kosinski, 2024; Wu et al., 2023; Chen et al., 2024b), designing alternative (situated) tests for LLMs (Kim et al., 2023; Chan et al., 2024), and focusing on enhancing performance (Moghaddam and Honey, 2023; Jin et al., 2024). Large proprietary models in particular have been shown to perform on par with older children (van Duijn et al., 2023) and adults (Strachan et al., 2024). Trott et al. (2023) have used such a comparison to argue that part of what it takes to pass FBTs can be learned from exposure to large-scale language data, whereas other parts seem to rely on human qualities not possessed by LLMs. Concurrent work further strengthens this comparison by showing that, for a range of LLMs, distributional statistics are in principle *sufficient* to develop sensitivity to belief states in FBT reasoning (Trott et al., 2026). Approaches that apply interpretability techniques have been able to isolate special subnetworks in models that seem to have specialised for FBT reasoning (Wu et al., 2025).

Early implementations of the FBT for end-to-end memory network (MemN2N) (Grant et al., 2017; Nematzadeh et al., 2018) were already criticised for their sensitivity to problems with answer matching and data leakage that allowed superficial heuristics (Le et al., 2019). These problems persisted in LLM research. Ullman (2023) and Shapira et al. (2024) showed that small modifications in test formulations caused models to fail questions

they previously answered correctly, suggesting success was narrow and contingent on surface patterns. For a full overview of these issues, see Hu et al. (2025); for further theoretical reflection, see (Goldstein and Levinstein, 2024). Our work resonates with a broader movement calling for careful design and thorough analyses before drawing normative conclusions about the cognitive abilities of LLMs (Frank, 2023; Ivanova, 2025; Mitchell, 2026).

3 Methods

3.1 Tasks

Our analyses build on the FBT battery from (Trott et al., 2023), later incorporated into the EPITOME dataset (Jones et al., 2024). This is a suitable battery since (1) its scenarios are *not* present in the OLMo 2 and OLMo 3 training data, which we rule out through extensive investigation into data leakage (Appendix D), and (2) it is counterbalanced to avoid superficial task-solving based on training priors or surface-level heuristics (Liu et al., 2024).

The task contains 12 scenarios that conform to the original FBT structure (Wimmer and Perner, 1983), i.e., a main character places an object at a Start location and a second character moves the object to an End location. The experimental question then probes the test subject's (c.q. the LLM's) assessment of where the main character believes the object is. Crucially, for each scenario, there are 16 versions that differ in four dimensions (observation condition, knowledge cue, and two location mentions), resulting in 192 test variations:

Knowledge state – True Belief: the main character is present (and thus sees) when the second character is moving the object. False Belief: the main character is absent (and thus does not see) when the object is being moved.

Knowledge cue – Explicit: the propositional attitude of the main character is explicated in the experimental question (X *thinks* the book is in the . . .). Implicit: an action verb is used instead (e.g., X *goes* to get the book from the . . .).

Location mentions – Start and End locations are mentioned twice in each scenario, alternating first and most recent mentions (creating two axes).

An example scenario is shown in Figure 1. Trott et al. (2023) reported that humans correctly answered False Belief questions in 83% of the cases (which aligns with the original finding for the oldest group in Wimmer and Perner, 1983); we include their scores as a human baseline in our plots.

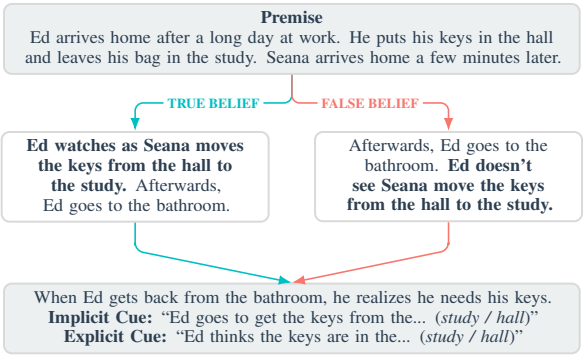

Figure 1: Example False Belief task from Trott et al. (2023). A detailed version is visible in Appendix B.

Our primary manipulation investigates the effect of changing the knowledge state and the knowledge cue. Location mentions serve as a control for the potential effect of recency or primacy biases (Liu et al., 2024; Mina et al., 2025).

3.2 Computing probabilities

We probe an LLM's assessment of the scenario by computing the probability of preset completions. Concretely, we compute the conditional probability of the Start and End locations, corresponding to the False and True Belief, given the cue context. Since all tested models use Beginning-of-Word tokenisers that may split the target words into variable-length subword units, we must ensure that the computed probability reflects the target word as a complete, whitespace-delimited unit rather than a prefix of a longer word. We correct for this following Pimentel and Meister (2024), by normalising subword token probabilities with a marginalisation factor (see Appendix A for details). The location word with the highest probability acts as the model's prediction.

3.3 Models

We test a suite of 17 different models from 6 families at varying levels of openness (see Appendix B for a full list and the detailed computational setup). Our subsequent analysis is split into two steps. First, in Section 4, we assess performance on the FBT across all models to enable LLM-generic analyses. Here, we distinguish between model variants based on *intended capability*, i.e., base, instruct, and reasoning. We do so since models differ in how they achieve said variants, using different data and training regimes. For example, Llama 3 instruct models were finetuned using `SFT` and `DPO`, while K2-V2 instruct only used `SFT`, and OLMo

2 and 3 underwent SFT, DPO, and RLVR. Second, in Section 5, we trace OLMo 2's acquisition of FBT competence across pre- and mid-training, and unravel how this is affected during post-training.

3.4 Analyses

Our analyses use Bayesian Regression Models as implemented in the brms package (Bürkner, 2021) and Bayesian two-sample t-tests as implemented in the BayesFactor package (Morey and Rouder, 2023) in R (R Core Team, 2025). We focus on correctly predicted answers (*correct*) for a True Belief or False Belief *knowledge state* that is prompted with a specific *knowledge cue*.

We fit a Bayesian Multilevel Logistic Regression model with a logit function:

$$\begin{aligned} correct \sim\ & model\_size * variant * \\ & knowledge\_state * knowledge\_cue \\ & + (1 \mid model\_family\_version) \end{aligned}$$

It predicts whether a given model size (in Billions), training variant (Base, Instruct, Reasoning), knowledge state (False Belief, True Belief), knowledge cue (Explicit, Implicit), and their interactions will correctly answer the question (binary). A random intercept for the model family and its version (e.g., llama_3.1, olmo_3) is included because these versions may differ in baseline accuracy. Each model is fitted using 4 chains, each with 4000 iterations and a warm-up of 2000. Variable coefficients indicate the direction and magnitude of an effect; we consider effects reliable (i.e., significant) when their 95% credible intervals (CI) *exclude* zero, indicating that at least 95% of the posterior probability mass falls on one side of zero.

4 Results

Despite substantial variance across model families and versions ($\beta = 0.34$), likely reflecting differences in architecture, training data, and optimisation techniques, we still observe LLM-generic effects that we discuss now.

4.1 Does model size predict ToM reasoning?

It is generally claimed that scaling model size and pre-training data positively correlate with downstream performance and alignment with humans (e.g., Hoffmann et al., 2022; Ren et al., 2025). Here, we address whether these carry over to ToM reasoning as suggested in prior work (Kosinski, 2024).

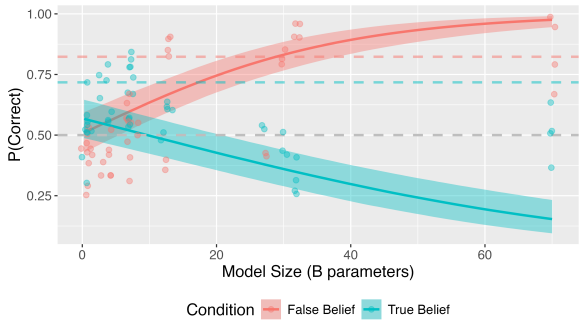

Figure 2: The effect of model size on the probability of being correct given a knowledge condition *without* any other interaction effects. Scaling positively influences False Belief performance but not True Belief. Dashed coloured lines indicate human performance.

We argue that to assess first-order ToM, performance across both False and True Belief scenarios should be taken into account, since flexibility to solve the task irrespective of the Knowledge State warrants a much higher degree of generalisability than considering False Belief scores alone.

The Bayesian model confirms that scaling up model size improves the log-odds of correct responses. Specifically, *without* accounting for any other effects, each additional unit (Billions) of model size increases log-odds performance by 0.05 ($\beta = .05, CI = [.04, .07]$). This aligns squarely with findings reported by Trott et al. (2026). Interestingly, however, this effect is not uniform across knowledge states, revealing a stark difference between False and True Belief tasks (Figure 2). Scaling strongly benefits False Belief, though it *harms* performance in the True Belief cases. While the classical developmental pattern is the reverse (false beliefs being harder than true beliefs in young children; Wellman et al. 2001), like models, adults appear to find the True Belief scenarios somewhat more difficult than the False Belief ones (Trott et al., 2023). This asymmetry is far more pronounced in models than in humans. Although Trott et al. (2026) show that larger models are better predictors of human responses, this asymmetry may explain why even these larger models cannot fully account for the human data. Returning to the question of whether model size predicts ToM reasoning, we observe that increasing model size does not have a strictly positive effect when both True Belief and False Belief scenarios are taken into account.

4.2 Explicating propositional attitudes

Earlier work reported that for GPT-3, FBT performance improved when propositional attitudes were explicated (e.g., "John thinks the wine is in the…") as opposed to when they were left implicit (e.g., "John goes to get the wine from the…"; Trott et al., 2023). Our results corroborate this effect, expand it to a broader range of models, and further specify its occurrence. Across all model sizes, implicit cues are consistently harder than explicit cues: *without* taking other effects into account, the log-odds of a correct response decrease by 1.11 units for implicit versus explicit cues ($\beta = -1.11, CI = [-1.40, -0.84]$). In our sample, this effect is equally strong across model sizes ($\beta = -0.01, CI = [-0.03, 0.00]$). However, Trott et al. (2026) use *additional* models and find that this effect even grows with parameter count.

Informed by the effects of scaling and implicit cues, we now aim to unravel why these cause the stark difference between True and False Belief scenarios. The answer lies in a striking crossover interaction between knowledge state and cue ($\beta = 2.07, CI = [1.70, 2.46]$, Figure 3), meaning that implicit cues impair False Belief performance but facilitate True Belief performance. Vice versa, explicit cues facilitate False Belief but hurt True Belief performance. This suggests that explicating propositional attitudes (X *thinks*) fundamentally alters response patterns and may explain why solving TB fails to benefit from scaling. Since the performance gap between FB and TB grows with model size, it is likely driven by a pattern learned from the data that is amplified as the model scales. Though it is tempting to attribute this solely to surface-level heuristics or pattern-matching (e.g., Ullman, 2023; Shapira et al., 2024), our results suggest a more nuanced mechanism. The crossover pattern indicates that models are sensitive to two interacting factors: (1) learned associations with stereotypical False Belief scenarios, and (2) the presence of explicit propositional attitudes themselves.

Critically, this does not necessarily imply an absence of genuine ToM reasoning. Rather, it suggests that explicit propositional attitudes and learned scenario patterns interact with such reasoning in ways that can either facilitate or interfere with the process of arriving at a correct prediction. Interestingly, Trott et al. (2026) report that the use of non-factive verbs affects human FBT performance *in similar ways* to those of LLMs. They

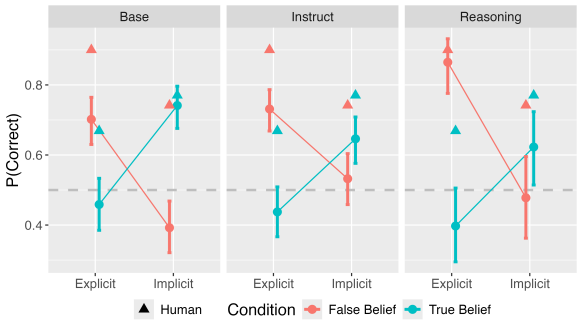

Figure 3: The probability of predicting the location correctly for model variant, knowledge state, and cue. Triangles indicate the average human performance.

ascribe this effect to *anti-presupposition*, i.e., the idea that 'Alex thinks X' weakly implies 'NOT X' (Chemla, 2008). This is in line with White et al. (2018)'s analysis of *think* as a representational non-factive propositional attitude verb. Compare the following two sentences:

(a) Alex thinks that Bo went to the store.
(b) Alex knows that Bo went to the store.

Sentence (a) is pragmatically fitting when the speaker does not know whether Bo went to the store or not, or knows for certain that Bo did not go. Yet it does *not* fit a context where the speaker knows for certain that Bo went to the store; here (b) would be more appropriate (c.f. also Stalnaker, 1973). Because of this distinction, *think* can be expected to be frequently used to mark epistemic divergence rather than congruency within communicative settings. Applied to LLMs and the narrow context of the FBT, when *think* appears, it may activate a learned pattern of opposing what is actually known to be the case. The presence of the non-factive propositional attitude verb *think* may thus bias the LLM towards predicting a *contrast* between the agent's belief and the object's location ('NOT X'). This benefits test performance in the False Belief condition but impairs it in True Belief scenarios. We further investigate this account in Section 5.

4.3 Does post-training act as a cooperative pressure?

Existing work draws a parallel between post-training and the cooperative pressures that shape human communication (van Duijn et al., 2023). For humans, communication is fundamentally cooperative and relies on the ability and willingness to engage in mental coordination (e.g. Grice, 1975; Ver-

hagen, 2015). This willingness to engage in cooperation to achieve successful communicative interactions is constantly rewarded (Tomasello, 2008), while failing to do so results in punishment in the form of social exclusion (David-Barrett and Dunbar, 2016). Post-training essentially induces analogous cooperative principles, rewarding helpful responses and penalising failures to coordinate with the user's intent. van Duijn et al. (2023) therefore hypothesised that instruction-tuned LLMs, like humans, bank on the capacity to coordinate with an interaction partner's perspective, benefiting ToM tasks that depend on precisely this capacity.

Here, we empirically test whether this parallel holds by examining the effects of different model variants. If the parallel exists, we hypothesise that post-trained model variants that follow instructions will show this most notably, as they are rewarded for interactive behaviour. While the analogy is less clear for models that have been post-trained for reasoning, one might expect these models to attend more carefully to prompt details ("...think step-by-step ..."). However, recent research has shown revealed that reasoning remains fragile to surface-level input variations (Mirzadeh et al., 2025), suggesting that reasoning may also amplify heuristic (i.a., pattern matching, pick the last location) responses.

Since post-training is done sequentially (i.e., first pre-training `Base`, then `SFT`, then `DPO`, etc.), we use checkpoints that reflect this structure (i.e., Base, Instruct, and Reasoning). Importantly, the effects reported should be interpreted as the cumulative effect of additional stages after pre-training, *not* the incremental effect of each training stage individually. Moreover, reported values represent performance at the mean model size, averaging over models in that condition and variant.

Given the interaction between knowledge states and propositional attitudes, we now examine how this differs after post-training. We find that post-training affects FBT accuracy. Both Instruct and Reasoning variants improve relative to Base models (Instruct: $\beta = 0.33$, CI= $[0.07, 0.57]$; Reasoning: $\beta = 0.74$, CI= $[0.10, 1.39]$), aligning with work showing that post-trained models (i.a., OLMo 2) are among the most human-aligned (Studdiford et al., 2025). However, these overall gains mask important differences in how training variants interact with task characteristics. Figure 3 also reveals that the crossover pattern discussed above is not uniform across training variants. In Base models, it is most pronounced as implicit cues strongly impair False Belief while substantially facilitating True Belief. Instruct training partially rescues the implicit False Belief deficit ($\beta = 0.37, CI = [0.02, 0.72]$), attenuating the crossover. Training to induce Reasoning, by contrast, worsens the implicit False Belief impairment ($\beta = -1.09, CI = [-1.93, -0.24]$), while also reducing True Belief performance ($\beta = -0.93, CI = [-1.73, -0.11]$), amplifying sensitivity to knowledge cue type dramatically. This suggests that reasoning training specialises models for explicit False Belief reasoning at the cost of broader ToM robustness.

Thus, the answer to whether post-training acts as a cooperative mechanism (van Duijn et al., 2023) is nuanced: while improving FBT performance overall, the gains are modulated by stronger interaction effects, and only robust performance across conditions would indicate true ToM reasoning abilities.

5 Case study: Traces of social intelligence in OLMo 2

While previous analyses focused on a large sample of models, we now narrow the analyses down to OLMo 2 models only. This allows us to (1) limit data contamination risks, as we are certain that the stimuli are *not* present in any of the training data (Appendix D), and (2) deepen our investigation into the influence of fine-tuning techniques (e.g., `SFT`, `DPO`, `RLVR`) on the observed cross-over effect, since we know exactly how OLMo is trained.

5.1 Pre-training dynamics

We follow Mahowald et al. (2024) and ask whether acquiring low-level formal language competencies naturally extends to functional linguistic capabilities. In line with previous analyses revealing early and sudden syntax acquisition (Chen et al., 2024a), OLMo 2 acquires formal linguistic capabilities, as measured by achieving $> 80\%$ accuracy on the BLIMP benchmark (Warstadt et al., 2020), after observing only 0.05% of the total number of training tokens. In contrast, it takes about 25 times longer (12.5% of the data observed) for the model to achieve above-random performance on the False Belief task. Moreover, BLIMP performance remains stable throughout training, whereas ToM performance fluctuates, further hinting at distinct underlying drivers (Hanna et al., 2026, more details in Appendix C.2). Though language acquisition may scaffold ToM in children (Milligan et al.,

2007; De Villiers and de Villiers, 2014), we cannot conclude that merely exposing LLMs to additional tokens will automatically lead to the emergence of ToM. However, our findings suggest that formal linguistic capabilities may be required for functional understanding and that post-training objectives can further elicit them. This aligns with recent evidence that LLM-brain alignment tracks formal competence earlier and more strongly than functional competence (AlKhamissi et al., 2025).

This leaves us with the question of when OLMo 2's FBT capability will evolve, as investigated through extensive analysis of its learning dynamics. The training of OLMo 2 consists of a pre-training phase, a mid-training phase that uses model merging, and a fine-tuning phase using `SFT`, `DPO`, and `RLVR` (OLMo et al., 2025). For the pre-training stage of each model size, we sample 50 checkpoints evenly spaced by token count, after confirming that our sampled checkpoints yield a representative learning curve of an exhaustive sweep over all checkpoints for the 7B model (Appendix C). Similarly, we established that mid-training, intended to embed models with specialised knowledge through model merging, does not yield meaningful changes, and we omit this stage from further analyses.

Moving beyond the effect of knowledge cues, we now aggregate across them to assess robustness to knowledge state. This means a response is considered correct only if both the implicit and explicit cue conditions are answered correctly for a given scenario and location ("strictly" correct). Figure 4 reveals the learning traces for all OLMo 2 model sizes during pre-training. Consistently, True and False Belief performance differ significantly and show a cross-pattern that evolves during training. For small models, False Belief scenarios are more difficult than True Belief scenarios, while larger models show the opposite effect. This indicates that models are sensitive to the main character's knowledge state but in different respects. Following the call for robust human-centred evaluation (Ivanova, 2025; Mitchell, 2026), we argue once again that attributing ToM to LLMs is fair only if the knowledge state has no impact on the evaluation.

To unravel why we observe such stark differences in performance across different knowledge states, we analyse them by knowledge cue. Bayesian T-tests testing whether the mean accuracy differs across knowledge cues at the last checkpoint for each model confirm our earlier finding (Section 4.2) that explicating propositional attitudes hurts False Belief reasoning, whilst implicit reasoning improves True Belief tasks (Table 1). This shows tentative similarities to the finding that LLM representations of explicit prompts better correlate with brain activations than those of nonsensical, noisy prompts (Ren et al., 2025). Importantly, we want to stress that the collapse is not due to an apparent primacy or recency bias of OLMo 2 since location mentions are balanced. Our observation is largely consistent across model sizes and again confirms that scale helps with False Belief but hurts True Belief, except for the 1B model. Yet, given its stable low performance on the False Belief task, it may simply not be up to the task at all and thus not affected by changing the knowledge cue.

Returning to the question of how OLMo 2's ToM ability evolves, we observe that strict False and True Belief reasoning diverge during pre-training. As such, pre-training leads to the earlier-observed interaction between knowledge states and cues, driven by OLMo 2's ToM ability, which is not robust to adversarial cases involving explicit True Belief or implicit False Belief. While OLMo 2 does not encounter the exact test scenarios during

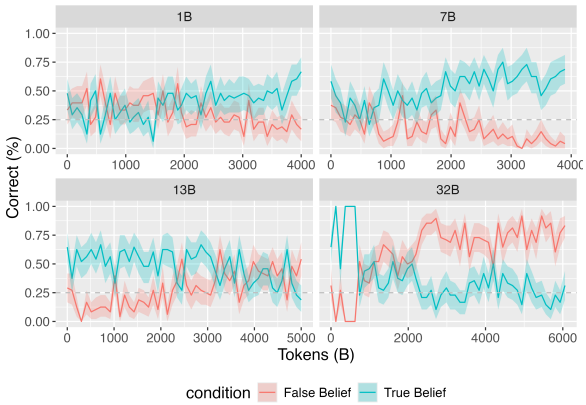

Figure 4: Strict performance during pre-training for different model sizes. Shaded areas indicate 95% CI.

		1B	7B	13B	32B
FB	Imp	.250	.042	.542	.833
	Exp	.292	.562**	.938**	.979*
TB	Imp	.771	1.00	.896	.750
	Exp	.750	.688**	.188**	.354**

Table 1: Mean accuracy at final pre-training checkpoint. Significance markers indicate a moderate ($*BF_{10} > 3$) or strong ($**BF_{10} > 30$) difference between paired Explicit (Exp) and Implicit (Imp) questions.

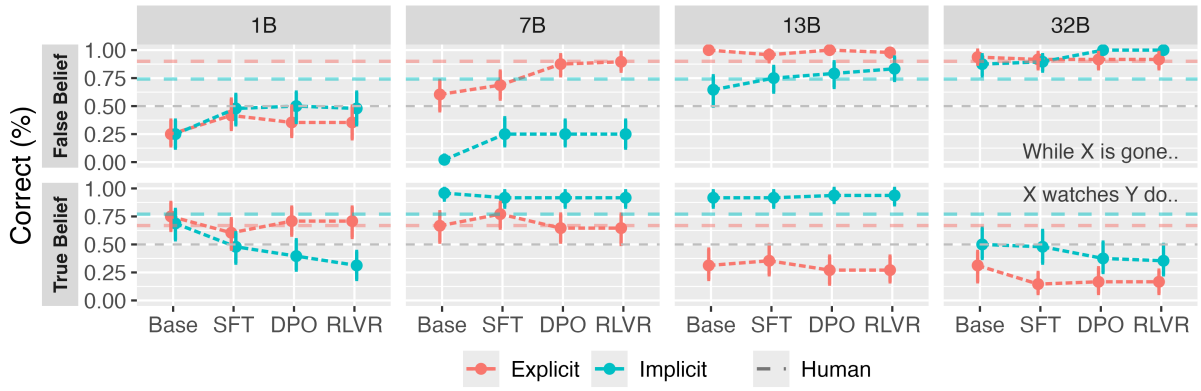

Figure 5: The percentage of correct answers in different base or post-training phases for differently sized OLMo 2 models. Coloured dashed lines indicate human performance, and the bars indicate 95% confidence intervals.

pre-training, our scenarios follow a format very similar to the canonical Sally-Anne format (Baron-Cohen et al., 1985). As argued in Section 4.2, explicating propositional attitude verbs also drives model responses. The combination of both implies that OLMo 2 may acquire stereotypical response patterns tied to mental-state vocabulary that can outweigh other scenario semantics. Pre-training language models on carefully curated datasets, excluding canonical FBT formats, would be required to further crystallise this hypothesis.

5.2 Post-training dynamics

Earlier LLM-generic analyses showed that post-training has a nuanced effect on the cross-over pattern between the knowledge state and the implicit/explicit cues (Section 4.3). The open-source nature of OLMo 2 enables us to investigate which post-training methods drive the cross-over effect.

Again, we observe a dichotomy between True and False Belief accuracy across model sizes (Figure 5). For False Belief tasks, smaller models benefit somewhat from post-training, especially using `SFT` and `DPO`, where we observe the smallest effect on the largest model. Similar findings have been described in a similarity judgement task (Studdiford et al., 2025) and in investigations towards generating diverse stories (Peeperkorn et al., 2025). For True Belief tasks, however, post-training has little effect, with two notable exceptions. First, post-training has a pronounced negative effect on the True Belief performance of the 32B model. Even though our exact task is not present in the training data, this model may be so large that it has learned the general FBT structure and is doing pattern matching. Using the pattern as a heuristic could also explain why post-training does not add, and sometimes hurts; if the heuristic is too strong, additional, deviating signals are ignored (Ullman, 2023; Shapira et al., 2024). Second, the 1B model under the implicit cue also shows a considerable drop in True Belief accuracy after post-training. Since this model has not exactly acquired what it takes to solve the FBT (i.e., it performs below chance for False Belief scenarios), further 'pushing' it with additional cooperative signals likely only leads it to form misconceptions about knowledge states. Thus, for OLMo 2, False Belief Scenarios seem to benefit from `SFT` and `DPO` post-training. While for True Belief scenarios, post-training hurts or has no effect.

6 Steering with *think* vectors

Motivated by the strong influence of knowledge cues, we apply model steering (Subramani et al., 2022; Rimsky et al., 2024; Siddique et al., 2025) to test whether OLMo 2 7B Instruct, selected for its stark explicit–implicit gap (Figure 5), encodes a representational direction that can causally influence predictions. Model steering (intuitively) captures a model's contrast between two sets of opposing prompts by embedding them and computing their difference. The resulting vector can be used to steer future predictions. Specifically, we extract steering vectors by computing the difference in last token hidden-state activations between paired explicit and implicit scenarios at layers 8–16, as this region proved most susceptible to steering in preliminary tests and in previous work (Subramani et al., 2022). Since pairs differ only in whether or not propositional attitudes are explicated, these vectors isolate the representational contrast, or di-

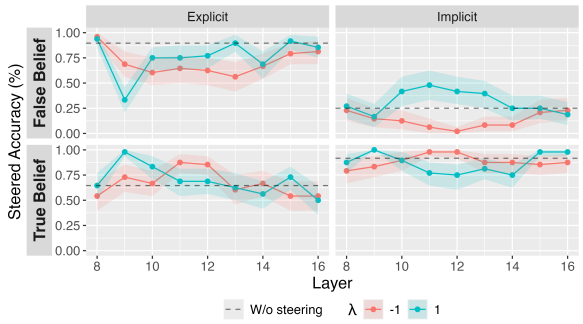

Figure 6: Model steering results. Dotted lines indicate *unsteered* accuracy and coloured lines indicate positively (green) or negatively (red) *steered* performance.

rection, between knowledge states—hence *think* vector. We inject these scaled (λ) vectors during inference at intermediate layers: $\lambda = 1$ adds the think vector to scenarios, supposedly pushing the model toward explicit-like behaviour, while $\lambda = -1$ subtracts it. To prevent information leakage, we use a leave-one-out procedure to ensure no scenario information is encoded in the think vector. This means that, for each scenario, we construct a think vector using the remaining 11 scenarios. Steered accuracy is compared against the unsteered baseline in Figure 6.

Model steering appears most effective at layers 10 to 12, to which we confine our discussion here. Adding 'explicitness' to *implicit* False Belief increases accuracy, while subtracting it decreases accuracy. Subtracting 'explicitness' from *explicit* False Belief decreases performance. Addition also hurts, but since base performance is at the ceiling and these scenarios are already explicit, it introduces redundancy and may therefore be disruptive. Our analyses showed that, for explicit True Belief cases, performance was lower (Section 5.2). In line with those findings, adding a think vector hurts for implicit True Belief cases, while subtracting it here boosts performance. Adding the think vector in explicit True Belief cases has little effect, whereas subtracting it leads to higher performance, again in line with expectations. These effects in the expected directions corroborate our suggestion that answering patterns can be rather strongly influenced by information encoded in the propositional attitude verb *think*, most likely contingent on its use in contexts of epistemic divergence in the training data.

7 Conclusion

Given ToM's central role in social interaction, we systematically tested the performance of 17 LLMs on the False Belief Test. Our analyses amount to the main finding that FBT performance of LLMs in our sample shows nuanced associations with scale and post-training, but is not robust to False vs. True and explicit vs. implicit variations. It appears that the interaction between learned scenario patterns and information contained in the verb *think* interferes with genuine social reasoning, sometimes in productive and sometimes in detrimental ways. Our findings inform discussions of the nature of social intelligence in LLMs and humans, and highlight the need for careful evaluation on high-quality tests.

Limitations

We highlight three noteworthy limitations of our work. First, we focus solely on ToM, while the ability to attribute mental states to others is clearly only a subset of socio-cognitive competencies. Social cognition involves a suite of mechanisms and strategies, ranging from low-level, implicit processes that rely on situated and embodied forms of cognition to reliance on norms, conventions, and common ground. All of these mechanisms amount to social competence and are important in everyday human-LLM interactions. Furthermore, the FBT is only a limited subset of ToM that, in our case, involves first-order mindreading. Beyond investigating higher-order cases, to speak of social competence in LLMs, it is pertinent that models perform robustly across a range of tests (e.g., recursive mindreading, strange stories, imposing memory; Jones et al., 2024).

Second, our finding that explicating propositional attitudes interacts with knowledge states currently relies on two verbs. These analyses should be extended and corroborated by adding different propositional attitude verbs (e.g., *believe*, *assume*) as test variants. Circuit analyses as in Wu et al. (2025) may further reveal whether propositional attitude verbs are encoded differently from non-propositional attitudes.

Lastly, recent work has revealed that LLMs may recognise when vector steering is applied (Fornasiere et al., 2026), attributed to introspection by some (Lindsey, 2026; Pearson-Vogel et al., 2026). As of yet, it is unclear whether and how this alters LLM inference. To rule out whether our think

vectors influence predictions, further stress testing with attention checks or evaluating factual questions in our scenarios can be conducted in additional analyses. We leave these analyses to future work.

References

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. 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.

Ian Apperly. 2010. *Mindreaders: the Cognitive Basis of "Theory of Mind"*. Psychology Press.

Burcu Arslan, Niels A. Taatgen, and Rineke Verbrugge. 2017. Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in Psychology*, 8.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Pamela Barone, Guido Corradi, and Antoni Gomila. 2019. Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, 57:101350.

Cindy Beaudoin, Elizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10.

Paul-Christian Bürkner. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

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

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

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

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024a. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. *Preprint*, arXiv:2505.00026.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Common Crawl Foundation. 2026. Common crawl. `https://commoncrawl.org/`. Accessed: 2026-02-03.

Tamas David-Barrett and Robin I. M. Dunbar. 2016. Language as a coordination tool evolves slowly. *R. Soc. open sci.*, 3:160259.

Jill G De Villiers and Peter A de Villiers. 2014. The role of language in theory of mind development. *Topics in Language Disorders*, 34(4):313–328.

Daniel C. Dennett. 1978. Beliefs about beliefs. *Behavioral and Brain Sciences*, 1(4):568–570.

J. H. Flavell, P. T. Botkin, C. L. Fry, J. W. Wright, and P. E. Jarvis. 1968. *The Development of Role-Taking and Communication Skills in Children*. Wiley, New York.

Damiano Fornasiere, Mirko Bronzi, Spencer Kitts, Alessandro Palmas, Yoshua Bengio, and Oliver Richardson. 2026. Language models recognize dropout and gaussian noise applied to their activations. *Preprint*, arXiv:2604.17465.

Michael C. Frank. 2023. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452.

Kanishk Gandhi, Dorsa Sadigh, and Noah Goodman. 2023. Strategic reasoning with language models. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Simon Goldstein and Benjamin A. Levinstein. 2024. Does ChatGPT have a mind? *arXiv preprint arXiv:2407.11015*.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and semantics. Vol. 3: Speech acts*, pages 41–58. Academic Press, New York.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2024. Machine psychology. *Preprint*, arXiv:2303.13988.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2026. Are formal and functional linguistic mechanisms dissociated in language models? *Computational Linguistics*, pages 1–41.

Cecilia Heyes. 2014. False belief in infancy: a fresh look. *Developmental Science*, 17(5):647–659.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

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

Claire Hughes, Anna Adlam, Francesca Happé, Jan Jackson, Alan Taylor, and Avshalom Caspi. 2000. Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry*, 41(4):483–490.

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

Pierre Jacob. 2020. What do false-belief tests show? *Review of Philosophy and Psychology*, 11(1):1–20.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. 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.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

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

Tom Kouwenhoven, Max Peeperkorn, Roy de Kleijn, and Tessa Verhoef. 2025a. Shaping shared languages: Human and large language models' inductive biases in emergent communication. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10298–10306. International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI.

Tom Kouwenhoven, Max Peeperkorn, and Tessa Verhoef. 2025b. Searching for structure: Investigating emergent communication with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, UAE. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

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.

Jack Lindsey. 2026. Emergent introspective awareness in large language models. *Preprint*, arXiv:2601.01828.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

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.

Jörg Meinhardt, Beate Sodian, Claudia Thoermer, Katrin Döhnel, and Monika Sommer. 2011. True- and false-belief reasoning in children and adults: An event-related potential study of theory of mind. *Developmental Cognitive Neuroscience*, 1(1):67–76.

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. 2007. Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-Belief Understanding. *Child Development*, 78(2):622–646.

Mario Mina, Valle Ruiz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor Gonzalez-Agirre. 2025. Cognitive biases, task complexity, and result interpretability in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE. Association for Computational Linguistics.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Melanie Mitchell. 2026. On evaluating cognitive capabilities in machines (and other "alien" intelligences). NeurIPS 2025 keynote.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Richard D. Morey and Jeffrey N. Rouder. 2023. *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.6.

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

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2025. Training on the benchmark is not all you need. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Soyoung Oh and Vera Demberg. 2025. Robustness of large language models in moral judgements. *Royal Society Open Science*, 12(4):241229.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. 2026. Latent introspection: Models can detect prior concept injections. *Preprint*, arXiv:2602.20031.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2025. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models. *Preprint*, arXiv:2507.20956.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

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

Zenon W. Pylyshyn. 1978. When is attribution of beliefs justified? [p&w]. *Behavioral and Brain Sciences*, 1(4):592–593.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2025. Large language models meet nlp: A survey. *Preprint*, arXiv:2405.12819.

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

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001, Abu Dhabi, UAE. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

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.

Rose M. Scott and Renée Baillargeon. 2017. Early false-belief understanding. *Trends in Cognitive Sciences*, 21(4):237–249.

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

Zara Siddique, Irtaza Khalid, Liam D. Turner, and Luis Espinosa-Anke. 2025. Shifting perspectives: Steering vectors for robust bias mitigation in llms. *Preprint*, arXiv:2503.05371.

Dan Sperber, editor. 2000. *Metarepresentations: A Multidisciplinary Perspective*. Oxford University Press USA, New York, US.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Zach Studdiford, Timothy T. Rogers, Kushin Mukherjee, and Siddharth Suresh. 2025. Uncovering the computational ingredients of human-like representations in llms. *Preprint*, arXiv:2510.01030.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Alistair Sutcliffe, R.I.M. Dunbar, Jens Binder, and Holly Arrow. 2014. Relationships and the social brain: Integrating psychological and evolutionary perspectives. In *Lucy to Language: The Benchmark Papers*. Oxford University Press.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

K2 Team, Zhengzhong Liu, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Seungwook Han, Bowen Tan, Gurpreet Gosal, Xudong Han, Varad Pimpalkhute, Shibo Hao, and 20 others. 2026. K2-v2: A 360-open, reasoning-enhanced llm. *Preprint*, arXiv:2512.06201.

Michael Tomasello. 2008. *Origins of human communication*. MIT Press, Cambridge, Mass.; London.

Michael Tomasello. 2014. The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.

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

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Sean Trott, Samuel Taylor, Cameron Jones, James A. Michaelov, and Pamela D. Rivière. 2026. Language statistics and false belief reasoning: Evidence from 41 open-weight lms. *Preprint*, arXiv:2602.16085.

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

Elske van der Vaart, Rineke Verbrugge, and Charlotte K. Hemelrijk. 2012. Corvid re-caching without 'theory of mind': A model. *PLoS ONE*, 7(3):e32904.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Arie Verhagen. 2015. *Grammar and cooperative communication*, pages 232–252. De Gruyter Mouton, Berlin, München, Boston.

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.

Henry M. Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684.

Aaron S. White, Valentine Hacquard, and Jeffrey Lidz. 2018. Semantic information and the syntax of propositional attitude verbs. *Cognitive Science*, 42(2):416–456.

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.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Computing probabilities

Since all tested models use Beginning-of-Word (BOW) tokenisers that might split the target words into variable-length subword units, we need to ensure that the computed probability reflects the target word as a complete, whitespace-delimited unit rather than a prefix of a longer word. We correct for this following Pimentel and Meister (2024), as shown in Eqn 1.

$$p(w \mid \mathbf{w}_{<t}) = p(\mathbf{s}^w \mid \mathbf{s}^{\mathbf{w}^{<t}}) \frac{\sum_{\{s \in \overline{\mathcal{S}}_{\text{bow}}\}} p(s \mid \mathbf{s}^{\mathbf{w}^{<t}} \circ \mathbf{s}^w)}{\sum_{\{s \in \overline{\mathcal{S}}_{\text{bow}}\}} p(s \mid \mathbf{s}^{\mathbf{w}^{<t}})} \quad (1)$$

We compute the probability of word w being generated following its preceding context $\mathbf{w}_{<t}$ by multiplying the probability of generating its constituent subword tokens $\mathbf{s}^w$ conditioned on the subword tokens for the previous words $\mathbf{s}^{\mathbf{w}^{<t}}$, with a marginalisation factor for all possible continuations of the target location word. This ensures that we compute the probability of the target location word being delimited by white space rather than continued by additional subwords, which would yield a different word than w. The marginalisation factor, introduced by Pimentel and Meister (2024), is computed as the ratio between the total probability of tokens that begin a new word $s \in \overline{\mathcal{S}}_{\text{bow}}$ following the predicted token appended to its preceding context $\mathbf{s}^{\mathbf{w}^{<t}} \circ \mathbf{s}^w$, divided over the total probability of all possible word-starting tokens at that position. This normalisation ensures that we compute a contextual sequence probability, where individual token probabilities are normalised to account for BOW tokenisers.

B Detailed False Belief task and tested Models

The open weight and open source models used in this work are listed in Table 2.

A more detailed overview of the task dimensions and example as explained in Section 3.1 can be found in Table 3.

Computational setup We run our models on heterogeneous GPU-enabled hardware, including H100 and A100 GPUs. The total running time for the FB dataset with the largest models is 15 minutes per model and experiment condition. Models up to 16B are run in full-precision mode (32-bit floats), larger model sizes are run in 16-bit precision mode.

C Traces

C.1 Detailed pre-training & mid-training traces

This appendix presents detailed learning curves for OLMo 2. We provide a subsampled overview (i.e., we sample 50 checkpoints evenly spaced by token count) of pre-training across all model sizes in Figure 7 and a complete view (all checkpoints) for the 7B model in Figure 8. The detailed run using all pre-training checkpoints shows considerable variation across checkpoints. However, we observe the same general trend when subsampling checkpoints: over the course of training, FB perfor-

Model	Authors	Sizes	Variants
OLMo 2	OLMo et al. (2025)	1B, 7B (exh), 13B, 32B	Ckpts, Base, SFT, DPO, RLVR
OLMo 3	Olmo et al. (2025)	7B, 32B	Base, SFT, DPO, RLVR, Thinking
Gemma 3	Team et al. (2025)	270M, 1B, 4B, 12B, 27B	Base, IT
Qwen 3	Yang et al. (2025)	4B, 30B	Base, IT, Thinking
K2-V2	Team et al. (2026)	70B	Base, IT
Llama 3.1, Llama 3.2	Grattafiori et al. (2024)	1B, 3B, 8B, 70B	Base, IT

Table 2: Overview of the models used in our experiments. An exhaustive (exh) checkpoint sweep was done for OLMo 2 7B.

Dimension	Level	Example
Premise		Ed arrives home after a long day at work. He puts his keys in the hall and leaves his bag in the study. Seana arrives home a few minutes later.
Knowledge state	True: character can see that the object is being moved	Ed watches as Seana moves the keys from the hall to the study. Afterwards, Ed goes to the bathroom.
	False: character cannot see that the object is being moved	Afterwards, Ed goes to the bathroom. Ed doesn't see Seana move the keys from the hall to the study.
Location mentions	Start location mentioned first Start location mentioned last	. . . puts his keys in the hall and leaves his bag in the study . . . moves the keys to the study from the hall
	End location mentioned first End location mentioned last	. . . leaves his bag in the study and puts his keys in the hall . . . moves the keys from the hall to the study
Knowledge cue	Implicit: action verb "go to"	Ed **goes to** get the keys from the [mask]
	Explicit: propositional attitude verb "think"	Ed **thinks** the keys are in the [mask]

Table 3: Overview of task variables and their realisations. Each scenario is constructed by combining one level from each variable, yielding 192 unique items.

mance tends to increase or decrease. Hence, we use the subsampled results in the main paper to draw our conclusions about the learning dynamics. Figure 9 shows the strict performance of mid-training. There seems to be no clear difference between the start and end of mid-training. We therefore did not incorporate mid-training analyses into the main body of this paper.

C.2 Developing linguistic capability

We also compare how quickly OLMo 2 acquires FBT capability, a *functional* linguistic capability, with general *formal* linguistic skill, measured through BLIMP accuracy, in Figure 10. The BLIMP benchmark contains samples for 67 linguistic phenomena, each of which isolates and tests a particular capability in syntax, morphology, or semantics (Warstadt et al., 2020). We randomly subsample the dataset to 10% (100 samples per phenomenon), given the number of checkpoints we are repeating the benchmark across.

The model rapidly (within 25B tokens) achieves $> 80\%$ accuracy across all linguistic phenomena on BLIMP, whereas it takes at least 300B tokens to stabilise on above-chance performance for the FB tasks. Further, there is no deterioration on BLIMP as the model trains, revealing a robust understanding of formal components. The behaviour on the FB dataset is also erratic and worsens during training for the False Belief knowledge state. Therefore, similar to earlier claims (Mahowald et al., 2024), we conclude that functional and formal understanding should be measured separately. While the latter may be required for the former, we cannot conclude that simply exposing the model to additional tokens will automatically lead to mindreading capabilities that are required for the FB task.

D Data Leakage Analysis

It may be possible that the contents from any of our evaluation benchmarks have made their way into the training data for any of the phases of the models used in our experiments (Ni et al., 2025;

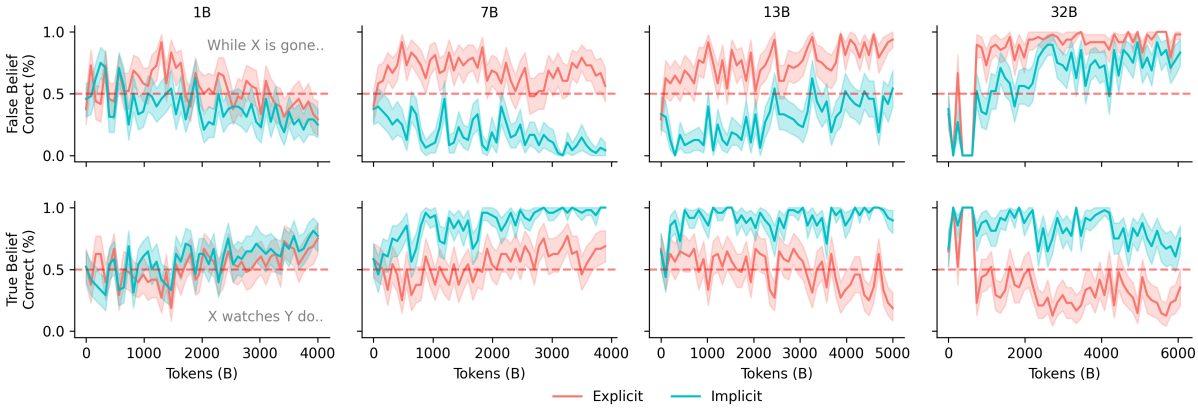

Figure 7: OLMo 2 traces of social intelligence for stage 1. Shaded areas are 95% confidence intervals.

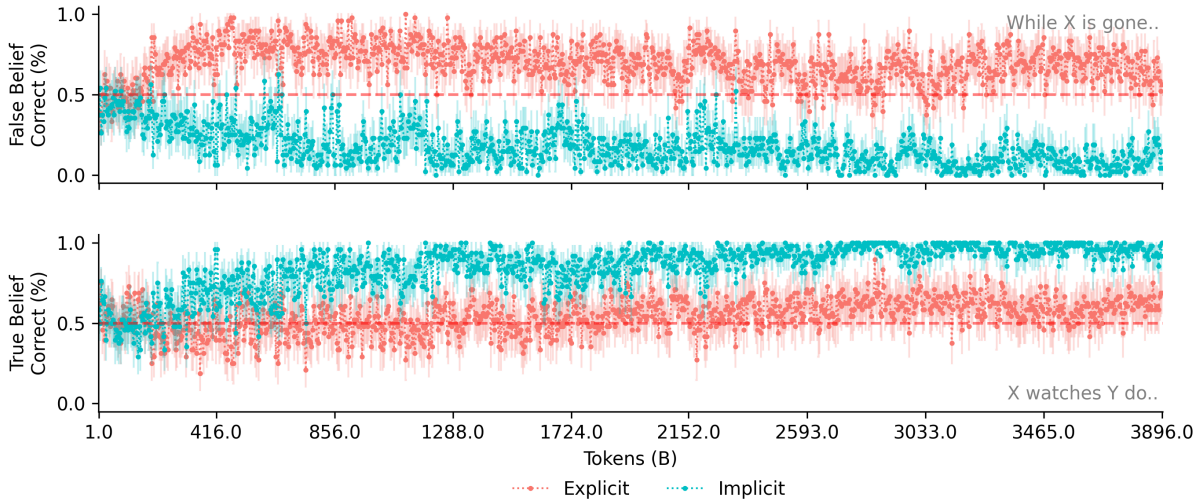

Figure 8: The exhaustive evaluation for stage 1 of OLMo 2 7B checkpoints to trace the emergence of social intelligence.

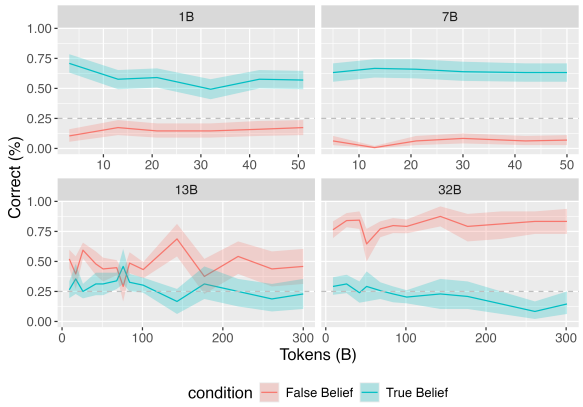

Figure 9: OLMo 2 mid-training traces of strict performance on our task.

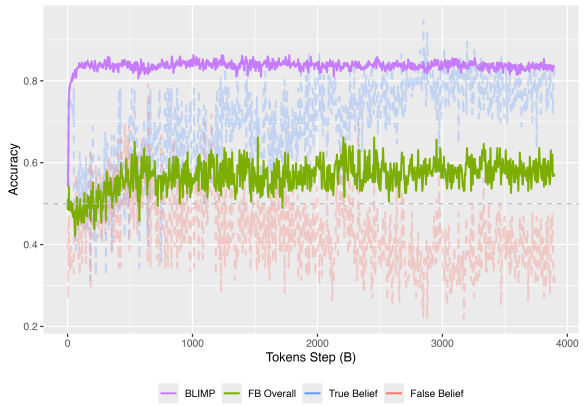

Figure 10: BLIMP vs FB task performance for OLMo 2 (7B). Solid lines show aggregated benchmark performance, whereas the dotted, lighter colours show FB performance split on knowledge state.

Balloccu et al., 2024). Our experiments investigate both open-weight models, as well as more completely open-source models, the main difference being access to their training data. For the open-source models, including `OLMo 2`, `OLMo 3`, and `K2-V2`, we can verify that the models did not directly observe test set samples at any point during training. Since our results mostly rely on results for the False Belief (FB) task, we focus our leakage analysis on this benchmark.

Pre-training data The aforementioned models publish their training data alongside the trained model. For OLMo 2, these are `olmo-mix-1124`[1] for pre-training, and `dolmino-mix-1124`[2] for mid-training. For OLMo 3, these are `Dolma 3`[3] for pre-, mid-, and long-context training, and various `Dolci 3` subsets for post-training. Each dataset is, in turn, a filtered combination of other datasets of various sources. Based on the document type in these sources, we can already disregard some of them as containing possible leakages, since they are very unlikely to include the benchmark data. For example, math web pages, math proofs, Wikipedia text, code, and academic paper content are unlikely to contain verbatim samples from the FB dataset. However, two sources potentially contain the dataset, as they are constructed by scraping the web freely (**CommonCrawl**) or by explicitly combining various datasets (**FLAN**). We discuss how we check each of these sources.

CommonCrawl We can check the index to see whether any known links to the FB data have been included in the CommonCrawl corpora (Common Crawl Foundation, 2026). We find seven hits for known links to the OSF archive (`https://osf.io/zp6q8`, `https://osf.io/agqwv/`, `https://archive.org/details/osf-registrations-agqwv-v1`, `https://archive.org/details/osf-registrations-zp6q8-v1`) for the FB dataset, though none of these contains the actual samples. For these pages, the crawler instead scraped the non-JS-enabled page text, in which the data is not rendered.

FLAN data FLAN data (Longpre et al., 2023) constitutes yet another mixture of many different datasets, and is included in `dolmino-mix-1124` and `Dolmo 3`. The bulk of this data stems from the 2021 iteration of FLAN, which predates the FB dataset and, therefore, cannot include it. However, the updated 2022 version might. Because this dataset is also included in the post-training dataset, we describe our checking mechanism there.

Post-training data The post-training dataset used by OLMo 2 is `allenai/tulu-3-sft-olmo-2-mixture`[4], or further filtered versions. As in the pre- and mid-training datasets, this dataset encapsulates many individual datasets as described in Tülu 3 (Lambert et al., 2024). Dolci 3, the post-training dataset

[1] `https://huggingface.co/datasets/allenai/olmo-mix-1124`

[2] `https://huggingface.co/datasets/allenai/dolmino-mix-1124`

[3] `https://huggingface.co/datasets/allenai/dolma3_pool`

[4] `https://huggingface.co/datasets/allenai/tulu-3-sft-olmo-2-mixture`

for OLMo 3 (specifically, `Dolci-Think-SFT`[5] and `Dolci-Instruct-SFT`[6]), shares many common sources with the Tülu dataset, but is expanded to include reasoning traces, and pools from additional datasets to more than double its size. Like before, we can already disregard certain sources as unlikely to contain the text samples from the FB benchmark, and avoid checking a set twice if it is included in both datasets. Since the resulting filtered set of samples to investigate is of a smaller order of magnitude (437K samples for OLMo 2, 3M samples for OLMo 3) than the pre-training data, checking for lexical overlap between FB data and the post-training samples is feasible.

We check all 192 FB samples to see whether they appear verbatim in the chat interactions provided in the post-training datasets. We select the complete story text (passages), both attention-check questions, and the first critical questions to see whether any of them literally appear in the dataset's logs. Our search took roughly 45 minutes on modern hardware and resulted in zero matches. Therefore, we conclude that none of the FB samples is included in the post-training data, and this benchmark can be safely used as a true test of generalised ToM ability as a surrogate for social intelligence.

[5] https://huggingface.co/datasets/allenai/Dolci-Think-SFT

[6] https://huggingface.co/datasets/allenai/Dolci-Instruct-SFT